NLP4DH 2026

The 6th International Conference on Natural Language Processing for the Digital Humanities

Proceedings of the Conference

July 4, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-427-9

Preface

We are delighted to welcome you to the 6th International Conference on Natural Language Processing for the Digital Humanities (NLP4DH 2026)! This year we are co-located alongside the Annual Conference of the Association for Computational Linguistics (ACL 2026) in San Diego, California.

NLP4DH brings together researchers from natural language processing and the humanities around shared questions: how computational tools shape interpretation, how language technologies treat under-represented languages and cultural materials, and what assumptions get baked into the systems we are continuously building. With LLMs now being used by so many people, these questions are especially urgent. Thus the question we posed this year was: looking beyond NLP helping DH, in what ways can DH help NLP?

We are happy to have accepted 36 wonderful papers. They span corpus construction for low-resource and historical languages, evaluation of LLMs against literary and cultural materials, OCR quality and provenance, computational narrative analysis, and methodological reflection on what NLP can tell us about cultural texts.

Alongside the main track, NLP4DH 2026 features a special track: "Reinterpreting NLP," which invited work that examines language-model training corpora, adapts mechanistic interpretability methods to humanities questions, and brings humanistic intuition back into language-technology design. We believe the papers in this proceeding speak to this important question.

We thank our program committee for their reviewing, our invited speakers for joining us, the ACL 2026 organizers for hosting NLP4DH, and the community for keeping the conversation going. We are furthermore thank our invited speakers, Sophie Hao (Boston University) and Luca Soldaini (Allen Institute for AI).

Thank you for joining us!

All best,

The NLP4DH 2026 Program Committee

Organizing Committee

Organizers

Sil Hamilton, Cornell University
Emily Öhman, Waseda University
Rebecca M. M. Hicke, Cornell University
Yuri Bizzoni, Aarhus University
Axel Bax, Cornell University
Jacob A. Matthews, Cornell University
Mika Hämäläinen, Metropolia University of Applied Sciences

Program Committee

Program Chairs

Sil Hamilton, Cornell University
Emily Öhman, Waseda University
Rebecca M. M. Hicke, Cornell University
Yuri Bizzoni, Aarhus University
Axel Bax, Cornell University
Jacob A. Matthews, Cornell University
Mika Hämäläinen, Metropolia University of Applied Sciences

Reviewers

Maria Antoniak, University of Colorado at Boulder
Frederik Arnold, Humboldt Universität Berlin
Mohammed Attia, Google
Won Ik Cho, Samsung Advanced Institute of Technology
Thibault Clérice, INRIA Paris - Almanach
Sourav Das, Indian Institute of Information Technology Kalyani
Stefania Degaetano-Ortlieb, Universität des Saarlandes
Anna Dmitrieva, Pushkin State Russian Language Institute
Quan Duong, University of Helsinki
Sebastian Oliver Eck, University of Oxford
Pascale Feldkamp, School of Communication and Culture, Aarhus University
Sijia Ge, University of Cincinnati
Luke Gessler, Indiana University
Balázs Indig, Eötvös Lorand University
Kenichi Iwatsuki, Mirai Translate
Antti Kanner, University of Turku
Yoshifumi Kawasaki, The University of Tokyo
Katerina Korre, Athena Research and Innovation Centre
Leo Leppänen, University of Helsinki
Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics
Aatu Liimatta, University of Helsinki
Dongqi Liu, Universität des Saarlandes
Kiara M. H. Liu, Cornell University
Pierre Magistry, Institut National des Langues et Civilisations Orientales
Enrique Manjavacas, Sigma Cognition
Sanghoon Oh, Cornell University
Shu Okabe, Technische Universität München
Hugo Gonçalo Oliveira, Universidade de Coimbra
Jaihyun Park, Indiana University
Niko Tapio Partanen, University of Helsinki
Lidia Pivovarova, University of Helsinki
Aynat Rubinstein, Hebrew University of Jerusalem
Martin Ruskov, University of Milan
Alejandro Sierra Múnera, Hasso Plattner Institute
Youngsook Song, Lablup

Jonne Sälevä, Brandeis University
Jouni Tuominen, University of Helsinki
Hisao USUI, Tokyo University of Agriculture and Technology, Tokyo Institute of Technology
Elissa Nakajima Wickham, Waseda University
Joshua Wilbur, University of Tartu
Sophie Wu, McGill University, McGill University
Shuo Zhang, Tufts University

Table of Contents

From OCR to Analysis: Tracking Correction Provenance in Digital Humanities Pipelines
Haoze Guo and Ziqi Wei 1

Frequency Accelerates Semantic Change: Evidence from 500 Years of Korean
Cheonkam Jeong and Yeeun Choi 13

Narrative Landscape: Mapping Narrative Dispositions Across LLMs
Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung 24

100,000+ Movie Reviews from Kazakhstan: Russian, Kazakh, and Code-Switched Texts
Rustem Yeshpanov 31

Quantifying Text Reuse Across Three Kṛṣṇa Yajurveda Recensions: Using Multi-Algorithm Computational Collation
So Miyagawa, Kyoko Amano, Yuzuki Tsukagoshi and Yuki Kyogoku 41

Data Contamination in Neural Hieroglyphic Translation: A Reproducibility Study
Ammar Toutou, Abdelrahman Harb and Christine Basta 50

Beyond Prompt-Sensitive Emotion Words: Stable Embeddings for Tang Poetry Analysis
Linyue Zhang and Feiyue Li 58

Measuring Embedding Sensitivity to Authorial Style in French: Comparing Literary Texts with Language Model Rewritings
Benjamin Icard, Lila Sainero, Alice Breton, Evangelia Zve and Jean-Gabriel Ganascia 69

Prompting the Past: Linguistic Transformations and Cultural Accuracy in AI-Generated Image Reconstructions for Multivocal Cultural Heritage
Ravini Wimalasuriya, Lea Krause and Gert-Jan Burgers 83

Temporal Text Classification with Large Language Models
Nishat Raihan and Marcos Zampieri 96

Evaluating Latin and Ancient Greek Sentence Alignment through Parallel Sentence Mining
Sebastian Reichbauer, Shu Okabe and Alexander Fraser 106

Modeling the "Dalet" Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis
Rachel Tal, Cheyn Shmuel Shmidman and Avi Shmidman 121

Beyond Genre Categories: How Narrative Pattern Coherence and Spanning Distance Shape Film Success
Zhichao Wang and ZEYU LYU 132

Register Mixing Is the Norm on the Web
Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner and Veronika Laippala . 138

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations
Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi and Shai Fine 150

PHMartialLawNER: A Tagalog Named Entity Recognition Corpus for the Philippine Martial Law Era
Abdiel Clarence Tabuzo, Vladimir Gray Velazco, Cassandra Cabral, Moneah Shaila Lacsam and Charmaine Salvador Ponay 167

Fluency and Faithfulness in Human and Machine Literary Translation
Sarah Griebel and Ted Underwood 178

Directional Alignment and Narrative Agency in Human–LLM Co-Writing
Halfdan Nordahl Fundal and Yuri Bizzoni 190

Bias Mitigation in Hiring-Related NLP: Interactions Between Masking, Rewriting, and Adversarial Debiasing
Alexandre Puttick and Rami El-Wazzi 202

Matching Meaning at Scale: Evaluating Semantic Search for 18th-Century Intellectual History through the Case of Locke
Yu Wu, Ananth Mahadevan, Filip Ginter, Michael Mathioudakis and Mikko Tolonen 214

Tracing Thematic Change in Early English-Language Science Fiction, 1818-1930
Jonathan Gordon 226

Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters
Tamás Kovács, Giuseppe Consolo and Georg Vogeler 236

Artistic Interventions for NLP Annotation Challenges: The Stress Test of Machinic Glossolalia
Tyler Grimes and Marshall Washington 242

In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica
Jean Barré 255

Computational Modeling of Educational Theory in Low-Socioeconomic Contexts
Jadon Swearingen, Mustafa Ocal, Md Tarique Hasan Khan and Labiba Jahan 264

Lost in Translation? Exploring the Shift in Grammatical Gender from Latin to Occitan
Ahan Chatterjee, Matthias Schöffel, Matthias Aßenmacher, Marinus Wiedner and Esteban Garces Arias 276

From Traditional Taggers to LLMs: A Comparative Study of POS Tagging for Medieval Romance Languages
Matthias Schöffel and Esteban Garces Arias 297

Statistical Structure in Indus Sign Sequences
Tanishk Tiwari 314

Exploring Topological Invariance in Semantic Embeddings
Fangzhou Gao and Justin Brody 320

MADRAG: Multi-Agent Debate with Retrieval-Augmented Generation for Training-Free Analytic Essay Scoring
Ali Keramati, Shiyuan Zhou, Sharad Mehrotra and Mark Warschauer 325

Never Care For What They Say ? Platform vs Genre Rules in Online Horror Narratives (2007–2024)
Alexandre Lionnet-Rollin and Florian Cafiero 346

StoicLLM: Preference Optimization for Philosophical Alignment in Small Language Models
Ishmam Khan, Sindhuja Thogarrati and Shuo Zhang 355

Between Whispers and Screams: Loudness Standard Deviation as a Proxy for Explicit Content Detection in US Romance Novels
Svenja Guhr 368

Computational Authorship Attribution in the Children's Tales of Oscar and Constance Wilde: The Case of "The Selfish Giant"
Liviu P Dinu, Alina Iacob and Cosmin Ciotlos 381

Evaluating Open-Source LLMs for Text Summarization and Named Entity Recognition in Long, Unstructured Text
Pauline Kister and Miriam Schirmer 390

Perspectives – Interactive Document Clustering for Qualitative Data Analysis
Tim Fischer and Chris Biemann 411

Program

Saturday, July 4, 2026

09:00 - 09:30 *Virtual Posters*

From OCR to Analysis: Tracking Correction Provenance in Digital Humanities Pipelines
Haoze Guo and Ziqi Wei

Frequency Accelerates Semantic Change: Evidence from 500 Years of Korean
Cheonkam Jeong and Yeeun Choi

100,000+ Movie Reviews from Kazakhstan: Russian, Kazakh, and Code-Switched Texts
Rustem Yeshpanov

Quantifying Text Reuse Across Three Kṛṣṇa Yajurveda Recensions: Using Multi-Algorithm Computational Collation
So Miyagawa, Kyoko Amano, Yuzuki Tsukagoshi and Yuki Kyogoku

Data Contamination in Neural Hieroglyphic Translation: A Reproducibility Study
Ammar Toutou, Abdelrahman Harb and Christine Basta

Beyond Prompt-Sensitive Emotion Words: Stable Embeddings for Tang Poetry Analysis
Linyue Zhang and Feiyue Li

Measuring Embedding Sensitivity to Authorial Style in French: Comparing Literary Texts with Language Model Rewritings
Benjamin Icard, Lila Sainero, Alice Breton, Evangelia Zve and Jean-Gabriel Ganascia

Evaluating Latin and Ancient Greek Sentence Alignment through Parallel Sentence Mining
Sebastian Reichbauer, Shu Okabe and Alexander Fraser

Beyond Genre Categories: How Narrative Pattern Coherence and Spanning Distance Shape Film Success
Zhichao Wang and ZEYU LYU

Register Mixing Is the Norm on the Web
Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner and Veronika Laippala

Saturday, July 4, 2026 (continued)

PHMartialLawNER: A Tagalog Named Entity Recognition Corpus for the Philippine Martial Law Era
Abdiel Clarence Tabuzo, Vladimir Gray Velazco, Cassandra Cabral, Moneah Shaila Lacsam and Charmaine Salvador Ponay

Fluency and Faithfulness in Human and Machine Literary Translation
Sarah Griebel and Ted Underwood

Matching Meaning at Scale: Evaluating Semantic Search for 18th-Century Intellectual History through the Case of Locke
Yu Wu, Ananth Mahadevan, Filip Ginter, Michael Mathioudakis and Mikko Tolonen

Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters
Tamás Kovács, Giuseppe Consolo and Georg Vogeler

Computational Modeling of Educational Theory in Low-Socioeconomic Contexts
Jadon Swearingen, Mustafa Ocal, Md Tarique Hasan Khan and Labiba Jahan

Never Care For What They Say ? Platform vs Genre Rules in Online Horror Narratives (2007–2024)
Alexandre Lionnet-Rollin and Florian Cafiero

Computational Authorship Attribution in the Children's Tales of Oscar and Constance Wilde: The Case of "The Selfish Giant"
Liviu P Dinu, Alina Iacob and Cosmin Ciotlos

Evaluating Open-Source LLMs for Text Summarization and Named Entity Recognition in Long, Unstructured Text
Pauline Kister and Miriam Schirmer

Perspectives – Interactive Document Clustering for Qualitative Data Analysis
Tim Fischer and Chris Biemann

09:30 - 09:45 *Opening Remarks*

09:45 - 10:15 *Invited Talk: Sophie Hao (Boston University)*

10:30 - 11:30 *Coffee Break*

Saturday, July 4, 2026 (continued)

11:30 - 13:00 *Session 1: Historical and Multilingual Texts*

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations
Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi and Shai Fine

Modeling the "Dalet" Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis
Rachel Tal, Cheyn Shmuel Shmidman and Avi Shmidman

Tracing Thematic Change in Early English-Language Science Fiction, 1818-1930
Jonathan Gordon

From Traditional Taggers to LLMs: A Comparative Study of POS Tagging for Medieval Romance Languages
Matthias Schöffel and Esteban Garces Arias

Lost in Translation? Exploring the Shift in Grammatical Gender from Latin to Occitan
Ahan Chatterjee, Matthias Schöffel, Matthias Aßenmacher, Marinus Wiedner and Esteban Garces Arias

In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica
Jean Barré

13:00 - 14:00 *Lunch Break*

14:00 - 14:30 *Invited Talk: Luca Soldaini (Allen Institute for AI)*

14:30 - 15:30 *Session 2: LLMs and Cultural Analysis*

Directional Alignment and Narrative Agency in Human–LLM Co-Writing
Halfdan Nordahl Fundal and Yuri Bizzoni

MADRAG: Multi-Agent Debate with Retrieval-Augmented Generation for Training-Free Analytic Essay Scoring
Ali Keramati, Shiyuan Zhou, Sharad Mehrotra and Mark Warschauer

Saturday, July 4, 2026 (continued)

Narrative Landscape: Mapping Narrative Dispositions Across LLMs
Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung

Between Whispers and Screams: Loudness Standard Deviation as a Proxy for Explicit Content Detection in US Romance Novels
Svenja Guhr

15:30 - 16:30 *Coffee Break*

16:30 - 17:30 *In-Person Posters*

Narrative Landscape: Mapping Narrative Dispositions Across LLMs
Donghoon Jung, Jiwoo Choi, Songeun Chae and Seohyon Jung

Prompting the Past: Linguistic Transformations and Cultural Accuracy in AI-Generated Image Reconstructions for Multivocal Cultural Heritage
Ravini Wimalasuriya, Lea Krause and Gert-Jan Burgers

Temporal Text Classification with Large Language Models
Nishat Raihan and Marcos Zampieri

Modeling the "Dalet" Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis
Rachel Tal, Cheyn Shmuel Shmidman and Avi Shmidman

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations
Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi and Shai Fine

Directional Alignment and Narrative Agency in Human–LLM Co-Writing
Halfdan Nordahl Fundal and Yuri Bizzoni

Bias Mitigation in Hiring-Related NLP: Interactions Between Masking, Rewriting, and Adversarial Debiasing
Alexandre Puttick and Rami El-Wazzi

Tracing Thematic Change in Early English-Language Science Fiction, 1818-1930
Jonathan Gordon

Artistic Interventions for NLP Annotation Challenges: The Stress Test of Machinic Glossolalia
Tyler Grimes and Marshall Washington

In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica
Jean Barré

Lost in Translation? Exploring the Shift in Grammatical Gender from Latin to Occitan
Ahan Chatterjee, Matthias Schöffel, Matthias Aßenmacher, Marinus Wiedner and Esteban Garces Arias

From Traditional Taggers to LLMs: A Comparative Study of POS Tagging for Medieval Romance Languages
Matthias Schöffel and Esteban Garces Arias

Statistical Structure in Indus Sign Sequences
Tanishk Tiwari

Exploring Topological Invariance in Semantic Embeddings
Fangzhou Gao and Justin Brody

MADRAG: Multi-Agent Debate with Retrieval-Augmented Generation for Training-Free Analytic Essay Scoring
Ali Keramati, Shiyuan Zhou, Sharad Mehrotra and Mark Warschauer

StoicLLM: Preference Optimization for Philosophical Alignment in Small Language Models
Ishmam Khan, Sindhuja Thogarrati and Shuo Zhang

Between Whispers and Screams: Loudness Standard Deviation as a Proxy for Explicit Content Detection in US Romance Novels
Svenja Guhr

From OCR to Analysis: Tracking Correction Provenance in Digital Humanities Pipelines

Haoze Guo
University of Wisconsin - Madison
College of Engineering
hguo246@wisc.edu

Ziqi Wei
University of Wisconsin - Madison
College of Letters and Sciences
zwei232@wisc.edu

Abstract

Optical Character Recognition (OCR) is a critical but error-prone stage in digital humanities text pipelines. While OCR correction improves usability for downstream NLP tasks, common workflows often overwrite intermediate decisions, obscuring how textual transformations affect scholarly interpretation. We present a provenance-aware framework for OCR-corrected humanities corpora that records correction lineage as base-anchored span edits, including edit type, correction source, confidence, and review status. Using a pilot corpus of historical texts, we compare downstream named entity extraction across raw OCR, fully corrected text, and provenance-filtered corrections. Our results show that correction pathways can substantially alter extracted entities and document-level interpretations, while provenance signals help identify unstable outputs and prioritize human review. We argue that provenance should be treated as a first-class analytical layer in NLP for digital humanities, supporting reproducibility, source criticism, and uncertainty-aware interpretation.

1 Introduction

Optical Character Recognition (OCR) is an important part of the digital humanities (DH) text analysis workflow which allows researchers to build searchable and analyzable corpora from scanned material (Smith, 2007; Nguyen et al., 2021). OCR is the first step in constructing resource-wide indices and retrieval systems for archival and newspaper collections and provides input for corpus-scale analysis processing pipelines (Ehrmann et al., 2020a; Düring et al., 2021, 2024). Unfortunately, OCR output produced from historical sources is often noisy because many historical materials are degraded, have non-standard characters, have unusual layouts, and have typographically varied characters (Nguyen et al., 2021; Neudecker et al., 2021). Consequently, many researchers use one or more methods to clean up their OCR output prior to performing natural language processing (NLP) such as using normalizations through rules, post-correction through neural networks, and/or manually editing (Evershed and Fitch, 2014; Nguyen et al., 2020; Lyu et al., 2021).

These correction steps improve readability and frequently positively impact downstream NLP performance. However, they can also make it difficult to understand how the original text may have evolved throughout the process and what changes are still uncertain. Corrections can erase analytical history from a text by replacing the original OCR output without any indication of the original content, the nature of the change(s), or which change(s) are uncertain to the analyst. This case represents a methodological challenge for Digital Humanities (DH) because future interpretations of data (e.g., named entities, dates, topics, and trends) will be based on textual transformations that cannot be seen by the analyst (Strange et al., 2014; van Strien et al., 2020). This issue is particularly important for historical named entity recognition (NER) because noisy OCR and historical variation are known to have impacted the quality, reliability, and the ability to evaluate the data extracted from text (Hamdi et al., 2019; Ehrmann et al., 2024; Boroş et al., 2020).

We suggest that OCR correction be not just one way; but should be modeled as a repeated sequence of traceable editorial decisions; the provenance of which are also able to be examined with downstream NLP outputs. In the DH context this framing is especially important since core DH values (Ockeloen et al., 2013; Romein et al., 2024) include source criticism, uncertainty and interpretive transparency. Previous DH infrastructure work (Ockeloen et al., 2013; Guo and Wei, 2026a) has already emphasised the need for multi-level provenance as a significant requirement of scholarly workflows; therefore, broader provenance standards provide a

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 1–12
July 4, 2026 ©2026 Association for Computational Linguistics

formal representation of such traces (Moreau et al., 2013; Lebo et al., 2013; Moreau and Groth, 2013). This context can also be seen by looking at broader audit-focused research that discusses the need for preserving transformation traces for longitudinal analysis and assessment (Guo and Wei, 2026c; Guo, 2026; Guo and Wei, 2026b), even across application domains.

In this paper, we introduce a provenance-aware representation for OCR-corrected humanities corpora and demonstrate its utility in a pilot study. We make three contributions:

- We introduce a span-level provenance schema for OCR correction that records edit lineage, correction source, confidence, and review status.
- We present a pilot empirical comparison of downstream named entity recognition (NER) across raw OCR, fully corrected text, and provenance-filtered corrected text.
- We provide a DH-oriented error analysis lens showing how provenance signals can identify unstable outputs and prioritize human review.

2 Related Work

Historical OCR pipelines face challenges beyond standard contemporary OCR settings. Noise can arise from page damage, bleed-through, historical fonts, line segmentation errors, footnotes, marginalia, and spelling conventions that differ from modern forms (Nguyen et al., 2021; Neudecker et al., 2021). As a result, OCR errors are not uniformly distributed across documents or regions of a page, and "correction" may involve both error repair and editorial normalization (Evershed and Fitch, 2014; Nguyen et al., 2021).

For DH scholars, this distinction matters. A correction that restores a malformed token may improve fidelity to the source, while a normalization step may improve computational consistency but alter historically meaningful variation (Strange et al., 2014). In both cases, downstream NLP outputs can shift: entities may appear or disappear, names may merge or split, and frequencies may change (van Strien et al., 2020; Hamdi et al., 2019; Ehrmann et al., 2024). If the intermediate correction decisions are not preserved, it becomes difficult to audit why a downstream result changed.

This issue is especially salient for named entity analysis on historical texts. Shared tasks and survey work (e.g., HIPE) document the difficulty of robust NER and linking in multilingual historical materials, including challenges from noisy OCR, diachronic variation, and annotation heterogeneity (Ehrmann et al., 2020b, 2022, 2024). Parallel work on historical newspaper infrastructures such as *impresso* has also emphasized the need to connect computational processing with scholarly interpretability and interface-level transparency (Ehrmann et al., 2020a; Düring et al., 2021, 2024).

Our motivation is therefore not only to improve OCR correction quality, but to preserve and expose the lineage of correction decisions so that computational findings remain inspectable and reproducible. Provenance research offers a strong conceptual foundation for this goal, including explicit modeling of entities, activities, and agents (Moreau et al., 2013; Lebo et al., 2013), but DH OCR pipelines need a lightweight, text-first representation that can operate at the span level and integrate directly with downstream NLP outputs.

We summarize how common provenance/annotation and OCR-output practices relate to the requirements of replayable span edits and policy-driven variant construction in Table 1. This comparison motivates our focus on base-anchored span-edit events coupled with explicit trust policies for reconstruction and audit.

3 Provenance-Aware Correction Schema

3.1 Design Principles

We design the schema around three principles: (1) **traceability**, so each correction can be linked to a document location and prior text span; (2) **tool-agnostic interoperability**, so the representation can capture rule-based, model-based, and human edits; and (3) **analytical usefulness**, so metadata can support downstream filtering and audit. Our design is informed by both DH provenance modeling practices and the broader PROV family of standards (Ockeloen et al., 2013; Moreau et al., 2013; Lebo et al., 2013).

We represent corrections at the **span level**, which provides more flexibility than token-only records while remaining simpler than character-level edit traces. A span may correspond to a token, multi-token phrase, or split/merge operation. This design is compatible with OCR post-correction workflows ranging from noisy-channel approaches to neural sequence models (Evershed and Fitch, 2014; Nguyen et al., 2020; Lyu et al., 2021; Nguyen et al.,

Approach	Span-level edits	Base-anchored offsets	Deterministic replay	Trust-policy filtering	Downstream trace links	Pipeline-ready artifact
PROV-DM / PROV-O (Moreau et al., 2013; Lebo et al., 2013)	✗	✗	✗	✗	✓	✗
TEI stand-off (`<standOff>`) (Text Encoding Initiative, 2026)	✓	✓	✗	✗	✗	✓
ALTO XML (Library of Congress, 2022)	✗	✓	✗	✗	✗	✓
hOCR (Bohnet et al., 2024)	✗	✓	✗	✗	✗	✓
PAGE-XML (Pletschacher and Antonacopoulos, 2010)	✗	✓	✗	✗	✗	✓
OCR-D (Neudecker et al., 2019; OCR-D Consortium, 2026b,a)	✗	✓	✗	✗	✗	✓
Span-edit events + provenance-aware filtering (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of our span-edit provenance + policy-driven reconstruction with common DH/OCR provenance and annotation practices.

2021).

3.2 Schema

We represent OCR correction provenance at the span level using a compact record schema that links each edit to a document location and revision step. Each record stores: document/page identifiers, span offsets, original and corrected text, edit type (e.g., substitution, split, merge), correction source (rule-based, model-assisted, or human), optional confidence, optional review status, and optional layout-zone metadata (e.g., body, header, footnote). This representation is sufficient to reconstruct correction lineage and to trace downstream NLP outputs (e.g., extracted entities) back to the edits they depend on (Ciccarese et al., 2013).

3.3 Illustrative Record

A single provenance record links an OCR span to its corrected form together with the metadata needed for audit (e.g., edit type, correction source, confidence, and review status). For example, if the OCR span *Madifon* is corrected to *Madison* by a model-assisted step with confidence 0.74 and no human approval, any downstream entity extracted from *Madison* can be traced back to that specific correction event. This span-level linkage is central to our analysis of entity volatility and uncertainty-aware filtering in the pilot study.

3.4 Serialization and Interoperability

Our span-level provenance schema is designed to be tool-agnostic and easy to integrate into existing NLP pipelines. The correction records can be serialized as **JSONL**, as a **tabular file** (CSV/Parquet), or as **stand-off annotations** that reference the underlying text by character offsets. Stand-off serialization is particularly useful for DH settings because it preserves the original OCR text while enabling multiple correction layers to be applied or compared without overwriting intermediate states.

In JSON-based NLP pipelines, provenance records can be embedded as an auxiliary field alongside the raw and corrected text variants, enabling downstream tasks to (i) reconstruct a chosen variant under a specified trust policy, and (ii) trace extracted entities back to the correction events that influenced them. For interoperability across tools, we recommend storing: stable document/page identifiers, explicit span offsets, explicit revision steps, and correction-source metadata, so that records remain portable even when tokenization or sentence segmentation changes. This makes the schema compatible with common DH workflows and with standard NLP processing stacks that expect JSON-serializable artifacts.

3.5 Offset Semantics and Edit Application

All correction events are anchored to a **base revision** (typically raw OCR) to avoid cascading offset drift and to preserve stable linkage from downstream outputs back to scan-derived text. Offsets (`span_start`, `span_end`) are Unicode-codepoint indices over the base text using half-open intervals `[start,end)`; `orig_text` must exactly match the base substring at that interval as an integrity check. To construct a variant, we select events under a trust policy (e.g., confidence threshold or `review_status=approved`), sort by `span_start` (tie-break by `event_id`), and apply them as base-anchored replacements, explicitly detecting overlaps for resolution (e.g., prefer human over model or defer to adjudication). Boundary-sensitive edits (split/merge) and OCR normalizations (line-break/whitespace reflow, hyphenation repair such as `"inter- national"`→`"international"`, paragraph merges) are represented as the same span-replacement events with explicit boundaries; large reflow regions are recorded as a single replacement to avoid brittle micro-edits.

4 Pilot Study Design

4.1 Corpus and Unit of Analysis

We conduct a pilot study on a small corpus of historical texts drawn from a digitized humanities collec-

tion. The corpus includes scanned materials with OCR output and a subset of passages that undergo correction. We treat this study as a methodological pilot rather than a benchmark-scale evaluation, with the goal of testing whether provenance signals provide useful analytical leverage in downstream NLP (van Strien et al., 2020).

Our primary unit of analysis is the **document-level NER output** for each text variant, with additional linkage to **span-level correction records**. For each document, we preserve the raw OCR text and record every applied correction as a provenance event, including span offsets, edit type, correction source, and optional confidence/review status. This design supports both aggregate comparison, how the extracted entity inventory changes across variants, and auditability which correction events are associated with unstable entities.

4.2 Text Variants and Provenance Filtering Policy

For each document, we construct three text variants from the same OCR source: **Raw OCR** (no post-correction), **Fully corrected** (all available corrections applied), and **Provenance-filtered** (only corrections meeting a provenance criterion).

The provenance-filtered variant operationalizes a conservative "trust policy" over corrections. In the main comparison, this policy is confidence-based (e.g., confidence ≥ 0.70) and may optionally require human approval when such metadata is available. We additionally report a threshold sensitivity analysis (Table 3) that varies the filter strictness, including higher confidence thresholds and a human-approved-only condition. This provides a transparent characterization of the coverage–stability tradeoff induced by provenance-aware filtering, rather than presenting a single corrected corpus as an implicit ground truth.

This design separates two pipeline questions that are often conflated in DH OCR workflows: (i) whether correction improves downstream extraction at all, and (ii) which corrections should be treated as sufficiently reliable for interpretive analysis.

4.2.1 Where Confidence Comes From

In our pilot, `confidence` is an optional provenance attribute attached to each correction event and emitted by the correction pathway that proposed the edit. For **model-assisted** corrections, `confidence` corresponds to the model's own edit score, used to rank suggested substitutions and split/merge operations. For **rule-based** corrections, `confidence` is derived from rule certainty (e.g., high-confidence deterministic patterns versus lower-confidence heuristic matches). For **human** corrections, we treat `review_status=approved` as the primary reliability signal; when an explicit reviewer score is available, we store it as `confidence`.

These scores are not assumed to be calibrated probabilities, and they are not assumed comparable across different tools or correction sources. Accordingly, we use `confidence` only as a *within-pipeline ranking signal* for sensitivity analysis. Thresholds (e.g., confidence ≥ 0.70 in the main provenance-filtered condition) are applied under a fixed correction stack and interpreted operationally rather than as a universal notion of correctness. The threshold sweep (Table 3) is therefore intended to make the coverage–stability tradeoff explicit under a fixed scoring regime, not to claim that a given numeric threshold generalizes across tool stacks.

4.3 Downstream Task and Quantitative Metrics

We use named entity recognition (NER) as the downstream task because entity extraction is common in DH workflows and is highly sensitive to OCR noise, normalization, and segmentation decisions (Hamdi et al., 2019; Boroş et al., 2020; Ehrmann et al., 2020b, 2022, 2024; Clausner et al., 2011). We run the same NER pipeline over all text variants and compare the resulting entity outputs.

4.3.1 NER Pipeline Configuration

We use a single, fixed NER pipeline across all text variants to isolate the effect of OCR correction and provenance filtering. Concretely, we sentence-segment the text, tokenize using the model's native subword tokenizer, and apply a transformer-based NER model fine-tuned on CoNLL-2003 (implemented via the HuggingFace Transformers library) (Devlin et al., 2019; Liu et al., 2019; Wolf et al., 2020). We keep all inference settings fixed across variants.

Because historical corpora exhibit diachronic spelling and orthographic variation, we avoid additional spelling modernization or normalization inside the NER pipeline beyond what is already implied by the OCR correction variants under study. This design prevents conflating "downstream robustness tricks" with the provenance-aware correction analysis. We acknowledge that domain

adaptation can substantially affect absolute NER accuracy in historical settings; our goal here is to characterize how correction pathways change extracted entities under a consistent downstream model (Ehrmann et al., 2024; Hamdi et al., 2019; Boroş et al., 2020).

To quantify differences across variants, we report:

- **Entity mentions**: total number of extracted entity mentions per variant.
- **Unique entities**: number of distinct extracted entity surface strings per variant.
- **Entity overlap**: Jaccard similarity between unique-entity sets across variants.
- **Entity volatility**: entities that (i) appear in one variant but not another, or (ii) appear in both but with changed surface form or span boundary.

In addition, we compute the share of volatile entities linked to **low-confidence or unreviewed correction spans**. This measure provides an audit-oriented view of instability: instead of treating volatility as an opaque downstream artifact, it tests whether provenance fields can identify subsets of edits that disproportionately contribute to unstable outputs.

4.4 Entity Linking as a Secondary Downstream Task

To test whether provenance-aware correction affects not only entity extraction but also downstream *resolution* of entities to knowledge bases, we include a lightweight entity linking (EL) analysis on a pilot subset. For each text variant, we take extracted entity mentions and apply an off-the-shelf EL system to link mentions to a reference knowledge base (e.g., Wikipedia/Wikidata identifiers). We focus on mentions that are volatile across variants, since these are the cases where small surface-form or boundary changes are most likely to alter linking decisions.

We report (i) **linking coverage** (fraction of mentions assigned a KB identifier), (ii) **link stability** across variants, fraction of mentions that resolve to the same KB identifier, and (iii) a small **manual audit of disagreements** to characterize whether changes reflect improved disambiguation, harmful normalization, or overconfident linking on noisy strings. This EL check is designed as a methodological stress test rather than a benchmark-scale evaluation.

4.5 Attribution of Entity Changes to Correction Events

To associate downstream entity differences with provenance records, we compute an **association** (not causal proof) between each entity mention and nearby correction events. We first attempt **span overlap**: an entity mention is associated with a correction record if the mention character offsets overlap the correction span offsets in that same variant. For edits that may shift offsets (notably split/merge operations), we use a bounded local fallback: we search for correction events within a window of $\pm W$ characters around the entity mention (we set $W = 50$ in the pilot) and choose the closest event by absolute offset distance. Ties are broken by preferring (i) overlap over proximity, (ii) edit types that can directly affect boundaries (split/merge) over substitutions, and (iii) higher-confidence events when available.

We use this association to support audit and summarization (e.g., "how many volatile entities are linked to low-confidence edits"), and we interpret results as identifying *likely contributing edits* rather than definitive causes.

4.6 Validation of the Attribution Heuristic

Because proximity-based association can produce spurious links (nearby edits may be unrelated), we validate the attribution heuristic on a labeled sample. We randomly sample linked (entity, correction-event) pairs and ask annotators to judge whether the correction plausibly contributed to the entity change (yes/no). We report the precision of the heuristic and use this validation to qualify interpretation of the signal-utility analysis.

4.7 Qualitative Coding Protocol

To complement the quantitative comparison, we manually inspect a pilot sample of volatile entities and assign each case to an error category (e.g., OCR noise, normalization shift, split/merge boundary issue, layout artifact). The goal of this coding step is not to produce an exhaustive taxonomy, but to evaluate whether provenance fields (confidence, edit type, layout zone, review status) provide actionable diagnostic value. We use the qualitative categories to interpret the mechanisms behind volatility and to assess which provenance signals are most useful for DH-oriented review and source criticism.

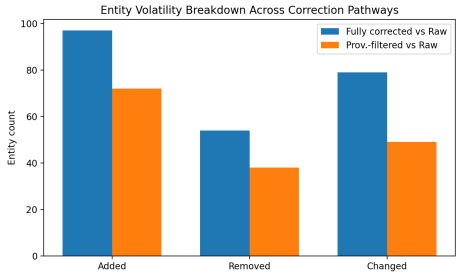


Figure 1: Breakdown of entity volatility across correction pathways.

Variant	Ment.	Uniq.	Jac. vs Raw	Vol.	% Unrev.
Raw OCR	1184	512	1.00	–	–
Fully corrected	1342	566	0.69	176	68%
Prov.-filtered	1287	548	0.76	121	76%

Table 2: NER output differences across OCR correction pathways. "Vol." counts entities that appear/disappear or change surface form/span boundary across variants; "% Unrev." is the share of volatile entities linked to unreviewed edits.

5 Results

5.1 Quantitative Differences Across Correction Pathways

Across the pilot corpus, correction activity is unevenly distributed across documents and page regions, and many edits occur in short spans that affect candidate entity strings. This matters because a relatively small number of span edits can produce disproportionately large downstream changes in NER outputs, particularly when edits touch capitalization, rare proper nouns, or token boundaries.

Figure 1 provides a compact view of entity volatility relative to raw OCR. The fully corrected condition produces more *added* and *changed* entities than the provenance-filtered condition, indicating stronger transformation of the extracted entity inventory. The provenance-filtered condition still preserves a substantial portion of correction gains, but reduces high-risk changes. In practice, this suggests that provenance filtering behaves less like "undoing correction" and more like selecting a correction pathway with a different analytical risk profile.

Table 2 quantifies these differences. Relative to raw OCR, the fully corrected variant increases both entity mentions (1184 $\rightarrow$ 1342) and unique entities (512 $\rightarrow$ 566), consistent with correction improving recognizability of names and places. However, these gains coincide with the highest volatility (176 volatile entities), meaning that correction alters not only the volume of extracted entities but also the stability of the extracted inventory. The provenance-filtered variant retains most of the coverage gains (1287 mentions; 548 unique entities) while reducing volatility (121), suggesting that provenance-aware filtering can improve analytical stability without reverting to raw OCR.

A key observation is that volatility is not uniformly distributed across edits: a substantial fraction of volatile entities in the corrected conditions is linked to low-confidence or unreviewed correction events (Table 2). This does not imply that all low-confidence edits are incorrect; rather, it indicates that these edits are most likely to produce downstream differences that analysts may want to inspect, qualify, or report explicitly.

5.2 Threshold Sensitivity of Provenance Filtering

Table 3 shows how downstream outputs change as the provenance filter becomes stricter. As the confidence threshold increases (or when only human-approved edits are used), entity volatility decreases monotonically, but entity coverage also declines. This tradeoff is expected and analytically useful: different DH workflows may prefer different operating points depending on whether the priority is recall-oriented exploration or conservative interpretation.

Figure 2 visualizes the same sweep as a coverage–stability curve. Moving from "All corrections" to stricter thresholds shifts the operating point toward lower instability, but also lower coverage. In this pilot, a mid-range threshold (e.g., confidence ≥ 0.70) yields a balanced regime: many of the coverage improvements from correction are retained, while a substantial portion of volatility is reduced. We emphasize that thresholds are not universal: confidence values are tool- and workflow-dependent. The contribution here is that provenance makes the operating point explicit and reportable, rather than implicit in a single overwritten corrected text.

5.3 Provenance Signals Predicting Instability

Beyond overall thresholds, provenance fields help identify which correction operations are most likely to affect downstream entity extraction. Table 4 summarizes the utility of several provenance signals for predicting volatility. Two patterns are par-

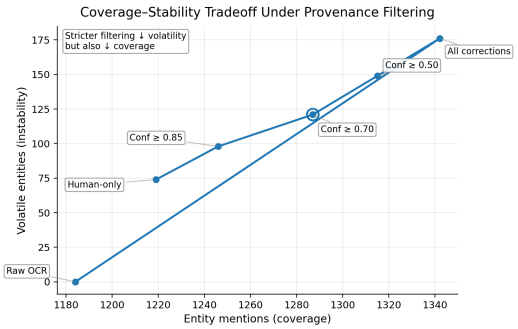


Figure 2: Coverage–stability tradeoff under provenance filtering.

Filter	Mentions	Unique	Jaccard vs Raw	Volatile
Raw OCR	1184	512	1.00	–
All corrections	1342	566	0.69	176
Conf ≥ 0.50	1315	559	0.73	149
Conf ≥ 0.70	1287	548	0.76	121
Conf ≥ 0.85	1246	531	0.79	98
Human-approved only	1219	523	0.82	74

Table 3: Threshold sensitivity for provenance-filtered correction.

ticularly salient. First, boundary-affecting edits (split/merge) exhibit the highest volatility lift despite being relatively infrequent, consistent with entity extraction being sensitive to segmentation decisions. Second, non-body layout zones (e.g., headers and footnotes) are instability hotspots, reflecting known OCR challenges in regions where typography and layout differ from running prose.

Low confidence and unreviewed status provide coarse but useful uncertainty flags. These fields are less diagnostic than edit type or layout zone, but they are broadly applicable across correction pathways and can support conservative analysis regimes (e.g., confidence-based filtering) and targeted inspection when human review capacity is limited.

These associations are produced by a deterministic overlap-plus-window heuristic and should be interpreted as likely contributing edits rather than definitive causal attributions; we therefore validate heuristic precision on a labeled sample and report it alongside the signal-utility analysis.

5.4 Qualitative Error Categories and Diagnostic Use

Table 5 summarizes qualitative categories observed among volatile entities. The most common category is false entities caused by OCR noise, followed by normalization shifts and boundary errors (split/merge cases). These patterns align with known OCR and historical NER challenges (van Strien et al., 2020; Ehrmann et al., 2024), but the key result in our setting is that provenance metadata helps localize and explain failures rather than only reporting aggregate quality.

Signal	%	Flag vol.	Lift
Low conf. (< 0.70)	31%	7.8%	2.7×
Unreviewed	90%	5.9%	1.7×
Split/merge	9%	12.5%	3.3×
Non-body zone	14%	10.1%	2.6×

Table 4: Utility of provenance signals for identifying instability. "Flag vol." is the volatility rate among entities associated with the flagged condition; "Lift" compares flagged vs unflagged.

Different provenance fields are useful for different diagnostic purposes. Confidence and review status help identify risky corrections for triage; edit type helps distinguish normalization shifts from segmentation effects; and layout-zone metadata supports page-region-based auditing of headers, footnotes, and other non-body text. Taken together, provenance enables analysts to answer not only "what changed" (entity volatility), but also "which editorial operations caused the change" and "which changes are plausible candidates for targeted review."

5.5 Entity Linking Sensitivity to Correction Pathways

Entity linking is particularly sensitive to surface form and boundary decisions: small spelling changes, whitespace insertion, or normalization can alter candidate generation and disambiguation. In our pilot EL analysis, fully corrected text increases linking coverage relative to raw OCR by producing more linkable surface strings, but it also introduces additional link instability for a subset of volatile mentions. Provenance-filtered text typically preserves much of the coverage gain while reducing the rate of link changes tied to low-confidence or unreviewed corrections.

Qualitatively, we observe three recurring EL disagreement patterns: (i) **repair-driven improvements** where correcting OCR noise enables linking to the intended entity, (ii) **normalization-driven shifts** where historical spellings or abbreviations are modernized and shift candidate selection, and (iii) **boundary-driven failures** where split/merge

Pattern	%	Diagnostic provenance cue(s)
OCR-noise false entity	29%	Low confidence; model-assisted; often disappears under filtering
Normalization shift	21%	`edit_type=substitution`; unreviewed; surface-form changes
Boundary (split/merge)	18%	`edit_type=split/merge`; span boundary drift
Layout artifact	17%	`layout_zone=header/footnote`; non-body text
Ambiguous correction	15%	Medium confidence; unreviewed; competing entity strings

Table 5: Qualitative categories among volatile entities.

edits change mention spans and trigger different candidate sets. These results reinforce the core claim that correction pathways are analytically consequential: downstream interpretations can shift not only in *which* entities are extracted, but also in *which real-world entities* those mentions are resolved to.

6 Discussion

The pilot study shows that OCR correction pathways can produce meaningfully different analytical corpora from the same scanned source. Raw OCR, fully corrected text, and provenance-filtered text do not merely differ in readability; they produce different entity inventories, different levels of volatility, and different evidentiary risks for downstream interpretation. This supports our central claim that OCR correction should be treated as an inspectable analytical process rather than a hidden preprocessing step.

Provenance makes these differences explicit. By preserving edit source, confidence, review status, edit type, and layout-zone information, analysts can treat correction as a set of documented editorial choices rather than as an invisible transformation. The threshold sweep and coverage–stability curve show how different trust policies expose different operating points: a recall-oriented analysis may prefer broader correction coverage, while a conservative interpretive analysis may prefer provenance-filtered text with lower volatility.

The qualitative categories highlight a recurring DH concern: some corrections restore fidelity to the source (repair), while others normalize historically meaningful variation (normalization). Provenance makes these transformations visible and therefore contestable. This is important for interpretive work where spelling variation, abbreviations, and typography may carry historical meaning.

Provenance also supports locating where instability originates. Our signal-utility analysis suggests that boundary edits and non-body layout zones are particularly consequential. This enables a more DH-aligned auditing practice in which analysts can attribute changes in extracted entities to specific editorial operations, rather than treating entity differences as opaque model behavior.

6.1 Limitations and Future Work

This study has several limitations. First, it is a methodological pilot rather than a benchmark-scale evaluation, and the corpus is not intended to represent the full diversity of DH materials. Future work should test the schema across multiple DH domains, including historical newspapers, literary prose, poetry, correspondence, and archival administrative records, because each genre introduces different OCR and correction failure modes. Second, we use NER as the primary downstream probe; additional tasks such as topic modeling, text classification, relation extraction, and entity linking at larger scale may expose different kinds of correction sensitivity. Additionally, although confidence scores and metadata can be utilized with some degree of familiarity to compare various corrections, they serve as a signal of provenance rather than as calibrated probabilities.

Future research into the analysis of correction pathways which account for provenance must be extended to cover multilingual collections and more complex historical page layouts. In these instances, greater differences in typography and segmentation between different collections will be present. A second area for further research is to expand the downstream tasks used for example with regard to entity linking where small changes in the surface (visibility) features result in significant impact on linking decisions. Lastly, the construction of provenance-aware pipelines creates an opportunity to establish standardized reporting criteria for DH-related NLP. Researchers will be able to use these documents as a means of reporting their own sensitivity to correction pathways while also documenting the provenance signals that directly contributed most to their instability.

7 Conclusion

We presented a provenance-aware framework for OCR-corrected humanities corpora and a pilot study showing how correction pathways alter down-

stream named entity extraction. By preserving correction lineage at the span level, our approach helps make NLP outputs more auditable, reproducible, and interpretable in DH workflows. We argue that provenance should be treated as a first-class analytical layer in OCR-to-NLP pipelines, not merely as implementation metadata.

References

Klaus Bohnet and 1 others. 2024. hocr: The embedded ocr workflow and output format (specification). Specification (GitHub repository).

Emanuela Boroş, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, pages 431–441.

Paolo Ciccarese, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair J. G. Gray, Carole Goble, and Tim Clark. 2013. Pav ontology: Provenance, authoring and versioning. *Journal of Biomedical Semantics*.

Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. 2011. Scenario driven in-depth performance evaluation of document layout analysis methods. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity: Introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(1):20–40.

Marten Düring, Roman Kalyakin, Estelle Bunout, and Daniele Guido. 2021. Impresso inspect and compare: Visual comparison of semantically enriched historical newspaper articles. *Information*, 12(9):348.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):27:1–27:47.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020a. Language resources for historical newspapers: The impresso collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 958–968.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020b. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2020)*, volume 12260 of *Lecture Notes in Computer Science*, pages 288–310.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Working Notes of CLEF 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 1038–1063.

John Evershed and Kent Fitch. 2014. Correcting noisy ocr: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014)*, pages 45–51.

Haoze Guo. 2026. Consentdiff at scale: Longitudinal audits of web privacy policy changes and ui frictions. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA. Association for Computing Machinery.

Haoze Guo and Ziqi Wei. 2026a. Behind the feed: A taxonomy of user-facing cues for algorithmic transparency in social media. *Preprint*, arXiv:2602.03121.

Haoze Guo and Ziqi Wei. 2026b. Hidden-in-plain-text: A benchmark for social-web indirect prompt injection in rag. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 8337–8340, New York, NY, USA. Association for Computing Machinery.

Haoze Guo and Ziqi Wei. 2026c. Temporal drift in privacy recall: Users misremember from verbatim loss to gist-based overexposure. In *Proceedings of the 2025 International Conference on Human-Engaged Computing*, ICHEC '25, New York, NY, USA. Association for Computing Machinery.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334.

Timothy Lebo, Satya Sahoo, Deborah McGuinness, and 1 others. 2013. Prov-o: The prov ontology. W3C Recommendation.

Library of Congress. 2022. Alto: Technical metadata for layout and text objects. Standard (Library of Congress).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv*.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural ocr post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Luc Moreau and Paul Groth. 2013. *Provenance: An Introduction to PROV*, volume 3 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool.

Luc Moreau, Paul Groth, and 1 others. 2013. Prov-dm: The prov data model. W3C Recommendation.

Clemens Neudecker, Konstantin Baierer, Don Borgman, Markus Federbusch, Matthias Boenig, Wolfgang Maier, Christian Reul, and Frank Puppe. 2019. Ocr-d: An end-to-end open source ocr framework for historical printed documents. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2019)*.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing (HIP 2021)*, pages 13–18.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickaël Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Computing Surveys*, 54(6):124:1–124:37.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, and Antoine Doucet. 2020. Neural machine translation with bert for post-ocr error detection and correction. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, pages 333–336.

C. J. Ockeloen, A. S. Fokkens-Zwirello, S. ter Braake, P. T. J. M. Vossen, V. de Boer, A. T. Schreiber, and S. Legene. 2013. Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the 3rd International Workshop on Linked Science (LISC 2013)*.

OCR-D Consortium. 2026a. Ocr-d workflow guide. OCR-D Documentation.

OCR-D Consortium. 2026b. Requirements on handling mets/page. OCR-D Specification.

Stefan Pletschacher and Apostolos Antonacopoulos. 2010. The page (page analysis and ground-truth elements) format framework. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*.

C. Annemieke Romein, Tobias Hodel, Femke Gordijn, Joris van Zundert, Alix Chagué, Milan van Lange, Helle Strandgaard Jensen, Andy Stauder, Jake Purcell, Melissa Terras, Pauline van den Heuvel, Carlijn Keijzer, Achim Rabus, Chantal Sitaram, Aakriti Bhatia, Katrien Depuydt, Mary Aderonke Afolabi, Anastasiia Anikina, Elisa Bastianello, and 1 others. 2024. Exploring data provenance in handwritten text recognition infrastructure: Sharing and reusing ground truth data, referencing models, and acknowledging contributions. *Journal of Data Mining and Digital Humanities*.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 629–633.

Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. Mining for the meanings of a murder: The impact of ocr quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1).

Text Encoding Initiative. 2026. The `standOff` element (tei p5 guidelines). TEI Guidelines reference.

Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, pages 484–496.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*.

A Schema Specification and Serialization

A.1 Normative field specification

We represent OCR correction provenance as a sequence of **span-edit events** anchored to a **base revision** of the text (typically raw OCR). Each event specifies *where* an edit applies in the base text, *what* replacement occurs, and *metadata* that supports audit and policy-based reconstruction.

Required invariants. Each event MUST satisfy *base anchoring* (offsets refer to `base_revision`, with `base_revision=0` for raw OCR in this paper), *half-open spans* (`[span_start, span_end)` with `span_start < span_end`), *Unicode-codepoint offsets*, and an *integrity check* (`orig_text` matches the base substring at `[span_start, span_end)`).

Event fields. Table 6 defines the canonical fields and constraints used by the schema described in §3 and referenced throughout the pilot study.

Versioning. We include `schema_version` in each event. Minor versions may add optional fields without changing existing field meanings; major versions may change field meanings or constraints.

Field	Type	Req.	Meaning / constraints
schema_version	string	Y	Semantic version, e.g., "1.0.0".
event_id	string	Y	Unique identifier within a corpus (UUID recommended).
doc_id	string	Y	Stable document identifier.
page_id	string/int	Y	Page identifier (or region id if pageless).
base_revision	int	Y	Revision index of anchor text (0 = raw OCR).
span_start	int	Y	Inclusive offset in base revision (Unicode codepoints).
span_end	int	Y	Exclusive offset; span_start < span_end.
orig_text	string	Y	Exact base substring at [span_start, span_end).
new_text	string	Y	Replacement string; empty string indicates deletion.
edit_type	enum	Y	{substitute, insert, delete, split, merge, normalize}.
source	enum	Y	{rule, model, human}.
confidence	float	O	In [0,1]; tool-specific ranking score (not assumed calibrated).
review_status	enum	O	{unreviewed, approved, rejected}.
reviewer_id	string	O	Pseudonymous reviewer id, if applicable.
layout_zone	enum/string	O	e.g., body, header, footnote, caption.
note	string	O	Free-form rationale/tag (e.g., "hyphenation repair").

Table 6: Canonical span-edit provenance event schema used in this paper.

All experiments in the current draft assume a fixed schema.

A.2 Canonical JSONL and OCR operation examples

We serialize events as JSONL (one event per line). This supports streaming pipelines and common NLP artifact formats while keeping events queryable and replayable.

Canonical example. The main text uses an illustrative correction *Madifon→Madison* (model-assisted; confidence 0.74; no human approval). In our schema this is represented as:

```
{"schema_version":"1.0.0","event_id":"c2f1...",
"doc_id":"doc_017","page_id":3,"base_revision":0,
"span_start":1284,"span_end":1291,
"orig_text":"Madifon","new_text":"Madison",
"edit_type":"substitute","source":"model",
"confidence":0.74,"review_status":"unreviewed",
"layout_zone":"body"}
```

OCR-specific operations as span edits. We encode OCR cleanup as span replacements: *hyphenation repair* ("inter-national"→"international") via normalize or substitute spanning the hyphen and break; *line-break normalization* (newline→space) via normalize; *whitespace reflow* (collapse repeated whitespace / remove intra-word spaces) via normalize; *split/merge boundaries* (e.g., "NewYork"→"New York") via split/merge; and *paragraph merges/reflow* as a single larger-span replacement to avoid brittle micro-edits while preserving auditability.

B Provenance-Aware Filtering Method

B.1 Inputs and outputs

Inputs. (i) Base text T (raw OCR; base_revision=0); (ii) a set of correction events E in the schema of Table 6; and (iii) a trust policy π over event metadata.

Outputs. A provenance-filtered variant T_π constructed by replaying the subset of events selected by π, alongside an application trace (applied/skipped/conflicted events) for audit.

B.2 Policy language and policies used in this paper

A policy π is a predicate on events. We use: **All corrections** $\pi(e) \equiv \texttt{true}$; **Confidence threshold** $\pi_\tau(e) \equiv (\texttt{confidence}(e) \geq \tau)$ with $\tau \in \{0.50, 0.70, 0.85\}$ (Table 3) and $\tau = 0.70$ as the main provenance-filtered condition; and **Human-approved only** $\pi_{\text{approved}}(e) \equiv (\texttt{review_status}(e) = \texttt{approved})$.

We emphasize that confidence is treated as a within-pipeline ranking signal, not a calibrated probability.

B.3 Deterministic replay and overlap handling

Ordering. Given selected events $E_\pi = \{e \in E : \pi(e)\}$, we sort by span_start ascending, break

ties by `event_id`, and replay as base-anchored replacements.

Conflicts. Because all events are anchored to the same base revision, overlaps can be detected explicitly. When overlapping events are incompatible, we resolve deterministically under a fixed rule (e.g., prefer human over model over rule; prefer `review_status=approved` over `unreviewed`; otherwise defer to adjudication). The key methodological point is that overlap handling is *explicit and reportable*, rather than silently overwritten.

Frequency Accelerates Semantic Change: Evidence from 500 Years of Korean

Cheonkam Jeong*
University of California, Irvine
Irvine, CA, United States
cheonkaj@hs.uci.edu

Yeeun Choi
Korea International School
Republic of Korea
lychoi0809@gmail.com

Abstract

The "law of conformity," the finding that frequent words are semantically stable, has been treated as a broad regularity of language change. We show it does not hold for Korean. Using diachronic word embeddings trained on historical corpora spanning 500 years (15th–20th centuries), we find a robust *positive* correlation between frequency and semantic shift: high-frequency Korean words change *more*, not less. The pattern survives six robustness controls and is validated against an English replication. Partial correlation analysis reveals that the role of polysemy in mediating the frequency–change relationship is not fixed but depends on time resolution and corpus homogeneity. We connect the reversal to frequency-driven reductive processes, including grammaticalization, semantic bleaching, and domain shift, that are especially productive in Korean. The frequency–change relationship is not a fixed regularity but varies with language typology and analytical conditions.

1 Introduction

A central question in historical linguistics is what drives words to change meaning over time. Quantitative approaches to this question have accelerated in recent years, enabled by diachronic word embedding methods that measure semantic shift across time periods (Kutuzov et al., 2018; Tahmasebi et al., 2021). Among the most influential findings is the "law of conformity" proposed by Hamilton et al. (2016), which states that more frequent words are more semantically stable: frequency and semantic change are negatively correlated ($\beta_f \in [-1.26, -0.27]$ across languages and corpora, 1800–2000). This has been interpreted as a statistical regularity of language change, alongside a parallel "law of innovation" linking polysemy to accelerated change.

However, this finding rests primarily on evidence from English and a small number of other Indo-European languages observed over roughly two centuries. Whether the law of conformity holds cross-linguistically, particularly for typologically distant languages and over longer time scales, remains an open question. Xu and Kemp (2015) provided early evidence that not all proposed regularities of semantic change generalize uniformly, while Dubossarsky et al. (2017) demonstrated that some observed correlations may reflect artifacts of the embedding process. Even within Hamilton et al. (2016)'s own English data, the relationship between frequency and polysemy is non-trivial: the two are highly correlated, and partial correlation analysis reverses the apparent direction of polysemy's effect, raising the question of whether the internal relationship among frequency, polysemy, and change is itself stable across languages.

Korean provides an ideal test case for these questions. It is agglutinative, head-final, and belongs to a distinct language family, with a well-documented written history spanning over 500 years from the promulgation of the Korean script (*Hunminjeongeum*) in 1446. Crucially, Korean has undergone extensive grammaticalization over this period (Rhee, 2019), a process that preferentially targets high-frequency items (Bybee, 2003) and may interact with the frequency–change relationship in ways not visible in shorter observation windows.

In this paper, we train diachronic Word2Vec models on Korean historical corpora from the 15th to 20th centuries, align them using Orthogonal Procrustes, and measure the correlation between word frequency and semantic shift. We find that the law of conformity does not hold in Korean: high-frequency words undergo significantly *more* semantic change ($r = +0.494$, $\rho = +0.527$). This pattern holds across controls for morphological artifacts, corpus size imbalance, polysemy, light verb

*Corresponding author.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 13–23
July 4, 2026 ©2026 Association for Computational Linguistics

compounds, genre, and period-specific effects, and is further validated by a direct comparison with English using Hamilton et al.'s own pre-trained embeddings (Appendix A). At century resolution, frequency is the independent driver and polysemy plays no role once frequency is controlled, but a controlled comparison across time resolutions and corpus types reveals that this null mediation is partly an artifact of coarse temporal grain and genre heterogeneity. We connect this reversal to Bybee (2003)'s observation that frequency exerts both a conserving force (entrenchment of stored representations) and a reducing force (phonological reduction, semantic bleaching, grammaticalization): in Korean, the reducing force dominates, yielding the positive correlation.

Our contributions are as follows.

1. We provide the first large-scale test of the law of conformity on a non-Indo-European, agglutinative language over a 500-year span.
2. We document a robust positive frequency–change correlation in Korean, reversing the predicted pattern.
3. We show that the role of polysemy in mediating frequency–change depends on time resolution and corpus homogeneity, reconciling apparently contradictory findings across analytical conditions.
4. We identify the dominance of frequency-driven reducing processes (semantic bleaching, grammaticalization, domain shift) over conserving processes as the mechanism underlying the Korean pattern.
5. We provide a direct English–Korean comparison using the same analytical pipeline, confirming that the reversal reflects a genuine cross-linguistic difference.

2 Related Work

2.1 Diachronic Word Embeddings

The computational study of semantic change has grown rapidly since the introduction of neural word embedding methods. Kim et al. (2014) trained neural language models on yearly snapshots of text to track word meaning trajectories, and Kulkarni et al. (2015) proposed statistical tests for detecting significant change using Word2Vec embeddings aligned across time. Building on this line of work, Hamilton et al. (2016) introduced Orthogonal Procrustes alignment (Schönemann, 1966) and proposed two statistical laws: the *law of conformity* (frequent words change less) and the *law of innovation* (polysemous words change more).

These findings have been both refined and challenged. Dubossarsky et al. (2017) showed that some observed correlations may be artifacts of the embedding training process, particularly for smaller corpora. More recent studies have employed contextualized embeddings (Hu et al., 2019; Giulianelli et al., 2020) and expanded analysis to languages beyond English, including German, Latin, and Swedish (Schlechtweg et al., 2020). Chen et al. (2023) introduced ChiWUG, the first graph-based evaluation dataset for Chinese semantic change detection, extending computational analysis to a non-Indo-European language. Despite this progress, large-scale tests of Hamilton et al.'s proposed laws on typologically distant languages remain scarce.

2.2 Frequency, Polysemy, and Semantic Change

The relationship between word frequency and meaning change has a long history in linguistics. Bybee (2003) argued that frequency exerts two opposing forces on linguistic items: a *conserving* force, whereby high token frequency entrenches stored representations and resists change, and a *reducing* force, whereby repeated use drives phonological reduction, semantic bleaching, and grammaticalization. Which force dominates may depend on the type of change, the language, and the time scale of observation. The law of conformity captures only the conserving side of this picture; the reducing side predicts the opposite (i.e., high-frequency words should change *more*), precisely the pattern we observe in Korean.

A subtle pattern within Hamilton et al. (2016)'s own analysis is that frequency and polysemy are highly positively correlated, yet after controlling for frequency, polysemy retains a significant positive effect on semantic change. This raises the question of whether the joint structure of frequency, polysemy, and change is itself stable across languages.

2.3 Grammaticalization and Language Change

Grammaticalization, the process by which lexical items develop grammatical functions over time (Kouteva et al., 2019), is one prominent instantiation of Bybee's reducing force: high-frequency items undergo semantic bleaching, category shift,

and phonological reduction as they grammaticalize (Bybee, 2003; Traugott and Dasher, 2001). However, grammaticalization is not the only frequency-driven reductive process; domain shift, meaning extension, and subjectification (Traugott, 1995) also target high-frequency items without necessarily producing grammatical elements.

Korean is known to have undergone extensive grammaticalization over its recorded history (Rhee, 2019): the verb *ha.ta* ('to do') developed from a full lexical verb into a productive light verb, and numerous spatial nouns grammaticalized into postpositions (Lee and Ramsey, 2011). Whether such productive reducing processes can override the conserving force behind the law of conformity has not been tested computationally; our study addresses this question.

2.4 Korean Historical Corpora

Korean has a continuous written tradition dating from the invention of the Hangul script in 1443, but historical texts require extensive orthographic normalization due to archaic jamo and variant spellings (Lee and Ramsey, 2011). To our best knowledge, computational work on Korean diachronic semantics remains limited to qualitative case studies; ours is the first large-scale quantitative analysis using diachronic embeddings.

3 Data

3.1 Corpus

We use the *eon* corpus, a diachronic corpus of Korean texts spanning six centuries (15th–20th), compiled from digitized historical Korean sources originally encoded in UTF-16LE with XML/SGML markup.[1] Texts were grouped by century to form six sub-corpora. Table 1 summarizes the corpus statistics and Table 2 provides the genre breakdown.

The corpus composition shifts from Buddhist and Confucian texts (15–17C) through royal and missionary materials (18–19C) to newspapers and novels (20C); details are in Table 2. The 20th-century sub-corpus is notably dominated by *sinsosel*[2] (60% of texts); we address genre and corpus size imbalances in our robustness analyses.

[1]The corpus was provided by a scholar of classical Korean literature (*hanmunhak*). We gratefully acknowledge their contribution.

[2]*Sinsosel* (신소설, lit. "new novel") is a transitional genre of early modern Korean fiction (ca. 1900–1920) bridging classical and modern Korean prose.

Table 1: Corpus statistics by century (after preprocessing and normalization, as used for model training).

Century	Files	Sentences	Tokens	Vocab
15th	61	29,858	487,967	9,864
16th	61	15,342	245,466	6,536
17th	58	20,384	458,690	12,028
18th	95	22,800	504,160	12,282
19th	65	76,677	1,450,944	31,781
20th	87	157,900	2,322,462	49,885
Total	427	322,961	5,469,689	—

Table 2: Genre distribution by century. Values indicate number of source texts per genre.

Genre	15C	16C	17C	18C	19C	20C
Buddhist	18	8	2	2	–	–
Confucian	3	23	12	15	4	–
Medical	1	7	6	–	2	–
Royal/court	15	2	3	22	6	–
Dict./reference	–	3	4	8	4	16
Translation	–	3	1	4	2	–
Christian	–	–	–	–	8	13
Diary/letter	1	3	9	1	1	1
Literary/poetry	9	–	9	8	–	–
Newspaper	–	–	–	–	–	1
New novel	–	–	–	–	–	52
Other	14	12	12	35	38	4
Total	61	61	58	95	65	87

3.2 Preprocessing

Historical Korean texts require extensive normalization due to orthographic variation across centuries. Our preprocessing pipeline includes Unicode normalization of PUA characters, archaic jamo[3] mapping to modern equivalents, markup removal, whitespace tokenization at the *eojeol*[4] level, and filtering of short tokens (<2 characters) and short lines (<3 tokens).

4 Methods

4.1 Word2Vec Training

We trained separate Skip-gram Word2Vec models (Mikolov et al., 2013) for each century using the Gensim library (Řehůřek and Sojka, 2010), with the following hyperparameters: embedding dimension = 100, context window = 5, minimum word count = 5, training epochs = 10. We used 100 dimensions given the smaller corpus size (cf. 300 in

[3]*Jamo* (자모) are the consonant and vowel letters that compose Hangul syllable blocks. Pre-modern texts contain archaic jamo (e.g., ㆍ *arae-a*) no longer used in modern Korean.

[4]An *eojeol* (어절) is the orthographic word unit delimited by whitespace in Korean; it typically contains a content stem followed by one or more bound morphemes (particles, suffixes, endings).

Hamilton et al. (2016)). All models were trained with a fixed random seed for reproducibility.

4.2 Alignment via Orthogonal Procrustes

We aligned all century models to a reference century (17th, chosen as the chronological midpoint with sufficient data) using Orthogonal Procrustes alignment (Schönemann, 1966; Hamilton et al., 2016), computing the optimal rotation $R^* = \arg\min_R \|X_c R - X_{\text{ref}}\|_F$ s.t. $R^T R = I$ over shared vocabulary words.

4.3 Semantic Shift Measurement

For each word present in at least 4 centuries and with an average frequency ≥ 5, we measured semantic shift as:

$$\text{shift}(w) = 1 - \cos(\mathbf{v}_w^{\text{first}}, \mathbf{v}_w^{\text{last}}) \quad (1)$$

where $\mathbf{v}_w^{\text{first}}$ and $\mathbf{v}_w^{\text{last}}$ are the aligned vectors of word w in its earliest and latest attested century, respectively.[5]

4.4 Frequency and Polysemy Measurement

Word frequency was computed as the average raw token count across all centuries in which the word appeared. Following Hamilton et al. (2016), we used $\log_{10}(\text{frequency})$ in all correlation analyses.

We estimated word polysemy using neighbor *dispersion* (the average pairwise cosine distance among the top 20 nearest neighbors), averaged across all attested centuries. This captures contextual diversity similarly to, though distinct from, Hamilton et al. (2016)'s local clustering coefficient: both measures quantify how diverse a word's usage contexts are, but neighbor dispersion operates directly in embedding space rather than on a co-occurrence network.

5 Results

5.1 Reversal of the Law of Conformity

Figure 1 shows the relationship between log frequency and semantic shift for 2,378 words present in at least 4 centuries. We observe a strong *positive* correlation: Pearson $r = +0.494$ ($p < 10^{-146}$), Spearman $\rho = +0.527$ ($p < 10^{-170}$).

This is the *opposite* direction from the law of conformity, which predicts a negative correlation. Our own English replication confirms the expected negative pattern ($r = -0.263$ for the full 1800–1990 span; Appendix A). In Korean, high-frequency words undergo significantly *more* semantic change over 500 years.

[5]We also computed Hamilton et al.'s displacement rate measure, $\Delta^{(t)}(w) = 1 - \cos(\mathbf{v}_w^{(t)}, \mathbf{v}_w^{(t+1)})$, with pooled regression across all word–period observations; results are reported in Appendix A.

Table 3: Pearson r between log frequency and semantic shift for adjacent century pairs.

Pair	r	p	n
15C–16C	$+0.537$	4.4×10^{-148}	1,974
16C–17C	$+0.228$	3.0×10^{-23}	1,850
17C–18C	$+0.330$	1.5×10^{-52}	2,024
18C–19C	$+0.403$	1.0×10^{-65}	1,658
19C–20C	$+0.445$	9.2×10^{-76}	1,537

5.2 Robustness Check 1: Period-Specific Consistency

To verify that the pattern is not driven by the long time span, we computed the frequency–shift correlation for each pair of adjacent centuries. As shown in Table 3, the correlation is positive and significant for all five pairs, ranging from $r = +0.228$ (16C–17C) to $r = +0.537$ (15C–16C).

5.3 Robustness Check 2: Morphological Control

Korean is agglutinative, so each eojeol may contain a stem plus multiple suffixes. We tested three frequency counting methods to check for tokenization artifacts: original eojeol frequency ($r = +0.494$), approximate stems ($r = +0.441$), and unique stem types ($r = +0.485$). All yield positive correlations ($p < 10^{-100}$).

5.4 Robustness Check 3: Corpus Size Equalization

The 20th-century sub-corpus (2.3M tokens) is nearly 10 times larger than the 16th-century one (245K tokens). We downsampled all centuries to 245,466 tokens, retrained and realigned all models: $r = +0.491$ ($p < 10^{-78}$, $n = 1{,}276$), a change of only -0.003, ruling out the corpus-size artifacts noted by Dubossarsky et al. (2017).

5.5 Robustness Check 4: Disentangling Frequency and Polysemy

Hamilton et al. (2016) proposed that polysemous words change faster (their law of innovation), and noted that frequency and polysemy are highly correlated in their English data; yet after mutual control, both variables retain independent effects

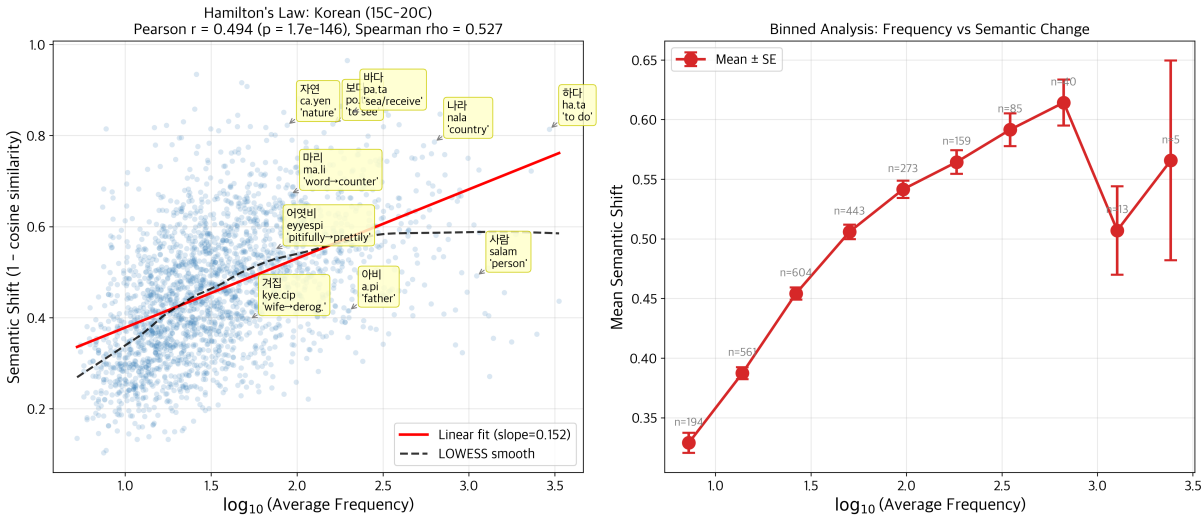


Figure 1: Log frequency vs. semantic shift for 2,378 Korean words (15C–20C). The positive correlation ($r = +0.494$) is the opposite direction from English ($r = -0.263$ in our replication; Appendix A).

Table 4: Pairwise and partial correlations among frequency, shift, and polysemy ($n = 2{,}378$).

Pairwise	r	p
Freq ↔ Shift	$+0.494$	$< 10^{-146}$
Freq ↔ Polysemy	$+0.788$	$< 10^{-300}$
Polysemy ↔ Shift	$+0.394$	$< 10^{-89}$
Partial	r	p
Freq → Shift | Poly	$+0.325$	$< 10^{-59}$
Poly → Shift | Freq	$+0.008$	0.708

(frequency negative, polysemy positive). If high-frequency Korean words are simply more polysemous, the observed pattern could likewise be a polysemy artifact.

We computed pairwise and partial correlations (Table 4). First, controlling for polysemy *reduces* but does not eliminate the frequency–shift correlation (r: $+0.494 \rightarrow +0.325$). Second, controlling for frequency *eliminates* the polysemy–shift correlation entirely ($r = +0.008$, n.s.). This pattern contrasts with Hamilton et al. (2016)'s English data, where both frequency (negative) and polysemy (positive) retain independent effects after mutual control. In Korean, frequency is the only variable that survives mutual control, and its effect is positive.

The pattern persists across polysemy tertiles: low ($r = +0.386$), mid ($r = +0.381$), and high ($r = +0.268$), all significant at $p < 10^{-15}$. Multicollinearity is not a concern (VIF = 2.29), and a permutation test (10,000 iterations) places the partial r 25 SDs above the null ($p < 10^{-4}$).

5.6 Robustness Check 5: Light Verb Exclusion

Korean has a highly productive light verb construction "X-*hata*" (하다 *ha.ta*, 'to do'), which combines with Sino-Korean nouns to form verbs (e.g., 사랑하다 *salang-ha.ta* 'to love', 공부하다 *kongpwu-ha.ta* 'to study'). The number of unique X-*hata* types expanded 22-fold from the 15th to the 20th century.

Excluding all words containing *hata*-related morphology *increases* the correlation slightly ($r = +0.504$, $n = 2{,}295$), confirming that the pattern is not driven by light verb constructions.

5.7 Robustness Check 6: Genre Control

Our corpus spans multiple genres whose distribution shifts across centuries (Table 2).

First, we computed per-genre correlations by restricting the analysis to words attested primarily within each genre. The correlation is positive and significant in all 11 genres, ranging from $r = +0.302$ (Newspaper) to $r = +0.605$ (Confucian) (Table 5).

Second, a linear mixed-effects model with log frequency as a fixed effect and genre as a random intercept confirms that frequency independently predicts semantic shift ($\beta = +0.109$, $z = 16.15$, $p = 1.09 \times 10^{-58}$). The genre random intercept variance is 0.002, indicating that genre contributes negligibly to baseline shift levels.

Third, we trained a separate Word2Vec pipeline exclusively on Confucian texts (15C–18C), which span four centuries with consistent representation.

Table 5: Per-genre frequency–shift correlations. All correlations are positive and significant.

Genre	r	p	n
Confucian	$+0.605$	2.0×10^{-11}	101
Translation	$+0.541$	7.9×10^{-4}	35
Diary/letter	$+0.495$	1.0×10^{-8}	119
Literary/poetry	$+0.386$	6.9×10^{-6}	128
Other	$+0.383$	1.1×10^{-5}	125
Dict./reference	$+0.377$	6.1×10^{-6}	136
Buddhist	$+0.351$	2.0×10^{-12}	378
New novel	$+0.344$	1.8×10^{-4}	114
Christian	$+0.337$	5.4×10^{-10}	323
Medical	$+0.309$	1.9×10^{-2}	57
Newspaper	$+0.302$	1.3×10^{-4}	155

Table 6: Mediation by time resolution and corpus type. Med. = proportion of total effect mediated by polysemy (Baron–Kenny).

Corpus	Resol.	Span	n	r	Med.
Literary	century	500 yr	2,378	$+0.494$	1.7%
Literary	decade	30 yr	1,448	$+0.323$	19.5%
News	decade	40 yr	40,785	$+0.547$	90.4%

This single-genre analysis yields $r = +0.733$ ($n = 388$), substantially stronger than the full-corpus result over the same time window ($r = +0.495$, $n = 1{,}749$). The mixing of genres in the full corpus thus dilutes rather than inflates the correlation.

5.8 Time Resolution and Mediation

The near-zero polysemy mediation at century resolution (§5.5) could reflect coarse temporal grain rather than a genuine absence of the polysemy pathway. To test this, we conducted a controlled comparison across three analytical configurations: (i) the literary corpus at century resolution (500 years, 15C–20C), (ii) the same literary corpus at decade resolution (30 years, 1880–1910), and (iii) a homogeneous newspaper corpus at decade resolution (40 years, 1920–1960).

Two factors jointly determine mediation strength (Table 6). First, *time resolution*: moving from century to decade resolution on the same literary corpus increases mediation from 1.7% to 19.5%. Second, *corpus homogeneity*: at the same decade resolution, the single-genre newspaper corpus yields 90.4% mediation versus 19.5% for the multi-genre literary corpus (which mixes religious texts, dictionaries, and fiction from 1880–1910). Genre heterogeneity injects noise into the polysemy proxy (neighbor dispersion), attenuating the mediated pathway.

The positive sign is stable across all three configurations ($r = +0.323$ to $+0.547$); what varies is how much of this effect routes through polysemy, a function of analytical resolution rather than the underlying linguistic phenomenon.

5.9 Contextualized Embedding Validation

To test whether the positive correlation is visible to contextualized models, we applied KLUE-BERT (Park et al., 2021) under two configurations. **Frozen BERT**: We extracted contextual embeddings for 1,493 target words (of the original 2,378; the remainder lacked sufficient attestation in BERT-tokenizable contexts) using up to 50 contexts per word per century and measured semantic shift via prototype distance (PRT) and average pairwise distance change (APD). Neither metric shows a significant correlation: PRT $r = -0.010$ ($p = 0.703$), APD $r = +0.032$ ($p = 0.222$). **Continual pretraining**: We further pretrained KLUE-BERT on each century's subcorpus via masked language modeling (3 epochs for 15–18C, 2 epochs for 19–20C). Results remain null: PRT $r = +0.011$ ($p = 0.669$), APD $r = +0.038$ ($p = 0.141$).

The null results are expected: the changes we detect (bleaching, domain shift, light verb development) operate at the *type level*, which static embeddings capture by design but BERT's token-level representations do not. Domain mismatch with pre-modern orthography further limits BERT's sensitivity. Two readings remain possible: (i) the reversal is genuinely type-level—a shift in distributional signature that static embeddings capture but BERT's token-level representations do not—or (ii) the Word2Vec signal partly reflects aligned-embedding artifacts. We favor (i) since corpus-size equalization (§5.4) and Hamilton-style replication (Appendix A) both reproduce the positive sign; the scope of our claim is therefore type-level lexical change.

5.10 Qualitative Case Studies

To illustrate the nature of the detected semantic changes, we examine the diachronic nearest neighbors of selected words (Table 7). The five words below are illustrative rather than exhaustive; the systematic, large-scale evidence for the reversal comes from the 2,378-word correlation analysis (§5.1), the 22-fold expansion of X-*hata* light verb constructions (§5.6), and the per-century shift heatmap (Figure 2).

Table 7: Diachronic nearest neighbors for selected words. Shift = $1 - \cos(\mathbf{v}^{\text{first}}, \mathbf{v}^{\text{last}})$. Yale romanization is provided for all Korean forms.

Word	Gloss	Shift	Span	Earliest neighbors	Latest neighbors
하다 *ha.ta*	'to do'	0.814	15–20C	일 도 *il.to* 'work-also', 그 러 타 *kule.tha* 'be so' (content verb)	것 *kes* 'thing', 엄중하게 *emcwung-ha.key* 'strictly' (light verb)
보다 *po.ta*	'to see'	0.825	15–20C	비취샨 *pichhwi.syan* 'shone upon', 상각하야 *sangkak.ha.ya* 'thinking' (visual)	깔보다 *kkal.po.ta* 'look down on', 살피다 *sal.phi.ta* 'examine' (evaluative)
바로 *pa.lo*	'straight'	0.965	15–20C	씩씩하샤미 *siksik.ha.sya.mi* 'being vigorous,' 너브시고 *nepu.si.ko* 'being broad' (physical)	지 체 *ci.chey* 'delay', 숨 기 지 *swumki.ci* 'not hiding' (abstract)
마리 *ma.li*	'word'	0.671	15–20C	달오다 *tal.wo.ta* 'to say', 듣자오미 *tut.ca.wo.mi* 'hearing' (speech)	송 아 지 *song.a.ci* 'calf', 곰 *kom* 'bear' (animal counter)
비록 *pi.lok*	'although'	0.323	15–20C	이시나 *i.si.na* 'but', 그러나 *ku.le.na* 'however' (concessive)	그러나 *ku.le.na* 'however', 할지언 뎡 *hal.ci.en.tyeng* 'even if' (concessive)

Grammaticalization The verb 하다 *ha.ta* ('to do') shows a clear trajectory from content verb to light verb (Rhee, 2019): from 일 *il* ('work') in the 15th century to the nominalizer 것 *kes* ('thing') in the 20th, reflecting its modern role in Sino-Korean constructions (e.g., 공부하다 *kongpwu-ha.ta* 'to study'). Similarly, 바로 *pa.lo* shifts from physical attributes to abstract/discourse functions, a textbook case of semantic bleaching (Bybee, 2003).

Semantic domain shift The noun 마 리 *ma.li* undergoes a complete domain shift: from speech-related neighbors in the 15th century (consistent with its original meaning of 'word/speech') to animal-related neighbors in the 20th, reflecting its modern use as a counting unit for animals.

Stability of function words In contrast, 비록 *pi.lok* ('although') maintains consistent concessive neighbors across centuries (그러나 *ku.le.na* 'however'), yielding a low shift score (0.323).

The qualitative picture matches the quantitative one: high-frequency content words undergo substantial change through diverse reductive processes, while already-grammaticalized function words remain stable. Figure 2 provides a fine-grained view of per-century-pair shift magnitudes.

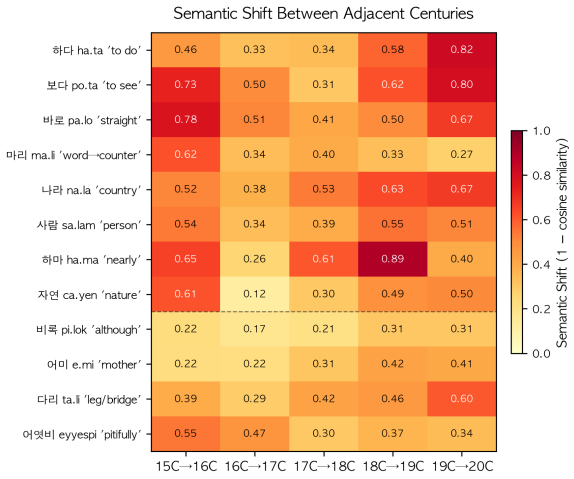


Figure 2: Pairwise semantic shift ($1 - \cos$) between adjacent centuries for selected words. Darker cells indicate greater shift. The dashed line separates high-shift content words (above) from more stable items (below).

6 Discussion

Our central finding is that the law of conformity appears not to hold for Korean: the positive correlation ($r = +0.494$) is robust across all six controls (§5.2–5.7), including corpus size equalization that partly addresses the embedding artifact concerns of Dubossarsky et al. (2017) ($r = +0.491$ vs. $+0.494$). We frame the linguistic interpretations below as compatible with our results rather than as causally established by them.

6.1 The Frequency–Polysemy–Change Triad

Our partial correlation results speak directly to the joint structure of frequency, polysemy, and semantic change identified by Hamilton et al. (2016). In their English data, frequency and polysemy are highly correlated, yet both retain independent effects after mutual control: frequency negative, polysemy positive—the latter opposite to what the positive frequency–polysemy correlation would predict. At century resolution, Korean shows a different structure: frequency retains an independent positive effect ($r = +0.325$ after controlling for polysemy), while polysemy's effect vanishes ($r = +0.008$, n.s.; VIF = 2.29, permutation test 25

SDs above null).

However, the time resolution experiment (§5.8) reveals that this near-zero mediation is partly an artifact of analytical conditions: at decade resolution mediation rises to 19.5%, and in a homogeneous newspaper corpus to 90.4%. Two implications follow. (i) Claims about *whether* polysemy mediates frequency–change are not separable from the conditions under which they are tested; century-level bins and genre mixing both attenuate the polysemy proxy, which may also explain divergent conclusions across English studies. (ii) The *positive* sign of the correlation is stable across all configurations ($r = +0.323$ to $+0.547$); what varies is the causal pathway, not the direction. Consistent with Bybee's dual-force account (§2), the reducing force appears to dominate in Korean.

6.2 Frequency-Driven Reductive Processes

Our case studies (§5.10) instantiate this pattern through grammaticalization, semantic bleaching, and domain shift, while already-grammaticalized items remain stable. These are frequency-driven reductive processes (Bybee, 2003; Rhee, 2019); the 22-fold expansion of X-*hata* (§5.6) provides population-level evidence that this pathway is productive across the corpus, not just in selected examples.

The 19C–20C pair ($r = +0.445$) spans roughly the same time depth as our English replication ($r = -0.668$ for 1850–1990; Appendix A), yet shows the opposite sign—ruling out observation window as the cause and pointing to typological differences (§6.3). The heatmap also shows shifts intensifying in the 18C–19C and 19C–20C transitions, coinciding with Korean modernization and Western-language contact, suggesting interaction with external sociolinguistic pressures.

6.3 Cross-linguistic Implications

The law of conformity was derived from English (1800–2000) and later replicated on Indo-European languages. Our English replication using HistWords (Appendix A; $r = -0.263$ to -0.776) confirms the negative pattern, while the Korean data yield a positive correlation of comparable magnitude. The frequency–change relationship is thus conditioned by both *language typology* and *time scale*.

On Bybee's dual-force account, every language exhibits both a conserving and a reducing force; the observed sign reflects which dominates. Agglutinative morphology may amplify the reducing force in two ways: (i) the productive X-*hata* construction and the grammaticalization of high-frequency verbs and spatial nouns into bound morphemes (§5.6; Rhee, 2019) create recurring constructional slots into which high-frequency items are repeatedly recruited, shifting their distributional neighborhood; (ii) eojeol-level contexts are dominated by bound-morpheme neighbors, so even modest functional re-analysis substantially changes a word's top neighbors. The reversal is thus a prediction of which force dominates given a language's morphosyntactic resources, not a denial of the conserving force.

Together with Xu and Kemp (2015)'s finding that regularities vary across word classes and Chen et al. (2023)'s work on Chinese, our results suggest that cross-linguistic variation in frequency–change dynamics may be more common than previously assumed. Whether other agglutinative languages (e.g., Turkish, Japanese, Finnish) show similar reversals remains an open question.

7 Conclusion

We have presented the first large-scale computational study of the frequency–semantic change relationship in Korean, spanning 500 years of recorded history (15th–20th centuries). Our central finding is that high-frequency Korean words appear to change *more*, not less, a pattern that reverses the law of conformity and is robust across six controls and validated against an English replication. The role of polysemy in mediating the frequency–change relationship is not absent but depends on analytical conditions: time resolution and corpus homogeneity jointly shape how much of the effect routes through polysemy, while the positive sign of the correlation remains stable. We interpret the reversal as consistent with frequency-driven reductive processes, including grammaticalization, semantic bleaching, and domain shift, that are especially productive in agglutinative Korean (Bybee, 2003; Rhee, 2019). The frequency–change relationship thus appears not to be a fixed regularity but to vary with language typology and analytical conditions.[6]

Limitations

Our study has several limitations. First, without a reliable morphological analyzer for historical Ko-

[6]Code and trained models will be made available upon publication.

rean, we tokenize at the eojeol (spacing unit) level, though our lemmatization check suggests this does not drive the main finding. Second, our BERT validation (§5.9) yielded null results, which we attribute to type-level vs. token-level sensitivity and domain mismatch; future work with historically pretrained language models may reveal more nuanced patterns. Embedding quality may also be lower for the smaller sub-corpora (15C: 488K tokens, 16C: 245K tokens), though our corpus-size equalization directly addresses this concern ($r = +0.491$ after downsampling all centuries to 245K tokens); an independent intrinsic evaluation of embedding quality on historical Korean is left to future work. Third, replication with independent historical Korean corpora would strengthen the findings. Fourth, we used average frequency across all attested centuries rather than lagged frequency at timepoint $t-1$; while our adjacent-century pair analysis (Table 3) partially addresses this by showing the positive correlation holds across all five pairs, a fully lagged design would be more appropriate for causal interpretation. Fifth, our neighbor-dispersion approach to polysemy provides only an approximate proxy for true lexicographic polysemy, reported here as a centuries-averaged score, and our time resolution experiment shows that mediation estimates are sensitive to temporal grain and corpus homogeneity. Sixth, the newspaper corpus used in the decade-resolution comparison (1920–1960) is limited in temporal range and may not be representative of all registers of modern Korean.

Acknowledgments

We thank the two anonymous reviewers for their thoughtful and constructive feedback, which substantially improved this paper. We thank the anonymous scholar of classical Korean literature (*hanmunhak*) for generously providing the historical Korean corpus used in this study. We also acknowledge the use of Claude Opus 4.6 (Anthropic) for assistance with manuscript revision and data visualization.

References

Joan Bybee. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 602–623. Blackwell Publishing, Oxford.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 130–138, Singapore. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD. Association for Computational Linguistics.

Tania Kouteva, Bernd Heine, Bo Hong, Haiping Long, Heiko Narrog, and Seongha Rhee. 2019. *World Lexicon of Grammaticalization*, 2nd edition. Cambridge University Press, Cambridge.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 625–635, Florence, Italy. ACM.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1384–1397, Santa Fe, NM. Association for Computational Linguistics.

Ki-Moon Lee and S. Robert Ramsey. 2011. *A History of the Korean Language*. Cambridge University Press, Cambridge. Originally published in Korean as *Kugosa Kaeseol* (Thaehaksa, 1972; revised editions through 2015).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*, pages 3111–3119. Curran Associates, Inc.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, and 18 others. 2021. KLUE: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Seongha Rhee. 2019. Grammaticalization in Korean. In Jaehoon Yeon and Lucien Brown, editors, *The Cambridge Handbook of Korean Linguistics*, pages 507–530. Cambridge University Press.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 1–23, Barcelona, Spain (Online). Association for Computational Linguistics.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Elizabeth Closs Traugott. 1995. Subjectification in grammaticalization. *Subjectivity and subjectivisation: Linguistic perspectives*, 1:31–54.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge University Press, Cambridge.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2703–2708.

Table 8: Frequency–shift correlations: Korean vs. English.

Window	**Language**	r	n
Full span	Korean (15C–20C)	$+0.494$	2,378
Full span	English (1800–1990)	-0.263	99,963
$\sim$150 yr	Korean (19C–20C)	$+0.445$	1,537
$\sim$150 yr	English (1850–1990)	-0.668	99,963
$\sim$100 yr	Korean (19C–20C)	$+0.445$	1,537
$\sim$100 yr	English (1900–1990)	-0.776	99,963

A English Replication with HistWords

To enable direct comparison, we replicated the frequency–shift correlation analysis on English using the pre-trained HistWords embeddings released by Hamilton et al. (2016).[7] These are decade-level Skip-gram embeddings (300d, 100K vocabulary per decade) trained on the Google Books English Fiction corpus. We aligned all decades to the 1900s reference via Orthogonal Procrustes, computed semantic shift as $1 - \cos(\mathbf{v}^{\text{first}}, \mathbf{v}^{\text{last}})$, and used vocabulary rank as a frequency proxy (following the Zipfian rank–frequency relationship).

Table 8 summarizes the results alongside the Korean findings from the main paper. The English data consistently show a *negative* correlation across all analysis windows, confirming the law of conformity for English, while the Korean data show a *positive* correlation of comparable magnitude.

The sign reversal is consistent across every adjacent time-period pair: all five Korean century pairs yield positive correlations ($r = +0.228$ to $+0.537$), while all 19 English decade pairs yield negative correlations ($r = -0.322$ to -0.822).

Hamilton-style pooled regression. To ensure that the reversal is not an artifact of our shift measurement (first–last cosine distance), we also replicated Hamilton et al.'s exact procedure: computing displacement rates between adjacent time periods, $\Delta^{(t)}(w) = 1 - \cos(\mathbf{v}_w^{(t)}, \mathbf{v}_w^{(t+1)})$, pooling all word–period observations, and regressing displacement on log frequency. Table 9 reports the results.

Our English replication yields $\beta_f = -0.263$, closely matching Hamilton et al.'s reported $\beta_f = -0.27$ and validating our pipeline. Korean yields $\beta_f = +0.435$ (literary, century resolution) to $+0.491$ (news, decade resolution), confirming that the sign reversal holds under the identical measure-

[7] SGNS embeddings from Google Books Fiction (1800–1990), available at https://nlp.stanford.edu/projects/histwords/.

Table 9: Hamilton-style pooled displacement rate regression ($\Delta^{(t)}(w) \sim \beta_f \cdot \log f(w)$). Our English β_f closely replicates Hamilton et al.'s reported value; Korean β_f is positive across all configurations.

Corpus	β_f	**Pooled** r	n
Hamilton (2016) Eng. SGNS	-0.27	—	—
English Gutenberg (ours)	-0.263	-0.263	633,157
Korean literary (century)	$+0.435$	$+0.435$	34,610
Korean literary (decade)	$+0.426$	$+0.426$	20,119
Korean news (decade)	$+0.491$	$+0.491$	467,370

ment procedure. These results rule out the possibility that the reversal is driven by differences in how semantic shift is operationalized.

Narrative Landscape: Mapping Narrative Dispositions Across LLMs

Donghoon Jung **Jiwoo Choi*** **Songeun Chae*** **Seohyon Jung†**
School of Digital Humanities and Computational Social Sciences, KAIST, South Korea
{donghoon.jung, jwchoi0515, songeun, seohyon.jung}@kaist.ac.kr

Abstract

This study proposes a quantitative framework for profiling LLM dispositions as stable, model-specific regularities in output under repeated, controlled elicitation. Using a structured narrative constraint-selection task administered across six frontier models and three instruction types, we operationalize disposition through two dimensions: "consistency", measured as cross-replication selection overlap via Jaccard similarity, and "diversity", measured as dispersion across options via the inverse Simpson index. We further introduce Narrative Landscape, a PCA-based visualization that maps each model's selection profile into a shared space for direct comparison. Results reveal a clear rigidity–exploration spectrum across model families and show that instruction types shift the geometry of selection spaces even when scalar metrics appear similar, indicating that comparable scores can mask qualitatively distinct selection topologies.

1 Introduction

During the development of Claude Opus 4.5, researchers reportedly used an internal "soul document" to specify the model's persona, values, and interaction style (Weiss, 2025). Anthropic (2024) characterizes this broader practice as character training, which aims to instill more nuanced traits such as curiosity and open-mindedness into the model's responses through design choices beyond standard instruction following. As large language models (LLMs) are now widely used across domains including narrative writing, scientific writing, and ideation, their engineered dispositions become consequential for the kinds of outputs they systematically produce. This study argues for the need to identify such dispositions as empirically observable regularities.

Recent empirical work uses replication-based elicitation to probe disposition stability across personality inventories (Serapio-García et al., 2023), political preference, and moral robustness and susceptibility under persona role-play (Rozado, 2024; Costa et al., 2025), and large-scale comparisons of personality versus political profiles (Goyanes et al., 2025), alongside measurement refinements via open-ended, AI-rated Big Five assessment (Zheng et al., 2025). Yet such profiles remain fragile with respect to test–retest reliability and paraphrase variation (Wang et al., 2025) and exhibit limited behavioral coherence, including word–deed inconsistency and weak cross-aspect transfer (Xu et al., 2025); moreover, they can be stabilized at the level of self-report without commensurate behavioral change (Han et al., 2025). Despite these measurement advances, prior work lacks a shared-space visualization that exposes model- and instruction-level structure in repeated selections.

This study introduces a replication-based framework for profiling LLM disposition as model-specific regularities in narrative constraint selection under controlled elicitation (Jung et al., 2025). With a fixed pool of narrative constraints and repeated replications, overlap in selected constraints captures the stability of commitments, operationalized as consistency via Jaccard similarity (Broder, 1997), while dispersion of selections across the pool captures coverage of the decision space, operationalized as diversity via the inverse Simpson index (Hill, 1973). To complement these scalar summaries and expose differences in selection structure, we also embed constraint-frequency profiles into a shared Principal Component Analysis (PCA) space for direct comparison across models and instruction types.

Results show a clear spectrum of narrative dispositions across model families, from rigid, high-overlap constraint selections with limited option-space coverage to low-overlap selections with

*Co-second authors.
†Corresponding author.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 24–30
July 4, 2026 ©2026 Association for Computational Linguistics

markedly higher diversity; instruction types further modulate the extent of the selection space. Narrative Landscape also indicates that comparable scalar metrics can correspond to distinct selection topologies, representing model- and instruction-level differences as geometric structure in a shared constraint space. The measurement framework and Narrative Landscape generalize beyond narrative, providing a transferable approach to profiling disposition in structured selection tasks.

2 Data

Model Selection and Constraint Pool. Building on Jung et al. (2025), we analyze six state-of-the-art LLMs spanning major provider families (OpenAI, Anthropic, Google, Alibaba). To maximize replicability across repeated runs, decoding parameters were held constant where available; temperature (temp) and top-p were fixed at 1.0 for all models, and vendor-specific controls (reasoning effort, verbosity) were set to high when present. Model identifiers and decoding settings are summarized in Appendix A.

All models were presented with the same constraint pool, which functions as a theory-grounded diagnostic taxonomy for making narrative preferences observable as authorial choices. The pool contains 200 narrative constraints systematically distributed across four narratological elements—Event, Style, Character, Setting—with each element subdivided into five categories of ten constraints each. To minimize surface-level selection bias and ensure that observed differences reflect model dispositions rather than prompt artifacts, each constraint is designed with structural regularity (uniform word length, parallel grammatical structure) and matched conceptual granularity within categories. The models operate on the natural-language constraints themselves, allowing selection patterns to be interpreted as behavioral traces of a model's implicit narrative logic.

Experimental Procedure. Each observation in our dataset corresponds to a single run in which a model produces a narrative plan by selecting constraints and generating justifications under a controlled instruction type. We operationalize "authorial orientations" through three broad instruction types—Basic, Quality-focused, and Creativity-focused—intended to capture stance rather than dependence on specific phrasing. Across runs, the persona is set via the instruction, and the user prompt instructs the model to read the full constraint pool, select a fixed number of constraints, justify each selection, and then evaluate cross-constraint dynamics in a final compatibility assessment.

In this paper, we focus on a pooled, unlabeled constraint-selection setting with a fixed selection budget. Models are presented with the full pool of 200 constraints without element labels and are required to select exactly 20 constraints they deem most useful for planning a single fictional narrative, providing a brief justification for each selection. The full prompt templates and the complete constraint pool are documented in Jung et al. (2025).

To enable rigorous replication and to control for order effects (Liu et al., 2024; Pezeshkpour and Hruschka, 2024; Shi et al., 2025), each run uses (i) a fresh random permutation of the constraint list and (ii) an isolated session state. Within any fixed experimental cell (model × instruction type, in our setting), only the constraint-order permutation and provider stochasticity vary; all other factors are held constant. We execute 160 independent replications per cell (2,880 runs total), producing dense samples for stability and sensitivity analyses.

3 Consistency and Diversity

To measure the consistency of constraint selection across repeated runs within the same model–instruction type cell, we use Jaccard similarity (J) (Broder, 1997):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, A and B are the constraint sets selected in two runs. For each model–instruction type cell, we compute Jaccard similarity over all run pairs and report the mean pairwise value as the consistency score.

To measure diversity, we use the Gini–Simpson index (GS) (Simpson, 1949), which corresponds to the probability that two randomly selected constraints are different. Because the Gini–Simpson index is bounded in $[0, 1]$ and can compress differences, we also report the inverse Simpson index in its effective-number form (EN) (Hill, 1973):

$$GS = 1 - \sum_k p_k^2, \qquad EN = \frac{1}{\sum_k p_k^2}$$

In both expressions, p_k is the proportion of selections assigned to constraint k. A higher effective number indicates that the model draws from a broader set of constraints more evenly.

Table 1: Narrative consistency and diversity by model and instruction type

Model	Instruction Type	Jaccard Similarity	Effective Number	Unique Constraints
gpt5	Quality-focused	0.3169	41.7581	113
gpt5	Basic	0.3070	42.7947	123
gemini	Creativity-focused	0.2597	48.7336	142
gpt5	Creativity-focused	0.2546	49.5021	128
gpt4.1	Creativity-focused	0.2458	51.1520	188
o4mini	Quality-focused	0.2420	51.7731	143
o4mini	Basic	0.2232	55.2856	159
claude	Creativity-focused	0.2166	56.5602	184
claude	Quality-focused	0.2160	56.7037	185
gemini	Basic	0.2006	60.2034	174
o4mini	Creativity-focused	0.2004	60.3261	165
gemini	Quality-focused	0.1936	61.9600	174
claude	Basic	0.1860	64.2425	196
gpt4.1	Quality-focused	0.1483	77.0911	199
gpt4.1	Basic	0.1344	83.9162	200
qwen	Quality-focused	0.0896	120.1963	200
qwen	Creativity-focused	0.0867	123.9079	200
qwen	Basic	0.0820	129.9954	200

Table 1 shows the means of Jaccard similarity and the effective number across six models and three instruction types. The results highlight a distinct spectrum of narrative behaviors. gpt5 demonstrates the most rigid narrative disposition. In the Quality-focused instruction type, it exhibits the highest consistency ($J = 0.3169$) and the lowest diversity ($EN = 41.7581$), utilizing only 113 unique constraints. This indicates a strong adherence to a focused set of narrative strategies. In contrast, qwen displays the opposite extreme. Across model–instruction type cells, qwen shows the lowest consistency ($J < 0.09$) and the highest diversity ($EN > 120$), using the maximum observed number of unique constraints (200). Other models such as gemini, o4mini, claude, and gpt4.1 occupy an intermediate position, balancing stability and exploration. These divergences suggest a rigidity–exploration spectrum in which each model reflects a structurally distinct selection logic rather than a variation along a single, shared trait.

In particular, while instruction types, such as Creativity-focused versus Quality-focused, induce relatively small fluctuations within models, between-model differences are typically larger than instruction-induced differences in narrative profiles under our experimental setup. Although prompt-level differences are generally smaller than between-model differences, for gpt5 the Creativity-focused is notably distinct: its difference from Basic ($\Delta J = 0.0524$) and from Quality-focused ($\Delta J = 0.0623$) both exceed the Basic–Quality-focused difference ($\Delta J = 0.0099$), and even some between-model gaps such as that between o4mini and claude.

4 Narrative Landscape

Despite the fact that consistency and diversity differentiate models, these scalar summaries provide only a partial view of disposition. Moreover, prior work can establish statistically reliable differences across models, but it does not reveal how repeated selections are structured when aggregated across replications in a shared representational space; consequently, similar scalar scores can correspond to qualitatively distinct selection organizations. We therefore introduce Narrative Landscape, a shared-space representation designed to make such structural differences directly comparable across models and instruction types.

We operationalize Narrative Landscape by vectorizing each constraint by its selection frequencies across model–instruction type cells, aggregated across replications. We then apply PCA to embed these constraint profiles into a two-dimensional space, yielding a landscape for comparing the geometry of cell-specific selection structure beyond scalar reduction.

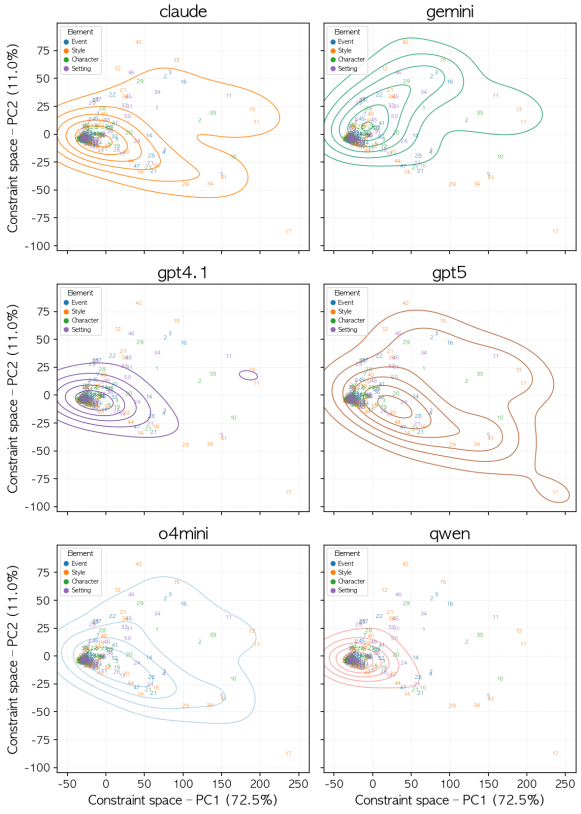


Figure 1: Narrative Landscape of six models under the Basic instruction type

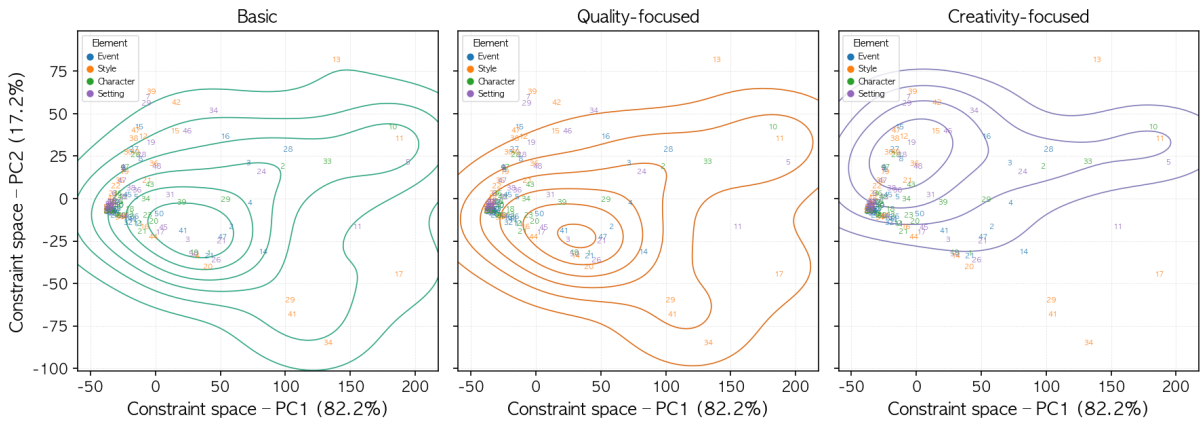


Figure 2: Narrative Landscape of three instruction types in gpt5

Figure 1 visualizes the projected Narrative Landscapes of six models under the Basic instruction type. By projecting the constraints onto the principal component space, the contour lines represent the smoothed density of model-weighted selection frequency, effectively illustrating the topography of each model's narrative disposition. The scattered labels represent the projected locations of individual constraints (Event, Style, Character, Setting), serving as landmarks within this abstract space.

This spatial analysis corroborates the results on consistency and diversity from the previous section while clarifying the structural nature of repeated selection. gpt5, which demonstrated the highest consistency ($J = 0.3070$), exhibits topological complexity, with irregular and stretched contours extending into specific regions of the constraint space, suggesting robust adherence to a particular narrative disposition and structural combinations that pull its profile away from the generic mean. In contrast, qwen, despite its high diversity ($EN > 120$), shows a simpler, more concentric topology centered near the origin, consistent with an averaging effect across replications in which broad, weakly structured selections average out in the aggregate. Narrative Landscape also reveals differences not recoverable from scalar metrics alone: gemini and claude have similar Jaccard similarity, yet their covered regions diverge, with claude extending further into the lower region and gemini covering more of the upper region. Thus, higher consistency corresponds to a more concentrated density, whereas extreme diversity yields a more diffuse distribution, and comparable scores can nonetheless reflect distinct selection geometries.

Figure 2 illustrates the Narrative Landscapes of gpt5 across three distinct instruction types. The comparison reveals that the specific contours and covered regions shift noticeably depending on the instruction. This visualizes how different instructions effectively alter the boundaries and focus of the model's Narrative Landscape, even within the same architecture. We further run an additional experiment on gpt5 with four contrasting additional instruction types. The four types, their metrics, and their Narrative Landscapes are reported in Appendix B–Appendix D. Notably, Optimistic vs. Pessimistic exhibits nearly matched consistency and diversity yet induces sharply different directional coverage in the landscape (see Appendix D).

5 Conclusion

This study (1) quantifies repeated constraint selection across two dimensions—consistency and diversity—and (2) introduces Narrative Landscape, a shared PCA space that makes model- and prompt-dependent selection structure directly comparable. Whereas prior work has primarily established statistically reliable differences in elicited behavior, our approach makes it possible to inspect how selections are organized across replications in a common space, revealing topological differences that scalar scores can obscure and extending naturally to other structured selection tasks.

As a next step, we plan to develop an interactive visualization tool that integrates significance testing and interpretable summaries with the landscape (e.g., hover-based inspection of points/regions and linked views for constraint-level effects). We also aim to formalize the correspondence between geometric signatures in the landscape (e.g., concentrated trajectories versus diffuse clusters) and underlying selection mechanisms, moving from de-

scriptive geometry to a more rigorous methodological account. Finally, we plan to apply this framework to the diverse selection-task datasets explored in prior work discussed in this paper, to assess transferability beyond narrative constraint selection.

Limitations

While our experiments cover six widely used proprietary models, a natural next step is to extend the analysis to a broader set of models (e.g., additional model families, sizes, and open-weight systems) to further assess the generality of the observed patterns. Similarly, although we consider multiple instruction types, exploring a wider prompt-framing space with more diverse and systematically varied prompts would help characterize how robust the profiles and landscape geometry are under richer elicitation conditions. The constraint pool, while carefully constructed through a theory-grounded approach, reflects specific narratological assumptions that may privilege certain storytelling conventions. Different theoretical frameworks and alternative sets or criteria of narrative constraints could yield different disposition profiles. Finally, we have so far evaluated the framework only in our narrative constraint-selection setting; applying the same methodology to other selection tasks remains an important direction for future work.

Acknowledgments

This work was supported by the Korea Advanced Institute of Science and Technology (KAIST) under the project "Quantifying Creativity: Developing Metrics for Evaluating AI-Generated Narratives" (Project No. 11250011 and No. N10260075).

References

Anthropic. 2024. Claude's character. Anthropic Research. Accessed: 2026-05-08.

A. Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE.

D. B. Costa, F. Alves, and R. Vicente. 2025. Moral susceptibility and robustness under persona role-play in large language models. *Preprint*, arXiv:2511.08565.

Manuel Goyanes, Adrián Domínguez-Díaz, and Luis de Marcos. 2025. Personality traits and political dispositions, ideology, and sentiment toward political leaders of 14 artificial intelligence large language models. *Human Behavior and Emerging Technologies*, 2025(1):5761832.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. In *NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models (ResponsibleFM)*.

M. O. Hill. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432.

Donghoon Jung, Jiwoo Choi, Songeun Chae, and Seohyon Jung. 2025. Style over story: Measuring LLM narrative preferences via structured selection. *Preprint*, arXiv:2510.02025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

David Rozado. 2024. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. Research Square preprint. Accessed: 2026-05-08.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

E. H. Simpson. 1949. Measurement of diversity. *Nature*, 163(4148):688–688.

Jiaqi Wang, Bo Wang, Fa Guo, Cheng Cheng, and Li Yang. 2025. A comparative study of large language models and human personality traits. *Preprint*, arXiv:2505.14845.

Richard Weiss. 2025. Claude 4.5 opus' soul document. LessWrong. Accessed: 2026-05-08.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. 2025. Large language models often say one thing and do another. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51(2):599–640.

A Model identifiers and decoding settings

Abbr.	Full Identifier / Release	Provider	Temp	Top-p	Reasoning Effort	Verbosity
o4mini	o4-mini-2025-04-16	OpenAI	1.0	1.0	high	—
gpt4.1	gpt-4.1-2025-04-14	OpenAI	1.0	1.0	—	—
gpt5	gpt-5-2025-08-07	OpenAI	1.0	1.0	high	high
claude	claude-opus-4-20250514	Anthropic	1.0	1.0	—	—
gemini	gemini-2.5-pro (2025-06-17)	Google	1.0	1.0	—	—
qwen	qwen-max-2025-01-25	Alibaba	1.0	1.0	—	—

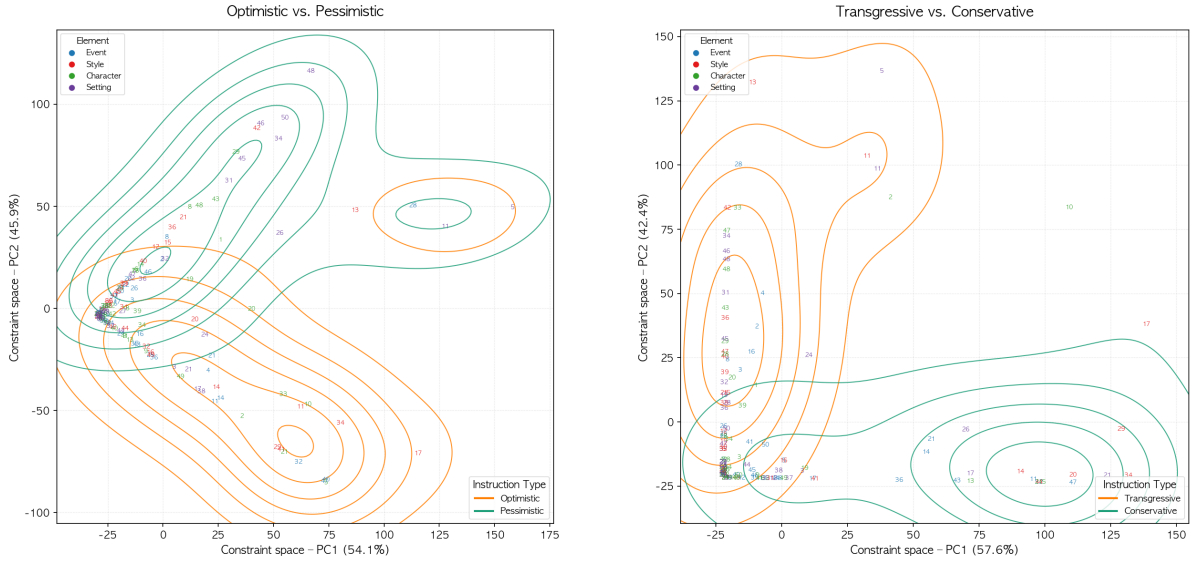
B Additional instruction types for gpt5

Instruction Type	Content
Optimistic	You are a hopeful writer who believes in human resilience and positive change. You write stories that explore challenges while emphasizing connection and possibility for growth. Your goal is to create narratives that leave readers uplifted, showing how people overcome difficulties or find redemption and satisfying success even under dire circumstances.
Pessimistic	You are a cynical writer who specializes in capturing harsh realities and systemic failures without sentimentality. You write stories that expose futility, moral compromise, and human limitations. Your goal is to create narratives that confront readers with difficult truths, showing how circumstances constrain people and good intentions can always fail or backfire.
Transgressive	You are a risk-taking writer unafraid to explore controversial subjects and morally complex situations with a progressive vision. You write stories that tackle difficult or uncomfortable situations with honesty and boldness. Your goal is to create narratives that push boundaries and challenge readers' assumptions without offering easy resolutions or sanitized outcomes.
Conservative	You are a safety-oriented writer who creates accessible narratives within comfortable thematic boundaries. You write stories that entertain the readers while maintaining broad appeal and widely-accepted moral frameworks. Your goal is to create narratives that provide satisfying experiences and unambiguous closure without venturing into disturbing, unstable, or divisive territory.

C Metrics for additional instruction types for gpt5

Instruction Type	Jaccard Similarity	Gini–Simpson	Effective Number	Unique Constraints
Conservative	0.4320	0.9700	33.3414	87
Optimistic	0.3372	0.9749	39.8887	109
Pessimistic	0.3239	0.9756	41.0342	110
Transgressive	0.3151	0.9761	41.9222	111

D Narrative Landscapes for additional instruction types for gpt5



100,000+ Movie Reviews from Kazakhstan: Russian, Kazakh, and Code-Switched Texts

Rustem Yeshpanov
Independent Researcher / Astana, Kazakhstan
yeshpanov.rustem@gmail.com

Abstract

We present a new publicly available corpus of 100,502 movie reviews from Kazakhstan collected from kino.kz, spanning 2001–2025 and covering 4,943 unique titles. The dataset is multilingual, consisting mainly of Russian reviews alongside Kazakh and code-switched texts. Reviews are manually annotated for language and sentiment polarity, and 11,309 reviews additionally contain explicit user-provided ratings. We define two sentiment tasks—three-way polarity classification and five-class score classification—and benchmark classical BoW/TF–IDF baselines against multilingual transformer models (mBERT, XLM-RoBERTa, RemBERT). Experimental results show that transformer models consistently outperform classical baselines on polarity classification, while score classification remains challenging under leakage-controlled evaluation due to severe class imbalance and subtle distinctions between adjacent rating levels.

1 Introduction

Movie reviews are widely used in sentiment analysis because they contain naturally occurring, explicitly evaluative language and typically provide more context than short social media posts. However, publicly available datasets of movie reviews in Kazakh remain scarce, limiting reproducible research on sentiment modelling in this under-resourced language. In addition, Kazakhstan provides a practically important multilingual setting in which user-generated reviews are predominantly written in Russian, while Kazakh reviews and code-switching also occur.

We introduce a new publicly available corpus of 100,502 movie reviews collected from kino.kz, spanning 25 years (2001–2025) and covering 4,943 unique titles. The dataset includes Russian, Kazakh, and code-switched texts, and is manually annotated for review language and sentiment polarity. A subset of 11,309 reviews additionally contains explicit user-provided ratings, enabling fine-grained score prediction.

We define two supervised sentiment classification tasks: polarity classification with three labels (negative/neutral/positive) and score classification based on user ratings. We report benchmark results for classical BoW/TF–IDF baselines and multilingual transformer models, including per-language evaluation to characterise performance under data imbalance and code-switching. The dataset, accompanying documentation, and trained models are released to support future work on multilingual sentiment analysis and culturally grounded user-generated text in Kazakhstan and comparable contexts.

While sentiment classification is not the only possible use of this corpus, it provides a widely understood and reproducible probe task for characterising dataset difficulty and establishing baselines on Kazakhstan-specific review discourse. Beyond aggregate scores, our experiments surface two dataset-specific issues that are easy to miss in cleaner benchmarks: (i) neutral polarity is rare and often expressed implicitly, which leads to systematic confusions, and (ii) fine-grained score prediction is highly susceptible to label leakage because users frequently state ratings verbatim in the text, motivating leakage-controlled evaluation. These baselines therefore serve as reference points for future work on multilingual modelling, code-switching, and robust sentiment inference in real-world review text from Kazakhstan.

2 Related Work

Movie reviews are a longstanding benchmark for supervised sentiment analysis, dating back to early polarity-classification work on review corpora (Pang et al., 2002). For English, widely used resources such as the IMDb Large Movie Review

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 31–40
July 4, 2026 ©2026 Association for Computational Linguistics

Dataset (Maas et al., 2011; Maas, 2011) and the Stanford Sentiment Treebank (Socher et al., 2013) have enabled extensive comparison of both classical and neural approaches across binary and fine-grained sentiment settings.

For Russian, sentiment datasets exist, but fewer have become standard movie-review benchmarks. A closely related resource is the Kinopoisk movie review corpus (Blinov et al., 2013). Other widely used Russian benchmarks focus on different domains, such as social media (e.g., RuSentiment (Rogers et al., 2018), and therefore differ from long-form reviews in length, register, and discourse structure.

For Kazakh, publicly available sentiment resources remain comparatively limited. KazSAnDRA (Yeshpanov and Varol, 2024) provides a large-scale Kazakh review dataset (180,064 items) with 1–5 star ratings from four domains (mapping/navigation, e-commerce marketplace, online bookstore, and Android app store). The dataset reflects naturally occurring Kazakh online text, including Kazakh–Russian code-switching and mixed Cyrillic/Latin writing practices, and the accompanying baselines report competitive performance for polarity classification (F_1 = 0.81) and substantially lower performance for fine-grained score prediction (F_1 = 0.39).

Finally, code-switched sentiment analysis has been studied primarily in short-form social media via shared tasks such as SemEval SentiMix (Patwa et al., 2020). In contrast, our corpus targets long-form movie-review discourse from Kazakhstan and provides a Kazakhstan-specific multilingual setting; while code-switched reviews form a small subset, they still allow targeted evaluation on naturally occurring Kazakh–Russian mixed-language reviews, within the broader review corpus.

Taken together, prior datasets provide limited coverage for Kazakhstan-specific movie-review sentiment with long temporal span and naturally occurring multilingual (Russian/Kazakh) review text, motivating the dataset release and the use of sentiment baselines as a diagnostic benchmark in this work.

3 Dataset Development

3.1 Source Data

Movie reviews were collected from `kino.kz`[1], a major Kazakh online ticketing and entertainment portal launched in 2000, using BeautifulSoup[2]. The platform allows users to browse showtimes, view trailers, access film information, leave reviews, and purchase e-tickets for films, concerts, theatre performances, sports events, and other cultural activities via both its website and mobile applications (Android and iOS). After removing duplicates, the data collected comprised 100,567 reviews, including review text, review date and author, movie title in Russian/Kazakh and English, screening year, genre, director, duration, age restriction, and production country.

Production-country labels are available for 600 of 4,943 titles (12.1%). Among titles with known country labels, the most frequent countries (by number of unique titles) are the United States (182), Kazakhstan (110), the United Kingdom (68), Russia (58), and France (48). Kazakhstan is listed as a production country for 18.3% of titles with known labels. Kazakh-language reviews are more common for these Kazakhstan-produced titles: the median share of Kazakh reviews per title is 0.10, compared to 0.00 for all other titles (mean shares: 0.19 vs 0.009).

3.2 Review Language Identification and Annotation

Since the language of the extracted movie reviews was not provided, the author manually identified the language of each review. Unlike Yeshpanov and Varol (2024), where reviews containing Kazakh-Russian words or grammar were labelled as Kazakh, in this study we aimed to annotate more granularly, labelling reviews as `kk` for Kazakh, `ru` for Russian, `en` for English, `cs` for instances of code-switching, and `ot` for all other languages. While `en` and `ot` reviews were found, they were extremely rare (65 in total) and were therefore excluded from subsequent analyses. Code-switched reviews include two or more languages within a single text, most commonly Kazakh–Russian, occasionally involving English or other languages.

We distinguish code-switching from loanword usage: the `cs` label is applied when a review contains a multiword segment from another language (e.g., an inserted phrase or clause), typically including the function words or grammatical marking of that language (i.e., an extended span in the other language). In contrast, isolated conventional

[1] `https://kino.kz/`

[2] `https://www.crummy.com/software/BeautifulSoup`

borrowings that are integrated into the surrounding language are treated as loanwords and do not, by themselves, warrant `cs`. Consider the following Kazakh–Russian code-switched review:

> *Уақыт аз болмаса, тема фильм. Звуктарды жақсы пайдаланған. Сюжет жаксы, но қысқа. Барып көруге стоит.*
> *Uaqyt az bolmasa, tema fil'm. Zvuktardy zhaqsy paidalanğan. Syuzhet zhaksy, no qysqa. Baryp köruge stoit.*
> "If you have some time, the film is solid. The sounds are used well. The plot is good, but short. It is worth going to see it."

As Table 1 indicates, Russian-language reviews constitute the vast majority of the corpus, with smaller subsets of Kazakh and code-switched texts. By whitespace-delimited word count, Russian reviews have a median length of 30 words (95th percentile: 108), Kazakh reviews 24 (95th percentile: 65), and code-switched reviews 33 (95th percentile: 73). The table also shows a strong skew towards positive reviews, a pattern reported for many review platforms (Pang and Lee, 2008); neutral labels are comparatively rare.

Language	Negative	Neutral	Positive	Total
cs	201	62	851	**1,114**
kk	178	127	2,334	**2,639**
ru	26,936	4,140	65,673	**96,749**
Total	**27,315**	**4,329**	**68,858**	**100,502**

Table 1: Distribution of movie reviews by language and sentiment polarity

Moreover, although the platform allows users to rate movies with stars (from one to ten), these ratings are not publicly displayed, complicating the assignment of polarity scores (positive, neutral, negative). Accordingly, the author manually labelled reviews following guidelines specifically devised for this purpose.

In the absence of additional human annotators, we employed gpt-4.1-nano-2025-04-14 as a compensatory measure to support annotation reliability, which was considered a practical solution under the circumstances. The model was instructed as follows:

> *You are a sentiment classifier for movie reviews (Russian or Kazakh). Return one digit only: 2 – clearly positive / recommends the movie; 1 – neutral, mixed, or unclear; 0 – clearly negative / does not recommend. Always choose one digit; never output anything else.*

GPT-generated labels achieved 89.54% accuracy relative to the single-annotator labels over the full corpus, with substantial agreement (Cohen's $\kappa = 0.78$), indicating strong consistency beyond chance (Landis and Koch, 1977). We report these figures to quantify label stability under single-annotator constraints; the released dataset uses the human annotations as the primary labels.

Furthermore, when available, user ratings were extracted from reviews (e.g., 3 out of 10). For reviews where ratings were provided on a 1–5 scale, scores were multiplied by 2 to align with the standard 1–10 scale. In some cases, users explicitly indicated that a movie was so unsatisfactory that it deserved a score of 0, rather than the minimum 1; these instances were accordingly assigned a rating of 0. Consequently, the final rating scale spans from 0 to 10.

In a small number of cases, the rating format was ambiguous (e.g., a user stating a score of "3" without specifying the scale), which could correspond either to 3/10 or to 3/5 (i.e., 6/10 after normalisation). To resolve such cases, we manually inspected the surrounding review content and inferred the most plausible interpretation based on the expressed sentiment. While we applied this procedure consistently and aimed to minimise errors, a limited number of borderline instances may remain, and the extracted scores should therefore be treated as approximate in rare ambiguous cases.

Overall, 11,309 reviews (approximately 11% of the dataset) contained an explicit user-provided score (e.g., "10/10", "9 из 10" [*devyat' iz desyati*, "9 out of 10"], "твердая семерка" [*tvyordaya semyorka*, "a solid seven"]). Table 2 presents the distribution of explicit user-provided scores across review languages.

In addition, during the review inspection, several recurring themes were occasionally noted, such as unmet expectations, whether the movie was a one-time watch, movie sections perceived as unsatisfactory, and cases where the overall impression was negative but the movie was still recommended for niche audiences. While these observa-

Language	Score											Total
	0	1	2	3	4	5	6	7	8	9	10	
cs	10	2	3	3	2	7	2	8	2	5	106	150
kk	3	4	1	3	3	5	4	7	10	14	212	266
ru	357	317	365	568	520	559	1,077	699	1,244	930	4,257	10,893
Total	**370**	**323**	**369**	**574**	**525**	**571**	**1,083**	**714**	**1,256**	**949**	**4,575**	**11,309**

Table 2: Distribution of user-provided ratings (0–10) by review language

tions were recorded, they are not the focus of the present analysis.

The language identification and annotation process, carried out single-handedly, spanned 110 days, from August 2025 to January 2026.

3.3 Collected Data Significance

We argue that the collected movie reviews are of substantial value to the natural language processing (NLP) community for several reasons. First, the dataset spans a period of 25 years, with the earliest reviews dating back to 2001 and the most recent to 2025. Such long-term temporal coverage makes it possible to trace changes in audience preferences and attitudes towards social phenomena and issues (e.g., traditions, domestic violence) over time, ranging from initial denial or avoidance to increased openness and willingness to engage with these topics. Changes in the role and use of the Kazakh language are also clearly observable. In particular, during manual language annotation, we found that although the earliest Kazakh-language review is associated with a film released in 2002, review creation timestamps indicate that the first Kazakh review in our data was authored in 2011, approximately a decade after the launch of the platform (Figure 1).

This likely reflects the initial predominance of Russian-language reviews and the gradual adoption of Kazakh for user-generated content on the platform. Earlier reviews frequently contain criticism of the quality of Kazakh dubbing and translations, or even explicit requests for permission to express opinions in Kazakh (e.g., *можно я на казахском* "May I speak in Kazakh?"), whereas later reviews increasingly express positive attitudes towards Kazakh-language film production and show greater confidence in using Kazakh to articulate opinions. Notably, the five films with the highest numbers of reviews were all produced in Kazakhstan.

Second, the dataset comprises reviews of 4,943 unique movie titles authored by 31,453 publicly visible reviewer identifiers, reflecting a large and diverse pool of contributors. While many identifiers correspond to self-selected usernames, 6,273 reviews (approximately 19%) are associated with a generic, platform-assigned label (e.g., "Kino.kz user", Russian: "Пользователь kino.kz"), indicating anonymous or non-registered reviewers. Although such entries cannot be distinguished at the individual level, they constitute a substantial portion of the dataset and further contribute to its overall diversity. For release, reviewer identifiers are anonymised by replacing each unique user string with a stable pseudonymous identifier, preserving within-user consistency while removing direct identifiers; reviews associated with the platform-generic label remain indistinguishable, consistent with the source platform.

Third, although the dataset is dominated by Russian-language reviews, the variety of Russian observed is of particular relevance. Specifically, the reviews frequently employ features of Kazakhstani Russian, a regional variety shaped by sustained contact with Kazakh and by local sociocultural context. This includes references to culturally specific events, institutions, and named entities, as well as lexical items and expressions uncommon or opaque to speakers of Russian outside Kazakhstan. Examples include *ажека, агашка, бастык, токалка, болашаковцы, шапалак, уят, Наурыз, Бауржан Шоу, Sulpak, Керуен, Kcell, Otau Cinema*, referring to kinship terms, social roles, cultural concepts, holidays, media productions, and local organisations specific to the Kazakhstani context, as well as regionally marked constructions such as *чёп-чёрный* ("pitch black"), which illustrates calquing of Kazakh reduplicative intensification patterns into Russian; *не уятьте* ("do not shame [someone]"), an example of contact-induced verb formation combining a

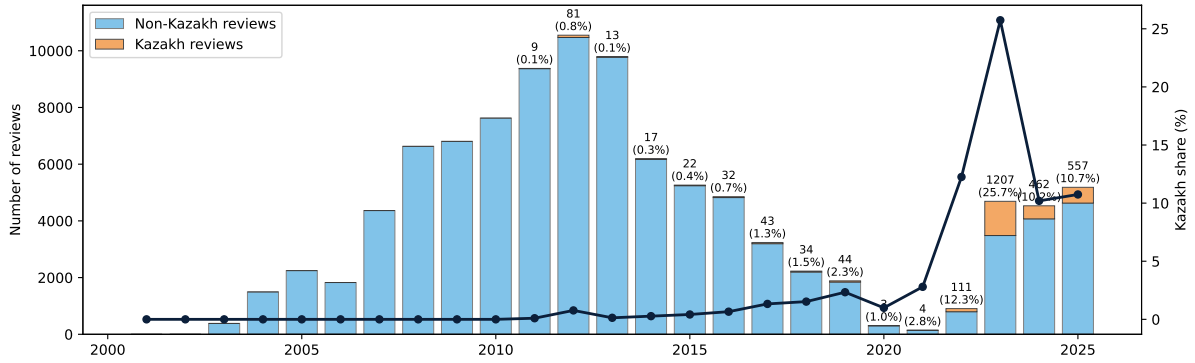


Figure 1: Kazakh-language review share over time

Kazakh lexical root with Russian negation and imperative morphology; and *ерекси* ("men"), formed using a Kazakh lexical root combined with a Russian plural inflection. Such phenomena make the data collected particularly valuable for studying regional language variation, code-switching, and culturally grounded named entity usage in real-world user-generated text.

3.4 Sentiment Classification Tasks

Following the design of prior work on Kazakh sentiment analysis, particularly KazSAnDRA, we formulate two primary sentiment classification tasks for our dataset. First, we define a polarity classification (PC) task, in which reviews are categorised into three broad sentiment categories: positive, neutral, and negative. Second, we consider a score classification (SC) task based on explicit user-provided ratings extracted from reviews.

During dataset construction, user ratings were normalised to a unified 0–10 scale. Accordingly, the initial formulation of the SC task involved predicting 11 discrete score labels (0–10). However, as shown in Table 2, the distribution of scores is highly imbalanced, with a substantial concentration of reviews assigned the maximum score and relatively few instances in lower score categories. Preliminary experiments with the 11-class setting resulted in unstable training and near-random macro-averaged F_1 scores, indicating that the fine-grained formulation is severely affected by data sparsity and long-tailed label distribution.

To obtain more reliable and statistically meaningful results, we therefore adopt a collapsed 5-class SC setting, where adjacent score ranges are grouped into broader ordinal bins (0–2, 3–4, 5–6, 7–8, 9–10). Furthermore, due to the highly imbalanced language distribution in the scored subset and the very limited number of Kazakh and code-switched reviews with explicit ratings, the SC task is restricted to Russian-language reviews.

3.5 Data Partitioning

The data for both tasks were divided into training (Train), validation (Valid), and testing (Test) splits in an 80/10/10 ratio. To reduce topical leakage, splitting was performed at the movie level, so that all reviews of a given film appear in exactly one split. Table 3 reports the distribution of reviews across splits by sentiment label and language for the PC task. Table 4 reports the distribution of Russian reviews across splits by score bin for the SC task.

Category	Train		Valid		Test	
	#	%	#	%	#	%
Sentiment						
Negative	21,767	27.21	2,890	26.69	2,658	27.50
Neutral	3,388	4.23	502	4.64	439	4.54
Positive	54,854	68.56	7,437	68.68	6,567	67.95
Language						
cs	809	1.01	196	1.81	109	1.13
kk	2,178	2.72	230	2.12	231	2.39
ru	77,022	96.27	10,403	96.07	9,324	96.48
Total	**80,009**	**100**	**10,829**	**100**	**9,664**	**100**

Table 3: Distribution of polarity labels and languages across the sets for PC

4 Experiment

4.1 Sentiment Classification Models

For the evaluation of sentiment classification tasks, we employed a set of multilingual transformer-based models that support both Kazakh and Rus-

Score	Train		Valid		Test	
	#	%	#	%	#	%
1	839	9.56	83	8.09	117	10.71
2	877	9.99	116	11.31	95	8.70
3	1,349	15.37	135	13.16	152	13.92
4	1,560	17.78	197	19.20	186	17.03
5	4,150	47.29	495	48.25	542	49.63
Total	**8,775**	**100**	**1,026**	**100**	**1,092**	**100**

Table 4: Scores across the sets for SC

sian and are readily available through the Hugging Face Transformers framework (Wolf et al., 2019).

mBERT[3] is a multilingual BERT model (Devlin et al., 2019) pre-trained on Wikipedia in 100+ languages, including Kazakh and Russian, with a shared WordPiece vocabulary (168M parameters).

XLM-RoBERTa[4] is a multilingual RoBERTa model (Conneau et al., 2020; Liu et al., 2019) pre-trained on CC-100 CommonCrawl data (100 languages; 270M parameters).

RemBERT[5] is a rebalanced multilingual BERT variant (Chung et al., 2021) trained on Wikipedia and web data in 110+ languages, designed to improve performance on underrepresented languages.

Evaluating instruction-tuned generative models and stronger multilingual encoders is a natural next step. Here we focus on widely used pre-trained multilingual transformers to provide a stable, reproducible supervised baseline that is less sensitive to prompting and instruction-tuning choices.

In addition to transformer-based models, we evaluated several classical machine learning baselines, including linear support vector machine (SVM), logistic regression (LR), and multinomial naïve Bayes (MNB), using bag-of-words (BoW) and Term Frequency–Inverse Document Frequency (TF–IDF) feature representations, which remain strong and widely adopted baselines for text classification tasks (Salton and Buckley, 1988; Joachims, 1999).

4.2 Experimental Setup

4.2.1 Transformer-based models

All three transformer models were fine-tuned separately for the PC and SC tasks using the corresponding training splits, while hyperparameters were selected based on performance on the respective validation sets. The final model configurations yielding the best validation macro-averaged F_1 were subsequently evaluated on the held-out test sets.

Fine-tuning was conducted on `Vast.AI`[6] using a single NVIDIA RTX 3090 GPU (24 GB VRAM). The total computational cost of all fine-tuning runs was approximately $2. The initial learning rate was set to 2×10^{-5} and the weight decay to 0.01. Training was terminated early when the validation loss showed a consistent increase across epochs, even in cases where the macro-averaged F_1 exhibited marginal fluctuations, in order to mitigate potential overfitting and promote stable generalisation.

Across fine-tuning runs, we used a maximum of three epochs with early stopping based on validation loss. For PC, mBERT and RemBERT were trained for two epochs, while XLM-RoBERTa was trained for three epochs. For SC, mBERT and XLM-RoBERTa were trained for three epochs, and RemBERT for one epoch. We used a batch size of 150 for mBERT and XLM-RoBERTa on both tasks. Due to GPU memory constraints, RemBERT was fine-tuned with smaller batch sizes (20 for PC and 16 for SC).

4.2.2 Classical Baselines

For all classical models, text was represented using either sparse BoW or TF–IDF features with a maximum vocabulary size of 50,000 and n-gram ranges determined via validation-based hyperparameter tuning.

For the PC task, the best-performing configuration relied on BoW features with 1–3-gram representations and lowercasing enabled. Linear SVM was trained with $C = 0.01$ and squared hinge loss, LR with $C = 2.0$, ℓ_2 regularisation, and the `saga` solver, and MNB with $\alpha = 0.01$ and disabled prior fitting. Class imbalance was addressed using class weights for SVM and LR.

For the SC task, different optimal configurations were observed. Linear SVM performed best with TF–IDF features and 1–3-gram representations ($C = 0.1$), whereas LR and MNB achieved stronger results with BoW features using 1–2-gram and 1–3-gram representations, respectively. The optimal LR setup used $C = 0.1$ with ℓ_2 regularisation and the `saga` solver, while MNB was configured with $\alpha = 0.5$ and enabled prior fitting. As in the PC task, class weights were applied to discriminative models to mitigate class imbalance.

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

[4] https://huggingface.co/FacebookAI/xlm-roberta-base

[5] https://huggingface.co/google/rembert

[6] https://vast.ai/

Model	Valid					Test				
	A	**P**	**R**	**F_1**	κ	**A**	**P**	**R**	**F_1**	κ
Polarity classification										
LR	0.88	0.70	0.71	0.70	0.75	0.89	0.70	0.72	0.71	0.76
MNB	0.87	0.68	0.70	0.69	0.73	0.88	0.68	0.72	0.70	0.74
SVM	0.90	0.71	0.71	0.71	0.77	0.90	0.73	0.74	0.73	0.79
mBERT	0.92	0.78	0.70	0.72	0.82	0.93	0.81	0.72	0.74	0.83
RemBERT	0.94	0.82	0.79	0.81	0.87	0.95	0.84	0.81	0.82	0.88
XLM-RoBERTa	0.94	0.80	0.78	0.79	0.86	0.94	0.81	0.81	0.81	0.87
Score classification										
LR	0.63	0.51	0.52	0.52	0.47	0.65	0.54	0.54	0.54	0.50
MNB	0.67	0.55	0.54	0.54	0.52	0.68	0.55	0.54	0.54	0.52
SVM	0.66	0.54	0.54	0.53	0.51	0.69	0.56	0.55	0.55	0.53
mBERT	0.64	0.52	0.52	0.50	0.48	0.67	0.53	0.51	0.51	0.50
RemBERT	0.67	0.55	0.54	0.52	0.53	0.69	0.56	0.55	0.54	0.55
XLM-RoBERTa	0.66	0.51	0.55	0.52	0.52	0.68	0.55	0.56	0.54	0.54

Table 5: Performance of classical and transformer-based models on the PC and SC tasks

All hyperparameters were selected via grid search on the validation set, with macro-averaged F_1 used as the primary optimisation criterion due to class imbalance.

4.2.3 Score Masking

For the SC task, explicit rating expressions often appear in the review text and may cause label leakage. We therefore replaced all explicit score mentions (identified via manual inspection of the scored subset, focusing on direct numeric and conventional rating expressions) with a placeholder token (`scoretoken`). The placeholder is alphanumeric for BoW/TF–IDF compatibility; it was also added to transformer tokenisers (with resized embeddings) to prevent subword splitting.

4.3 Sequence Length

To select an appropriate maximum input length for transformer models, we examined the token length distribution of the training, validation, and test splits after tokenisation with the respective model tokenisers (mBERT, XLM-RoBERTa, and RemBERT). The analysis showed that approximately 95–97% of reviews contain fewer than 256 tokens, depending on the tokeniser used. Only a small fraction of instances (about 2–5%) exceed this length and are consequently truncated when a maximum sequence length of 256 is applied.

Considering the relatively low proportion of longer reviews and the quadratic computational complexity of self-attention with respect to sequence length, we set the maximum input length to 256 tokens for all transformer-based models.

4.4 Performance Metrics

We evaluate model performance using accuracy (A), precision (P), recall (R), macro-averaged F_1-score (F_1), and Cohen's kappa (κ). Given the imbalanced class distributions in both PC and SC tasks, we treat macro-averaged F_1-score as the primary evaluation metric, as it assigns equal importance to all classes and offers a more balanced assessment than accuracy alone (Jurafsky and Martin, 2009; Yang, 2001). In addition, Cohen's κ is reported to measure the agreement between model predictions and gold labels while accounting for chance agreement.

4.5 Experiment Results

We report results for all models in Table 5. In general, transformer-based encoders still achieve the highest scores on PC, but the gap between them and classical methods narrows after masking explicit score mentions in the SC task. Since RemBERT performs consistently well across both PC and SC, we use it as the main reference system for per-class, per-language, and qualitative error analyses below.

4.5.1 Polarity Classification

On the PC task, the transformer models continue to set the standard. The test set macro-averaged F_1 scores for RemBERT and XLM-RoBERTa are 0.82 and 0.81, respectively. mBERT lags slightly behind with an F_1 of 0.74. Among the classical baselines, SVM and LR attain F_1 scores of 0.73 and 0.71, respectively, while MNB reaches 0.70. These results indicate that, although simple BoW

models remain strong baselines, contextualised encoders offer a consistent advantage for this three-way classification task.

Per-class analysis for RemBERT in Table 6 shows uniformly high precision and recall for positive and negative reviews (F_1 = 0.97 and 0.94), but markedly lower scores for the neutral class (F_1 = 0.56). The difficulty of detecting neutral sentiment is unsurprising given that neutral instances comprise only about 4–5% of the corpus and often contain ambiguous language. The per-language breakdown in Table 7 also confirms that performance correlates with data volume: Russian reviews dominate the dataset and yield the highest F_1 (0.81, κ = 0.88), whereas Kazakh (0.77, κ = 0.68) and code-switched (0.76, κ = 0.73) reviews achieve slightly lower but still respectable results.

Sentiment	P	R	F_1
Negative	0.94	0.94	0.94
Neutral	0.60	0.52	0.56
Positive	0.97	0.98	0.97

Table 6: Per-class results of RemBERT on the PC test set

Language	A	P	R	F_1	κ
cs	0.87	0.87	0.74	0.76	0.73
kk	0.89	0.89	0.73	0.77	0.68
ru	0.94	0.82	0.80	0.81	0.88

Table 7: Per-language results of RemBERT on the PC test set

Qualitative inspection of model errors suggests that most confusions involve borderline or implicitly neutral statements. For instance, the Russian review "*Неожиданный фильм для меня. Совершенно не похож на воспоминания из детства об индийском кино. Куда делись танцы?!!*" [*Neozhidannyy fil'm dlya menya. Sovershenno ne pokhozh na vospominaniya iz detstva ob indiyskom kino. Kuda delis' tantsy?!!*; "The film was unexpected for me. It is completely unlike my childhood memories of Indian cinema. Where did the dances go?!"] was labelled `Neutral` but predicted as `Negative`. Here, the author primarily expresses surprise and a mismatch with genre expectations (the absence of song-and-dance elements typical of Indian cinema), which can be interpreted as either mild criticism or a neutral observation; the model appears to treat the rhetorical question as negative sentiment.

A complementary error appears in a Kazakh review that combines obligation framing with a balanced appraisal: "*Озимиздин Казахтар тусиргесин колдап баруга тура келеди, бирак отиниш режиссёр сценаристы кишкене карау керек ау, актерлер оте жаксы ойнап шыкты.*" [*Ozimizding qazaqtar tüsirgesin qoldap baruğa tura keledi, biraq ötinish, rezhissior scenaristi kishkentai qarau kerek au, aktiorler öte zhaqsy oinap shyqty*; "We have to go and support films made by our Kazakhs, but the director and screenwriter really should improve a bit; the actors performed very well."]. Although the gold label is `Neutral` due to the combination of critique and praise, the model predicts `Positive`, likely overweighting explicit positive cues such as *"өте жақсы"* ("very good") while under-weighting the qualifying criticism. These cases illustrate that neutral reviews in this corpus often encode evaluation indirectly via expectations, rhetorical framing, or mixed praise and critique, making them particularly prone to polarity drift in automatic classification.

4.5.2 Score Classification

The SC task becomes considerably more challenging under the leakage-controlled setting with masked score mentions. Across all models, macro-averaged F_1 values fall into the 0.50–0.55 range, suggesting that predicting 5-level ratings from text alone is a much harder problem than PC. RemBERT and SVM achieve the highest test-set F_1 of 0.54–0.55, followed closely by XLM-RoBERTa and MNB (0.54). mBERT is the weakest among the transformers with an F_1 of 0.51. The similar performance of classical and transformer models indicates that the masking procedure effectively removed many of the lexical cues that deep models previously exploited, forcing them to rely on more subtle sentiment features in the text.

Per-score analysis in Table 8 highlights the extreme class imbalance. RemBERT performs well on the highest rating (score 5) with P = 0.87, R = 0.89, and F_1 = 0.88, but struggles on intermediate ratings; the F_1 for score 2, for example, is just 0.09 due to very few examples. This skew explains why overall accuracy is relatively high (69%) while macro-averaged metrics remain modest.

5 Discussion

The experiments provide two main takeaways. First, multilingual transformer encoders offer a

Score	P	R	F_1
1	0.72	0.65	0.68
2	0.24	0.05	0.09
3	0.44	0.67	0.53
4	0.53	0.51	0.52
5	0.87	0.89	0.88

Table 8: Per-class results of RemBERT on the SC test set

consistent advantage for PC on this corpus, but the margin over strong linear baselines is modest. The gap is most visible in the minority neutral class, where language is often ambiguous and underrepresented, while positive and negative sentiment are detected reliably by all models. This suggests that much of the polarity signal in movie reviews can be captured by surface lexical cues, yet contextual modelling remains beneficial for borderline cases and for improving robustness under class imbalance.

Second, leakage-controlled SC remains substantially more difficult than PC. After masking explicit rating mentions with `scoretoken`, performance converges across model families and drops to modest macro-averaged F_1 values, indicating that fine-grained rating inference depends on subtle and often implicit cues that are harder to learn than coarse polarity. The per-class behaviour further shows that the models perform well on the most frequent high-score bin but struggle on intermediate bins, reflecting both severe label imbalance and the inherently ordinal nature of ratings, where adjacent categories may be expressed with very similar language.

These findings highlight several limitations and directions for future work. The score distribution is strongly skewed toward favourable ratings, and even after collapsing to five bins, the mid-range classes remain sparse. More data in the lower and mid ranges, targeted rebalancing, or modelling approaches that explicitly account for ordinality (e.g., ordinal regression or regression-based formulations) may yield more stable improvements. In addition, fully eliminating score leakage is nontrivial because rating information can be expressed indirectly or idiomatically, and manual score extraction occasionally involves ambiguous cases; consequently, some residual noise in the score labels is likely. Finally, while the polarity results on Kazakh and code-switched reviews are encouraging, stronger conclusions about multilingual generalisation will require more balanced language coverage or dedicated evaluation subsets.

6 Conclusion

We introduced a new publicly available corpus of 100,502 movie reviews from Kazakhstan collected from `kino.kz`, spanning 2001–2025 and covering 4,943 titles, with Russian, Kazakh, and code-switched texts. Reviews were manually annotated for language and sentiment polarity, and 11,309 reviews additionally contain explicit user-provided ratings, enabling fine-grained sentiment modelling.

We defined two supervised tasks: three-way polarity classification and five-class score classification. On polarity classification, multilingual transformer encoders achieved the best results, with RemBERT performing strongest (macro-averaged $F_1 = 0.82$, $\kappa = 0.88$ on the test set). For score classification, we evaluated a leakage-controlled setting by masking explicit score mentions; under this setup, all models achieved modest macro-averaged F_1 scores (0.51–0.55), highlighting the difficulty of inferring rating levels from text alone under severe class imbalance.

To support reproducibility, we release the dataset, annotation guidelines, and trained models on the Hugging Face Hub.[7]

References

PD Blinov, Maria Klekovkina, Eugeny Kotelnikov, and Oleg Pestov. 2013. Research of lexical approach and machine learning methods for sentiment analysis. *Computational Linguistics and Intellectual Technologies*, 2(12):48–58.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

[7] https://huggingface.co/datasets/yeshpanovrustem/100k_movie_reviews_from_kz

Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Thorsten Joachims. 1999. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Andrew Maas. 2011. Large Movie Review Dataset. `https://ai.stanford.edu/~amaas/data/sentiment/`. Accessed 2026-02-26.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790. Association for Computational Linguistics.

Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 755–763. Association for Computational Linguistics.

Gerard Salton and Chris Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.*, 24:513–523.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1631–1642. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2019. Transformers: State-of-the-Art Natural Language Processing. In *Conference on Empirical Methods in Natural Language Processing*.

Yiming Yang. 2001. A Study on Thresholding Strategies for Text Categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–145.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9657–9667. ELRA Language Resource Association.

Quantifying Text Reuse Across Three Kṛṣṇa Yajurveda Recensions Using Multi-Algorithm Computational Collation

So Miyagawa[1], Kyoko Amano[2], Yuzuki Tsukagoshi[3], Yuki Kyogoku[4]

[1]Institute of Humanities and Social Sciences, University of Tsukuba, Japan,
[2]Graduate School of Letters, Kyoto University, Japan,
[3]Graduate School of Humanities and Sociology, The University of Tokyo, Japan,
[4]Institute for South and Central Asian Studies, Leipzig University, Germany

Correspondence: miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

The Kṛṣṇa Yajurveda survives in multiple recensions that share substantial ritual content, yet the degree and distribution of textual overlap across recensions have never been quantified systematically. This paper presents a computational analysis of text reuse across three recensions—the Maitrāyaṇī Saṃhitā (MS), the Kāṭhaka Saṃhitā (KS), and the Taittirīya Saṃhitā (TS)—for two ritual sections (Agnyupasthāna and Punarādhāna), using ICoMa (Intertextuality Collation Machine), a new web-based multi-algorithm collation tool. Five independent similarity algorithms consistently rank MS–KS as the most closely related pair, corroborating the philological consensus. Crucially, the two ritual sections exhibit strikingly different reuse profiles: Punarādhāna shows near-identical MS–KS overlap (up to 93.5%) with sharp divergence from TS, while Agnyupasthāna displays moderate, broadly distributed similarity across all three pairs. These contrasting patterns provide quantitative evidence that different ritual categories followed distinct paths of textual transmission within the Yajurvedic tradition. Thus, a way has been opened to explain the differences in similarity as reflecting differences in the periods at which the texts became fixed. ICoMa and the experimental data are freely available.

1 Introduction: The Kṛṣṇa Yajurveda and Its Recensions

The Vedas of ancient India constitute a vast corpus of literature centered on the religious rituals of their time. The oldest of these texts is the Ṛgveda, generally dated to the 15th to 12th centuries BCE. Following the Ṛgveda and the Atharvaveda, which consists mainly of hymns to the gods, the first prose texts emerged, focusing on the explanation of ritual actions. These are the Maitrāyaṇī Saṃhitā (MS), the Kāṭhaka-Saṃhitā (KS), and the Taittirīya-Saṃhitā (TS), which belong to what is known as Kṛṣṇa Yajurveda. A text referred to as Kapiṣṭhala-Kaṭha-Saṃhitā is also known, but its content largely coincides with that of the KS. These works are the most important sources for understanding the earliest form of Vedic ritual. The three recensions are the canonical texts of distinct schools (*śākhās*), each associated with a particular brahminical lineage and geographic region. The schools maintained their recensions through strict oral transmission designed to preserve the text including its sound shape.

The contents of the three texts are structured in parallel and share the same topics, namely various Vedic rituals. It has long been assumed that they derived from a common source, branching into an MS–KS group on the one hand and TS on the other, with MS regarded the oldest, followed by KS and then TS (Witzel, 1997).

However, Amano's studies of the MS (Amano, 2014–2015, 2020) have demonstrated that these texts were gradually compiled over a period of approximately one to two centuries. During this time, the relationships among the three schools changed, making it impossible to explain their formation by a simple branching model or by a straightforward relative chronology. The date of composition and the inter-school relationships differ from section to section; thus, inter-school relationships on a section-by-section basis is crucial for clarifying the formation of these texts and furthermore the development of Vedic rituals.

To address this issue, we have undertaken the detection of intertextual similarity. In Miyagawa et al. (2024), we analyzed intertextual similarity between MS and KS in three ritual sections using Word2Vec, the R package `stylo`, and TRACER. The results show that similarity varies by section and suggest that the degree of similarity may provide insights into the inter-school relationships and relative chronology at the time of composition.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 41–49
July 4, 2026 ©2026 Association for Computational Linguistics

2 Aims of the Present Comparison

In this paper, we examine intertextual similarity among the three texts MS, KS, and TS. By extending the analysis beyond a two-text comparison, we expect to gain further insight into the social and historical circumstances of the period. Previous scholarship has long recognized the close relationship between MS and KS, and more recent studies suggest that according to younger sections, KS and TS may in fact show a close relationship, while MS and TS do not show such a relationship (Amano, 2019, 2020). We therefore analyze all three pairwise combinations—MS–KS, MS–TS, and KS–TS.

However, the aim of this study is not merely to measure degrees of similarity, but to investigate how patterns of similarity differ from chapter to chapter. For this purpose, we selected two ritual sections for analysis: Agnyupasthāna (the daily worship of the ritual fires) and Punarādhāna (the re-establishment of the ritual fires). The Punarādhāna sections of MS and KS are known to exhibit particularly strong similarity compared to other sections. It has even been suggested that, rather than representing ordinary parallel transmission, they may reflect a direct borrowing relationship (Amano, 2014–2015). Whether this distinctive relationship can be detected through computational analysis constitutes a major test case for our method.

3 ICoMa: The Analysis Tool

ICoMa (Intertextuality Collation Machine) is a browser-based text reuse detection tool implemented in TypeScript and React. It was designed to address a methodological concern central to this study: different similarity algorithms capture different aspects of textual overlap, and relying on a single metric risks producing method-dependent conclusions. By integrating multiple algorithms in a single platform, ICoMa enables direct comparison of their outputs on identical data, providing a robustness check that has been absent from previous computational studies of Vedic intertextuality.

ICoMa currently supports ten script systems (Coptic, Greek, Latin, Devanāgarī, Syriac, Arabic, Hebrew, Ethiopic, CJK), with automatic script detection and script-appropriate tokenization. The tool is freely available as a web application.

3.1 Tokenization and Word Boundary Handling

ICoMa employs script-dependent tokenization. For alphabetic and abjad scripts (including romanized Sanskrit), word-level tokenization uses Unicode character properties ensuring correct handling of combining diacritics. For CJK scripts, character-level tokenization is applied.

The Sanskrit source texts contain phonological changes known as sandhi, in which, for example, a word-final vowel merges with the initial vowel of the following word. In the plain text, these words remain in their sandhi-merged surface form. Since the present study conducts a character-level analysis, we assume that sandhi does not interfere with the validity of the results and therefore use the plain text with sandhi retained. Word-level segmentation is realized in the lemmatized text, where forms are restored to their dictionary headword entries.

Potential issues concerning word boundaries arise primarily in two cases: whether nominal compounds should be split, and whether verbal prefixes should be separated from their verbs. Nominal compounds were separated as far as possible, while verbal prefixes were generally kept together with the verb. Since such decisions inevitably involve semantic interpretation, it is not possible to apply a single mechanical principle in every case. However, manual adjustments were made to ensure that no differences in segmentation policy arise among the three texts.

3.2 Similarity Algorithms

ICoMa implements six similarity algorithms. All sliding-window algorithms operate on windows of n consecutive tokens (set to $n = 3$ in this study).

Levenshtein distance (Levenshtein, 1966): normalized edit distance between windowed token sequences, yielding a percentage similarity score sensitive to substitutions, insertions, and deletions at the character level.

Character n-gram: token windows are concatenated into a single character string before computing Levenshtein distance. This is designed to mitigate word boundary inconsistencies by operating on continuous character sequences. Under the current parameters ($n = 3$), however, this algorithm produces results identical to the standard Levenshtein distance (see Section 4.4).

Word n-gram: Jaccard coefficient over normalized token sets within sliding windows, capturing

lexical overlap independently of word order.

Jaccard index (Jaccard, 1912): character-level Jaccard similarity over windowed text segments, measuring the ratio of shared to total unique characters.

Smith–Waterman alignment (Smith and Waterman, 1981): local sequence alignment at the token level with fuzzy matching (threshold: 80% character similarity), identifying the single best local alignment between two token sequences.

Script-aware normalization: script-specific preprocessing (Coptic supralinear stroke reduction, Greek polytonic accent normalization, Devanāgarī accent mark and sandhi stripping, and general diacritic normalization) followed by Levenshtein distance. This method is labeled "Coptic-Aware (Specific)" in the ICoMa interface, reflecting its origins in Coptic text reuse analysis (Miyagawa, 2022), but the normalization pipeline generalizes to other scripts.

3.3 Visualization

ICoMa provides six interactive visualization modules: (1) a parallel text viewer with highlighted matches, (2) a heatmap of pairwise token similarity, (3) a dispersion plot, (4) a similarity histogram, (5) a network graph of aligned tokens, and (6) an alignment flow diagram. Results are exportable in JSON format.

The dispersion plots are particularly informative for philological analysis. Each point represents a matched token pair, plotted by position in Witness α (x-axis) and Witness β (y-axis). Dense diagonal clustering indicates that the two texts preserve matching material in the same sequential order; off-diagonal scatter indicates rearrangement or formulaic reuse in non-corresponding positions.

3.4 Data Preparation

The corpus consists of MS, KS, and TS texts in romanized transliteration for the two ritual sections. Two text forms were prepared: a *plain* form preserving the original surface text, and a *lemmatized* form with words reduced to dictionary citation forms. The texts in lemmatized form are provided by Oliver Hellwig, who lemmatized the plain texts using the latest method described in Nehrdich et al. (2024).

3.5 Experimental Setup

All six algorithms were applied to each pairwise combination (MS–KS, MS–TS, KS–TS) for both ritual sections in both text forms. Window size was fixed at $n = 3$. Two metrics are reported: *coverage*, the percentage of tokens participating in at least one match above the similarity threshold; and *mean similarity*, the average score across all detected matches.

Thresholds were calibrated for each tool so that the similarity score for the MS–KS Agnyupasthāna (lemma text) would be close to 50%. The same thresholds were then consistently maintained for all analyses conducted in the same text form within each respective tool. This comparison was expected to fall roughly in the middle among the six pairs, thereby enabling an effective detection of both lower and higher degrees of similarity. As a negative control, a chapter from the MS that differs in both style and content (Darśapūrṇamāsa-Mantras) was included in the analysis.

4 Results

Table 1 presents coverage and mean similarity for each algorithm, text form, and text pair. Character n-gram results are omitted because they are identical to Levenshtein under the current parameters (Section 4.4).

4.1 Cross-Recension Similarity Rankings

All five independent algorithms and both text forms yield the same pairwise ranking without exception: MS–KS > KS–TS > MS–TS. This ranking holds for both Agnyupasthāna and Punarādhāna. The consistency is notable given that the algorithms measure fundamentally different properties—edit distance, character overlap, token-set intersection, local alignment, and diacritic-normalized edit distance—yet they converge on the same conclusion. All the results of the analysis are shown in Table 1.

The negative control confirms that the detected reuse is not attributable to general Sanskrit lexical overlap. Control coverage ranges from 1.9% to 3.1% across algorithms, well below the 47–55% observed for MS–KS Agnyupasthāna. Jaccard is the exception, with control coverage reaching 5.7% (lemmatized) and 13.0% (plain), reflecting its sensitivity to high-frequency character sequences common across any Sanskrit text. This indicates that Jaccard alone is insufficiently discriminating for Vedic text reuse analysis.

The ranking MS–KS > KS–TS > MS–TS is consistent across both coverage and mean similarity. However, it is coverage that highlights the dif-

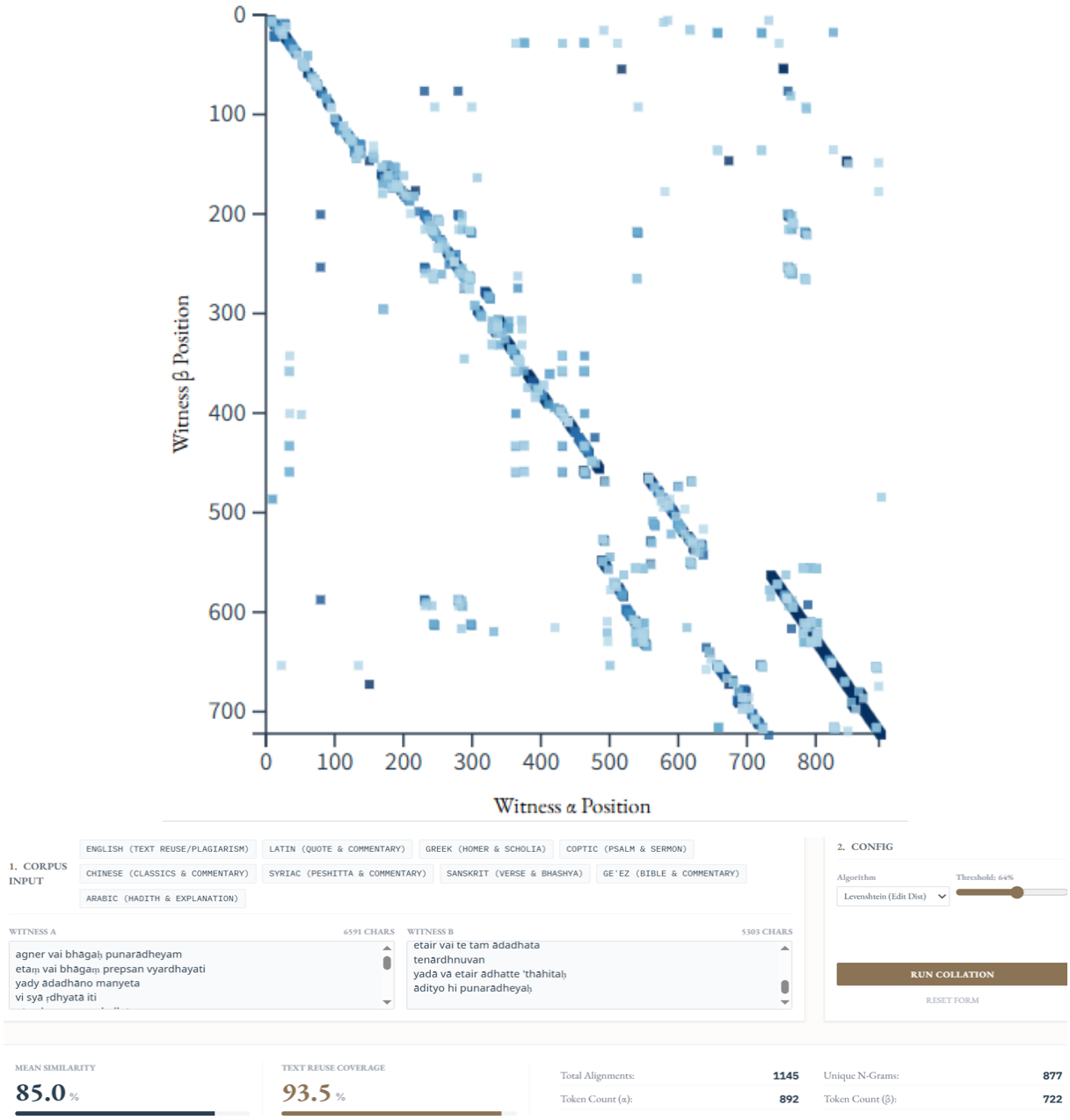


Figure 1: ICoMa interface: collation of MS and KS Punarādhāna (plain) using Levenshtein distance (threshold 64%). Coverage: 93.5%, Mean Similarity 85%, 1145 Alignments. Token counts: 728 (Witness α, MS), 722 (Witness β, KS). The dispersion plot shows dense diagonal clustering, confirming sequential correspondence.

		Reuse Coverage (%)				Mean Similarity (%)				
Algorithm	**Form**	MS–KS	MS–TS	KS–TS	Ctrl	MS–KS	MS–TS	KS–TS	Ctrl	**Thr.**
Levenshtein	lem/agnyup	50.2	30.3	36.3	2.5	91.7	91.4	90.4	91.0	78
	lem/punar	82.9	19.2	20.1	—	96.5	88.6	89.9	—	
Jaccard	lem/agnyup	49.4	31.9	33.4	5.7	99.4	99.4	99.3	99.2	92
	lem/punar	82.3	20.5	22.4	—	99.6	98.8	99.6	—	
Word n-gram	lem/agnyup	47.2	26.6	30.3	2.0	86.7	86.4	87.7	76.5	52
	lem/punar	80.5	18.8	22.2	—	88.4	77.2	76.8	—	
Script-aware	lem/agnyup	50.8	31.3	37.6	3.1	92.2	91.9	91.3	88.7	81
	lem/punar	82.9	20.5	21.0	—	96.8	89.9	91.2	—	
Levenshtein	pln/agnyup	50.1	31.9	36.2	1.9	80.2	77.8	77.5	85.2	64
	pln/punar	93.5	20.5	20.7	—	85.0	76.5	77.5	—	
Jaccard	pln/agnyup	51.0	31.0	33.2	13.0	93.4	93.1	93.4	91.4	86
	pln/punar	87.6	31.4	28.1	—	95.2	90.2	91.0	—	
Word n-gram	pln/agnyup	55.1	31.8	34.2	2.6	54.4	54.3	55.3	60.5	45
	pln/punar	89.6	23.5	25.4	—	59.0	51.6	52.4	—	
Script-aware	pln/agnyup	51.8	33.6	37.5	2.9	81.1	80.0	79.7	90.6	66
	pln/punar	93.4	21.5	23.1	—	87.2	78.1	78.9	—	
Smith-W.	lem/agnyup	3.8	—	—	—	100	—	—	—	20
	pln/agnyup	3.7	—	—	—	100	—	—	—	20

Table 1: Text reuse coverage and mean similarity across algorithms, text forms, and text pairs. "Ctrl" = negative control (MS Agnyupasthāna vs. MS Darśapūrṇamāsa mantra). "Thr." = similarity threshold (%). "lem" = lemmatized, "pln" = plain. Smith–Waterman returns the single best local alignment only. Character n-gram is omitted (identical to Levenshtein; see Section 4.4).

ferences between the pairs more clearly. Because mean similarity calculates the average score across all matches, it is leveled out by the low similarity values found in non-parallel portions of the texts. For this reason, in a study such as the present one—whose purpose is to compare parallel texts—mean similarity is not a particularly effective indicator.

4.2 Contrasting Reuse Profiles of the Two Ritual Sections

The most striking result is the sharp contrast between the two ritual sections (Table 2).

Punarādhāna shows extremely high MS–KS coverage (80.5–93.5%) but low MS–TS and KS–TS coverage (18.8–31.4%). The dispersion plot for MS–KS Punarādhāna (Figure 1) displays dense diagonal clustering, indicating that the two texts preserve matching material in the same sequential order. The small part of offset diagonal lines indicates passages where the order of the sentences has been rearranged, that is, similar descriptions occurring in a different sequence.

Agnyupasthāna shows moderate MS–KS coverage (47.2–55.1%) that is only 15–20 percentage points above MS–TS (26.6–33.6%) and KS–TS (30.3–37.6%). All three pairs share nontrivial textual material, but none approaches the near-total overlap seen in Punarādhāna. The KS–TS comparison (Figure 2) illustrates the corresponding dispersion pattern, with scattered rather than diagonal clustering, reflecting the diffuse, formulaic character of the shared material.

	Agnyupasthāna		Punarādhāna	
Pair	**Lem.**	**Plain**	**Lem.**	**Plain**
MS–KS	50.2%	50.1%	82.9%	93.5%
KS–TS	36.3%	36.2%	20.1%	20.7%
MS–TS	30.3%	31.9%	19.2%	20.5%

Table 2: Contrasting reuse profiles: Levenshtein coverage for the two ritual sections in both lemmatized and plain text forms.

4.3 Suitability of Lemmatized Text and Plain Text for Analysis

If we examine the difference between analyses based on lemmatized text and plain text using coverage as the more reliable indicator, the following result emerges (see Table 2). In the plain text, Punarādhāna MS–KS shows 93.5% compared to 50.1% for Agnyupasthāna MS–KS, while in the lemmatized text the former shows 82.9% compared to 50.2% for the latter (all numerical values based on Levenshtein). This suggests that the difference

Figure 2: ICoMa interface: collation of KS and TS Agnyupasthāna (plain) using Levenshtein distance (threshold 64%). Coverage: 36.2%, Mean Similarity 77.5%, 704 Alignments. Token counts: 1755 (Witness α, KS), 843 (Witness β, TS). The dispersion plot shows that the reuse is scattered throughout the section and is clearly different from the sequential correspondence observed in the MS–KS Punarādhāna (Figure 1).

in similarity is more clearly discernible in the plain text. This tendency can be observed across all tools. The higher similarity values in the plain text indicate a greater degree of surface-level correspondence, that is string identity or phonetic sequence identity between the texts.

Character-level analysis of plain text has the major advantage that the results are not affected by decisions such as how compounds are segmented in the data.

4.4 Algorithm Comparison

The convergence of five independent algorithms on the same pairwise rankings strengthens confidence that the observed patterns reflect genuine textual relationships.

The character n-gram algorithm produced results identical to the Levenshtein distance in every comparison. At the current window size ($n = 3$), concatenating three tokens into a character string before computing edit distance yields the same result as computing edit distance directly. Larger window sizes may break this equivalence and are planned as future work.

Smith–Waterman identifies only the single best aligned region per comparison (3.7–3.8% coverage, 100% similarity), providing limited value for quantifying overall text reuse in its current implementation. Among the sliding-window algorithms, Levenshtein and script-aware normalization offer the best combination of sensitivity and discriminating power.

Jaccard shows one anomaly: its MS–TS coverage rises from 20.5% (lemmatized) to 31.4% (plain), while other algorithms increase by less than 5 points. This parallels its elevated control coverage on plain text (13.0%) and reflects its set-based sensitivity to high-frequency character patterns.

The differences in intertextual similarity are most clearly pronounced by the two tools: Levenshtein shows Punarādhāna MS–KS 93.5% vs. MS–TS 20.5%, and Script-Aware shows Punarādhāna MS–KS 93.4% vs. MS–TS 21.5% (both based on the plain text).

It should also be noted that Miyagawa et al. (2024) analyzed the lemmatized text of the Punarādhāna chapters of MS and KS using a Word2Vec embedding approach, and obtained results that are almost identical to those of the present text-reuse analysis. In the Vedic parallel prose texts examined here, where surface-level correspondences are abundant, character- (i.e., sound-) based similarity analysis can therefore be considered sufficiently effective, to a degree comparable to semantic analysis.

5 Discussion: Implications for Textual History

The contrasting reuse profiles of Agnyupasthāna and Punarādhāna provide quantitative evidence that different ritual sections followed distinct transmission paths within the Kṛṣṇa Yajurveda tradition.

The near-identical Punarādhāna material in MS and KS (coverage up to 93.5%), combined with the sharp divergence from TS (coverage around 20%), suggests that the Maitrāyaṇī and Kāṭhaka schools maintained a particularly close textual lineage for this section. The dense diagonal clustering in the dispersion plot (Figure 1) further indicates that this shared material was preserved not merely as a common vocabulary but in the same sequential arrangement, consistent with the view that the texts had already been fixed and shared almost identical material at the surface-text level, most plausibly through direct borrowing. The TS school appears to have drawn on a different source tradition or substantially redacted the material.

Agnyupasthāna presents a contrasting pattern. The moderate, broadly distributed similarity across all three pairs (MS–KS: 47–55%, KS–TS: 30–38%, MS–TS: 27–34%) suggests that this ritual section circulated more widely among the three schools during an earlier period of shared transmission. The absence of any dramatically close pair implies that all three recensions appear to have shared the same ritual, but because this was a period in which the texts had not yet been fixed, there is relatively little agreement at the surface-text level. The scattered dispersion plot for KS–TS Agnyupasthāna (Figure 2) illustrates this diffuse pattern visually.

The robustness of these patterns across five independent algorithms (Section 4.4) ensures that the conclusions are not artifacts of a particular similarity metric.

6 Conclusion

This study applied multi-algorithm computational text reuse detection to three Kṛṣṇa Yajurveda recensions (MS, KS, TS) across two ritual sections. Three principal findings emerge.

First, all five independent similarity algorithms consistently rank MS–KS as the most closely related pair, followed by KS–TS and then MS–TS, in

both plain and lemmatized text forms. This convergence provides the first multi-algorithm quantitative confirmation of the hypothesis of Amano (2019, 2020) on recension relationships.

Second, it has become clear that, in comparing parallel texts of the Black Yajurveda, a character-level text-reuse analysis based on the plain text is effective. This is because these texts reflect the results of oral transmission and interaction, thereby highlighting a characteristic feature of Vedic literature: the strong emphasis on preserving sound without alteration.

Third, in comparison with other chapters and other intertextual pairings, Punarādhāna MS–KS exhibits a strikingly high degree of similarity and the same sequential arrangement. Unlike the level of parallelism observed in other chapters or combinations, Punarādhāna MS–KS provides strong evidence of the contact status where the texts were fixed, most plausibly a direct borrowing relationship. What had been hypothesized in Amano (2014–2015) can thus be said to have been substantiated by numerical data. In this manner, a way has been opened to explain the differences in similarity as reflecting differences in the periods at which the texts became fixed.

Until now, Vedic parallel texts have generally been understood as branching solely from a common archetype. If the existence of borrowing relationships can indeed be demonstrated, this would offer a new perspective on the formation history of Vedic literature.

The analysis was conducted using ICoMa, a freely available web-based collation tool supporting ten script systems, which is applicable to text reuse research across a range of historical traditions. Future work will implement iterative Smith–Waterman alignment for richer local analysis, and integrate semantic similarity measures through word embeddings to enable detection of paraphrastic reuse. However, the most significant prospect lies in having demonstrated the effectiveness of analysis on plain text, which makes it possible to extend the analysis to other chapters through rapid and streamlined processing with ICoMa. Only when all chapters of the three texts have been analyzed will it become possible to clarify the overall formation process of the Kṛṣṇa Yajurveda.

7 Limitations

ICoMa's word-level tokenization requires consistent word boundary conventions in input data (Section 3.1); the present study controlled for this manually. The negative control was applied only to Agnyupasthāna; extending it to Punarādhāna would strengthen the design. The similarity thresholds were set by manual histogram inspection rather than systematic optimization against gold-standard annotations. The analysis does not distinguish between mantra and brāhmaṇa content types within each section. Finally, the restriction to two ritual sections limits the generalizability of the transmission-historical conclusions.

Ethical Considerations

The texts analyzed are ancient liturgical compositions in the public domain. No human subjects were involved.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21KK0004.

References

Kyoko Amano. 2014–2015. Zur klärung der sprachschichten in der maitrāyaṇī saṃhitā. *Journal of Indological Studies*, 26/27:1–36.

Kyoko Amano. 2019. The development of the uses of *ha / ha vái / ha sma vái* with or without the narrative perfect and language layers in the old Yajurveda-Saṃhitā texts. *Lingua Posnaniensis*, 61:11–24.

Kyoko Amano. 2020. What is 'knowledge' justifying a ritual action? Uses of *ya evaṃ veda / ya evaṃ vidvān* in the Maitrāyaṇī Saṃhitā. In C. Redard and 1 others, editors, *Aux sources des liturgies indo-iraniennes*, volume 10 of *Collection Religions, Comparatisme – Histoire – Anthropologie*, pages 39–68. Presses Universitaires de Liège, Liège.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

So Miyagawa. 2022. *Shenoute, Besa and the Bible: Digital Text Reuse Analysis of Selected Monastic Writings from Egypt*. SUB Göttingen.

So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi, and Kyoko Amano. 2024. Exploring similarity measures and intertextuality in Vedic Sanskrit literature. In

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH), pages 123–131, Miami, USA. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751.

Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Michael Witzel. 1997. The development of the Vedic canon and its schools: The social and political milieu. In Michael Witzel, editor, *Inside the Texts, Beyond the Texts: New Approaches to the Study of the Vedas*, pages 257–345. Harvard University.

Data Contamination in Neural Hieroglyphic Translation: A Reproducibility Study

Ammar Toutou[1] **Abdelrahman Harb**[1] **Christine Basta**[2,3]
[1]Computer Science and Engineering, Alamein International University (AIU), Egypt
[2]HiTZ Center, University of the Basque Country, Spain
[3]Faculty of Computers and Data Science, Alexandria University, Egypt
{ammar.mohamed.2023, abdelraheman.shehata.2023}@aiu.edu.eg
christine.basta@alexu.edu.eg

Abstract

Ancient and endangered languages pose a unique challenge for NLP: their datasets are inherently scarce, difficult to expand, and built from formulaic corpora—making data-quality issues especially consequential yet rarely audited. Motivated by the need to understand what current NMT can realistically achieve for such languages, we investigate hieroglyphic-to-German translation, where a recent study reported 61.5 BLEU using fine-tuned M2M-100. Our reproduction yields only 37.0 BLEU with the released model. Investigating this gap, we find **32% of test targets appear identically in training** (16/50; 50% under 8-gram overlap at 70% threshold). This contamination inflates scores dramatically: contaminated samples achieve up to 83.8 BLEU / 0.924 COMET-22 versus 30.9–39.2 BLEU / 0.622–0.676 COMET-22 on clean samples across five model configurations spanning two architectures. Document-level decontamination reduces contaminated BLEU by only 4.6 points because 8/16 targets persist via other source documents—target-level deduplication is required. We release a decontaminated 34-sample test set and establish corrected baselines (30.9–39.2 BLEU), providing a realistic assessment of NMT capability for this endangered writing system.

1 Introduction

Ancient and endangered languages—including Ancient Egyptian, Akkadian, and Classical Latin—represent some of the most data-scarce domains in NLP (Sommerschield et al., 2023). Unlike modern languages where parallel corpora can be crowdsourced or web-scraped at scale, ancient-language datasets are inherently limited: they depend on surviving archaeological artifacts, require specialist expertise for annotation, and cannot be expanded on demand. The Thesaurus Linguae Aegyptiae (TLA), the largest digitized resource for Ancient Egyptian, yields only **18,669 usable training pairs** after filtering for samples with both digitized hieroglyphic source and German target—orders of magnitude smaller than modern MT benchmarks. This scarcity makes data-quality issues especially consequential: every contaminated or duplicated sample has outsized impact on evaluation.

Motivated by the need to establish what NMT can realistically achieve for such resource-limited languages, we investigated hieroglyphic-to-German translation. De Cao et al. (2024)[1] reported 61.5 BLEU using M2M-100 (Fan et al., 2021) fine-tuned on TLA data—a score approaching neural ceilings on high-resource benchmarks (Popel et al., 2020). Such performance would be remarkable for an endangered writing system with severely limited data. However, our reproduction using the publicly released model yielded only **37.0 BLEU**, a gap of over 24 points.

This discrepancy led us to examine the dataset itself, where we discovered that **32% of test samples have German target translations appearing identically in the training data**. This contamination—arising naturally from formulaic repetition in ancient corpora (medical instructions, offering formulae, royal epithets)—enables models to achieve high scores through memorization rather than genuine translation. On the 16 contaminated test samples, our best model achieves **83.8 BLEU**; on the 34 clean samples, only **37.3 BLEU**—a 47-point gap. This pattern holds across all five models tested (29–47 point gaps), confirming the problem is dataset-inherent.

These findings have broader implications: ancient corpora universally contain formulaic repetition, making train-test contamination likely under standard random splitting. Our corrected baselines (30.9–39.2 BLEU) provide a realistic assessment—useful for corpus triage but requiring

[1]Model publicly available on HuggingFace.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 50–57
July 4, 2026 ©2026 Association for Computational Linguistics

expert verification—and highlight the need for contamination auditing in all ancient-language NLP.

Contributions.

1. **First contamination audit for ancient-language NMT.** We show that formulaic repetition—pervasive in ancient Egyptian corpora—creates systematic data-leakage risks absent from modern-language benchmarks, and identify 32% exact target overlap in the public test set.

2. **Cross-architecture, cross-metric robustness.** The inflation is dataset-inherent, not model-specific: contaminated samples score 29–47 BLEU and 0.23–0.26 COMET-22 points higher than clean samples across five configurations spanning two architectures (M2M-100/mBART-50).

3. **Graduated decontamination analysis.** Beyond exact matching, we apply character 8-gram overlap thresholding, document-level leakage analysis, and per-item frequency cataloguing to characterize contamination severity.

4. **Corrected baselines and reusable evaluation protocol.** We release a decontaminated test set (34 samples, no target overlap), contamination-detection scripts, and corrected baselines (30.9–39.2 BLEU) as a realistic performance range for future work.

2 Background

Hieroglyphic Translation. Ancient Egyptian hieroglyphics (ca. 3200 BCE–400 CE) use a mixture of logographic and phonetic signs. The TLA provides the largest parallel resource of transliterations and German translations; hieroglyphs are encoded in Gardiner notation for computational processing. De Cao et al. (2024) fine-tuned M2M-100 (Fan et al., 2021) on TLA data, reporting 61.5 BLEU for hieroglyphic-to-German.

Data Contamination. Contamination occurs when evaluation targets leak into training. Magar and Schwartz (2022) and Kocyigit et al. (2025) showed 20–30 BLEU inflation in controlled experiments. Ancient corpora are particularly susceptible: formulaic phrases repeat across many texts, so random splitting distributes identical targets across partitions—**a methodological challenge rather than a researcher oversight**. Reproducibility challenges are endemic in ML (Pineau et al., 2021); to our knowledge, no prior study has audited contamination in ancient-language NMT.

Hieroglyphic NLP. Computational approaches have addressed recognition (Franken and van Gemert, 2013; Barucci et al., 2021), transliteration (Wiesenbach and Riezler, 2019), and translation (De Cao et al., 2024). Chen et al. (2024) present the first multi-task benchmark for ancient logographic systems; unlike LogogramNLP, our work audits contamination and provides splitting protocols.

Contamination Detection Methods. Beyond controlled experiments (Section 2), recent work has revealed contamination in deployed systems: Tan et al. (2026) showed cross-direction contamination in FLORES-200; Abbas et al. (2026) demonstrated obfuscation-hidden contamination in Arabic NLP; Enis and Hopkins (2024) identified test data in closed models' training. Detection methods such as permutation testing (Oren et al., 2023), distribution analysis (Dong et al., 2024), and generalization-gap testing (Dekoninck et al., 2024) would be complementary for future closed-model evaluations. Reproducibility challenges are well-documented (Pineau et al., 2021; Sommerschield et al., 2023); our contribution is the first contamination audit for ancient-language NMT.

3 Methodology

3.1 Data Sources

We use the publicly released data from the hiero-transformer repository:[2]

Table 1 summarizes the filtering pipeline. The TLA's digitization is incomplete: approximately two-thirds of German entries have a translated target but no digitized hieroglyphic source in Gardiner notation—they contain only transliterations, which cannot be used as model input. After filtering for entries with *both* non-empty hieroglyphic source and German target, only **18,669 usable pairs** remain—this is the actual training set for all fine-tuning. Similarly, each test hieroglyphic sentence appears twice (once with a German target, once with English); entries whose German target is empty have translations only in English. This yields just **50 valid test samples**—a size reflecting the fundamental scarcity of digitized hieroglyphic

[2] https://github.com/mattiadc/hiero-transformer

Statistic	Count
Training samples (total)	61,330
German-target (non-empty target)	55,397
German-target (non-empty src & tgt)	18,669
Unique German training targets	15,842
Test samples (total)	150
ea→de, valid	50
Validation samples (total)	125
ea→de, valid	75

Table 1: Dataset statistics. "Valid" = non-empty hieroglyphic source *and* non-empty German target.

data, where expansion requires new expert annotation of archaeological sources.

3.2 Contamination Detection

Normalized exact match. We compare each test target against all 15,842 unique German training targets after normalization (Unicode NFC $\rightarrow$ lowercase $\rightarrow$ punctuation removal $\rightarrow$ whitespace collapse; released as `normalize_translations.py`). Normalization is necessary because Egyptological annotations use bracket conventions (e.g., "[Werde]" vs. "Werde") that mask near-duplicates. This yields the contamination set $\mathcal{C} = \{(s,t) \in \mathcal{D}_{\text{test}} : \text{norm}(t) \in \{\text{norm}(t') : t' \in \mathcal{T}_{\text{train}}\}\}$, partitioning tests into **contaminated** ($|\mathcal{C}| = 16$) and **clean** (34) subsets.

Character n-gram overlap. To capture soft leakage beyond exact matches, we compute the fraction of character 8-grams in each test target appearing in any training target, reporting the number of samples flagged at overlap thresholds from 50% to 100%.

3.3 Models Evaluated

We evaluate five model configurations spanning two training regimes:

Original-regime models. **Released Model**: The publicly available HuggingFace checkpoint of De Cao et al. (2024), an M2M-100 (418M) model fine-tuned on TLA data with Adam (lr=3e-5, fixed schedule, batch size 16). This represents what independent researchers can access.

Script Reproduction: We retrained M2M-100 with its default hyperparameters (epochs=20, batch_size=16, lr=3e-5, Adam, no warmup, no label smoothing). This represents the closest possible replication of the original training procedure.

Our retrained models. The following three models use a modernized training recipe: AdamW optimizer, cosine learning-rate schedule with warmup, label smoothing (ϵ=0.15), weight decay (0.1), dropout tuning (0.2/0.05/0.05 for feedforward/attention/activation), gradient clipping (max norm 1.0), and 5$\times$ data upsampling—all applied identically across the three configurations. The effective batch size is 288 (vs. 16 in the original). These choices follow recent best practices for low-resource MT.

M2M-100 Hybrid: M2M-100 (418M) with lr=3e-5 and 1000-step warmup.

M2M-100 Conservative: Identical to Hybrid except lr=1e-5.

mBART-50: mBART-50 (611M) with lr=3e-5 and 500-step warmup, testing whether contamination effects generalize across architectures.

Using a stronger training recipe is deliberate: if contamination inflation persists even with improved optimization, the effect is clearly dataset-inherent rather than an artifact of under-training.

3.4 Evaluation Protocol

We evaluate on three subsets:

- **All**: Complete test set (50 samples)
- **Contaminated**: Samples with targets in training (16 samples)
- **Clean**: Samples with unseen targets (34 samples)

For each subset, we report:

- **BLEU**: Corpus-level BLEU using SacreBLEU (Post, 2018) with case-insensitive scoring (signature: `nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.6.0`)
- **chrF++**: Character n-gram F-score with word bigrams (Popović, 2015)

We report BLEU with `case:lc` (lowercased internally by SacreBLEU before tokenization) on unmodified model outputs. An ablation comparing case-sensitive versus case-insensitive evaluation confirms that contamination gaps are consistent regardless of text preprocessing.

Generation uses beam search with beam size 10 and maximum length 128, following typical configurations for M2M-100.

4 Results

4.1 Contamination Statistics

Exact matching finds 15/50 (30.0%) verbatim targets; after normalization (lowercasing, whitespace and punctuation standardization), **16/50 (32.0%)** match, where the 16th sample differs from its training counterpart only in parenthetical formatting. The validation set shows a comparable rate (23/75, 30.7%). At the 70% threshold with character 8-grams, 25/50 (50%) of test samples are flagged, confirming that 32% exact match is a conservative lower bound. The consistent rates across test and validation sets indicate a systematic characteristic of the splitting procedure.

4.2 Impact on Automatic Metrics

Table 2 presents translation quality stratified by contamination status.

The gap is substantial across all five models (+29 to +47 BLEU), holding across architectures (M2M-100 vs. mBART-50) and training strategies—the problem is inherent to the dataset. Despite wide 95% CIs on the small subsets (clean: $\pm$7–14 BLEU; contaminated: $\pm$15–20 BLEU), the contaminated and clean intervals do not overlap for any model, confirming the gap is statistically robust. The script reproduction outperforms the released model (+5.2 BLEU) despite identical hyperparameters; the gap is predominantly on contaminated samples (+17.2) versus clean (+2.9), consistent with different checkpoint selection or training seed rather than a systematic quality difference.

COMET-22 Validation. Table 3 confirms the gap using COMET-22 (Rei et al., 2022): contaminated samples score 0.871–0.924 while clean samples score 0.622–0.676 (a gap of 0.23–0.26 points). Model rankings on clean data are identical between BLEU and COMET-22, confirming genuine quality differences beyond contamination effects.

Retrieval Baseline. A retrieval baseline that, given access to the test target, copies the best-matching training target (by character 8-gram coverage) achieves 81.8 BLEU on exact-match items—rivaling the best neural model (77.9)—while scoring 0.0 on clean items. A realistic variant that uses only source similarity (KNN over source embeddings) achieves only 22.0 BLEU on exact-match items. The 59.8-point gap directly measures the contribution of target memorization over source-driven translation.

Granular Contamination Analysis. When we further split the 34 non-exact samples into soft leakage (70–99% 8-gram overlap; n=9) and fully clean (<70%; n=25), BLEU decreases monotonically across all five models—from exact to soft to clean—confirming that performance tracks contamination severity.

Source-Side Overlap. Exact-target samples have higher mean source overlap (0.883) than clean samples (0.543), but this reflects formulaic corpus structure, not causal attribution: **10 of 16 exact-target test items have at least one contaminating training sample with source similarity below 5%**, meaning the matched target was learned from a completely different hieroglyphic inscription. One test item ("Werde fein zerrieben.") has zero source overlap with any training sample yet achieves 100% target overlap. This directly demonstrates target memorization independent of source input.

English Direction. As a sanity check, ea→en evaluation (only 1/50 contaminated) yields 11.5–18.8 BLEU—below German clean scores, confirming contamination-driven inflation in the German direction.

Document-Level Decontamination. We remove every training sentence from the 33 test-source documents (28.4% of training), continue fine-tuning from M2M-100 Conservative for 3,000 steps at lr=3×10^{-6}, and evaluate the best checkpoint (step 300, Val BLEU 37.9). Table 4 reports results: exact-match BLEU drops only 4.6 points (77.9→73.3), while clean BLEU *increases* slightly (39.6→40.3). The residual contamination occurs because **8 of 16 exactly matched test targets persist in training from other source documents**—e.g., "Werde gekocht." (87 training occurrences) appears across many independent papyri. Document-level splitting is thus necessary but not sufficient: target-level deduplication is required.

4.3 Qualitative Analysis

Representative examples illustrate the contrast. Contaminated samples are reproduced verbatim; clean samples exhibit genuine translation errors: "hemsut" (a goddess) → "ausreiß" (escape), "uräen" (sacred cobras) → "rw.t-schlange" (a different serpent type), and "hunderte opfern ihm" (hundreds sacrifice to him) → "er zählt zu 100" (he counts to 100).

Manual categorization of errors in the 34 clean

Model	Subset	n	BLEU	95% CI	chrF++
Released	All	50	37.0	[24.5, 48.8]	56.2
	Contaminated	16	60.3	[37.9, 88.0]	81.5
	Clean	34	30.9	[16.2, 43.6]	48.1
Script Reproduction	All	50	42.2	[28.4, 55.0]	61.2
	Contaminated	16	77.5	[51.5, 93.4]	85.1
	Clean	34	33.8	[20.1, 48.2]	53.7
M2M-100 Hybrid	All	50	46.8	[33.9, 59.2]	63.3
	Contaminated	16	**83.8**	[62.5, 100.0]	**91.9**
	Clean	34	37.3	[23.6, 49.8]	54.0
M2M-100 Conservative	All	50	**47.3**	[35.0, 57.4]	**63.2**
	Contaminated	16	77.9	[59.2, 92.8]	90.6
	Clean	34	**39.2**	[25.0, 50.9]	**54.2**
mBART-50	All	50	41.9	[29.6, 51.8]	59.6
	Contaminated	16	72.8	[55.7, 92.4]	87.3
	Clean	34	33.7	[18.3, 47.1]	50.6

Table 2: Translation quality by contamination status across five models (`case:lc` BLEU, 95% CIs from 1,000 bootstrap resamples). All models show substantial contamination gaps (29–47 BLEU points). Bold indicates best performance per subset.

Model	All	Contam.	Clean
Released (HF)	0.702	0.871	0.622
Exact `train.py`	0.721	0.871	0.650
M2M-100 Hybrid	0.744	**0.924**	0.660
M2M-100 Conservative	**0.753**	0.915	**0.676**
mBART-50	0.732	0.905	0.651

Table 3: COMET-22 scores (`Unbabel/wmt22-comet-da`) on all 50 test samples, the 16 contaminated samples, and the 34 clean samples. Contaminated samples consistently score 0.17–0.26 points higher than clean samples, confirming the contamination-driven inflation observed in BLEU.

predictions reveals: domain-term substitution (35%), entity/deity confusion (24%), syntactic role reversal (18%), number/quantifier errors (15%), and missing content (9%). These error patterns are particularly concerning for DH applications requiring precise entity identification.

4.4 Phrase-Type Analysis

Of the 16 contaminated samples, 13 (81%) are medical formulae—short imperative instructions such as "Werde getrunken" (be drunk) and "Werde fein zermahlen" (be finely ground)—which repeat dozens of times in training. The remaining three are narrative markers (2) and a religious phrase (1). This concentration in formulaic genres confirms that contamination arises systematically from the data structure. The comparable contamination rates between test (32%) and validation (31%) sets further confirm this is a systematic characteristic of the split, not an isolated anomaly.

4.5 Illustrative Mixture Analysis

To illustrate how contamination inflates corpus-level BLEU, we simulate scores at controlled rates using the M2M-100 Conservative model's actual predictions (200 bootstrap trials per rate). At α=0 (fully clean) BLEU=40.2; at our observed rate (α=0.32) BLEU=46.5; at α=0.50 BLEU=52.0; at α=1.0 BLEU=77.9. Under these assumptions, the reported 61.5 BLEU would require a substantially higher contamination rate than the 32% we observe in the released test set, though we cannot estimate a precise rate without access to the original model's predictions and test composition.

5 Discussion

5.1 Sources of Contamination

Three factors produce 32% contamination: (1) **formulaic repetition** in ancient texts—medical instructions, offering formulae, and divine epithets recur across thousands of inscriptions; (2) **source-only deduplication**, which misses cases where different hieroglyphic sequences share the same German target; and (3) **sentence-level splitting**—all 33 unique test documents also appear in training (the highest-overlap document contributes 804 training and 5 test samples), confirming sentence-level rather than document-level partitioning.

5.2 Implications

Our contamination gaps (29–47 BLEU) exceed those in prior work: Kocyigit et al. (2025) found 20–30 point inflation in LLM-based MT, and Magar

Model	All (n=50)	Exact (n=16)	Soft (n=9)	Clean (n=25)
M2M-100 Conservative	47.3	77.9	35.3	39.6
Doc-Clean retrain	45.2	73.3	30.7	**40.3**
Δ	-2.1	-4.6	-4.6	$+0.7$

Table 4: BLEU scores before and after document-level decontamination (M2M-100 Conservative re-trained on 13,365 samples with test-source documents excluded). Exact: 16 samples with 100% 8-gram coverage; Soft: 9 samples with 70–99% coverage; Clean: 25 samples below threshold. Despite removing 28.4% of training data, exact-match BLEU drops only 4.6 points, because 8/16 contaminated targets persist in training from other source documents.

and Schwartz (2022) documented substantial inflation across NLP benchmarks. The gap between reported (61.5) and clean (30.9–39.2) BLEU represents plausible inflation of 22–31 points. Latin, Greek, Akkadian (Gutherz et al., 2023), and other ancient language corpora share the property of formulaic repetition; their benchmarks should be examined similarly.

5.3 Practical Guidance

Our corrected baselines (30.9–39.2 BLEU) indicate that NMT captures overall meaning but contains notable errors. For *corpus triage*, BLEU around 30–39 provides sufficient gist; for *philological analysis*, domain-term substitution (35% of errors) and entity confusion (24%) risk silent misinterpretation. NMT should be deployed as a screening tool with mandatory expert verification.

We recommend that **benchmark creators** check for train-test target overlap and report contamination statistics; **model developers** report performance separately on contaminated and clean subsets; and **consumers** treat reported scores as upper bounds pending contamination analysis.

5.4 Decontaminated Test Set and Splitting Protocol

We release a decontaminated test set of 34 samples (no exact target overlap) spanning medical/magical (41%), religious/funerary (32%), and administrative/literary (27%) content. Clean BLEU (30.9–39.2) represents a *lower bound* on contamination-removal effect and an *upper bound* on true generalization, since residual phrase-level overlap may provide partial memorization benefit.

To prevent contamination, we propose a **document-level splitting protocol**: (1) partition at the level of source texts rather than sentences; (2) deduplicate targets via frequency caps or cluster-and-drop; (3) stratify by genre; and (4) verify zero exact-match target overlap post-hoc. Future work should construct larger test sets from text collections not used in training.

6 Limitations

Test set size. Our clean set (34 samples) yields wide bootstrap CIs ($\pm$7–14 BLEU), though the consistent gap direction across all models confirms robustness. **Soft leakage.** Our primary detection uses exact matching (32%); n-gram analysis shows 50% at the 70% 8-gram threshold, but subset sizes preclude separate evaluation on graduated tiers. **Unavailable original artifacts.** We cannot access the original test set, checkpoint "ea9all", or "test_data" folder, so we cannot directly replicate the reported 61.5 BLEU evaluation. **Length confound.** Contaminated targets are shorter (mean 4.1 vs. 5.8 words), which may partly amplify BLEU inflation—though the retrieval baseline (81.8 BLEU with no translation) confirms the effect is real. **No human evaluation.** We rely on automatic metrics; Egyptologist judgements would strengthen assessment.

7 Conclusion

We reveal that 32% of test samples in a hieroglyphic-to-German NMT benchmark have targets matching training data (50% under 8-gram overlap at 70% threshold). This contamination inflates BLEU by 29–47 points across five model configurations: contaminated samples achieve up to 83.8 BLEU / 0.924 COMET-22 while clean samples achieve only 30.9–39.2 BLEU / 0.622–0.676 COMET-22. Document-level decontamination reduces exact-match BLEU by only 4.6 points, because 8/16 targets persist across other source documents—target-level deduplication is required.

We release a decontaminated 34-sample test set and establish corrected baselines for future work. More broadly, formulaic repetition in ancient corpora makes contamination particularly likely, and

we recommend future benchmarks report contamination statistics and partition results by contamination status.

Reproducibility Statement

All data used in this study is publicly available from the hiero-transformer repository. Our contamination detection code, the decontaminated test set (no exact target overlap, 34 samples), evaluation scripts, a per-item catalog mapping all 50 test items to document IDs, contamination status, training frequency, and source overlap scores, and all retrained model checkpoints are available at https://github.com/ammarlhassan/hiero-contamination-study.

Ethics Statement

Our work involves analysis of publicly released models and data. We present our findings as a constructive contribution to reproducibility rather than criticism of prior work—the contamination we identify likely arose from standard preprocessing procedures rather than intentional data manipulation.

Ancient Egyptian texts constitute irreplaceable cultural heritage. We encourage use of our corrected baselines to provide realistic expectations for NMT-assisted Egyptological research, avoiding overreliance on automated systems for scholarly work.

References

Chaymaa Abbas, Nour Shamaa, and Mariette Awad. 2026. Obscuring data contamination through translation: Evidence from arabic corpora. *arXiv preprint arXiv:2601.14994*.

Andrea Barucci, Costanza Cucci, Franco Franci, Marco Loschiavo, and Fabrizio Argenti. 2021. A deep learning approach to ancient Egyptian hieroglyphs classification. *IEEE Access*, 9:123438–123447.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14148–14167. Association for Computational Linguistics.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: A hieroglyphic transformer for translating ancient egyptian. *Digital Scholarship in the Humanities*.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. 2024. ConStat: Performance-based contamination detection in large language models. *arXiv preprint arXiv:2405.16281*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Morris Franken and Jan C van Gemert. 2013. Automatic Egyptian hieroglyph recognition by retrieving images as texts. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 765–768.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5):pgad096.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. *arXiv preprint arXiv:2501.18771*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 157–165.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

David Tan, Pinzhen Chen, Josef van Genabith, and Koel Dutta Chowdhury. 2026. When flores bloomz wrong: Cross-direction contamination in machine translation evaluation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*.

Philipp Wiesenbach and Stefan Riezler. 2019. Multi-task modeling of phonographic languages: Translating middle Egyptian hieroglyphs. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Beyond Prompt-Sensitive Emotion Words: Stable Embeddings for Tang Poetry Analysis

Linyue Zhang **Feiyue Li**

Abstract

Many Tang-poetry emotion studies still use coarse labels (e.g., positive/negative), while recent LLM-based attempts face a practical problem: one-word emotion outputs are highly sensitive to prompt wording. When labels shift with phrasing, historical interpretation becomes hard to reproduce and hard to trust. Focusing on Tang poetry around the An Lushan Rebellion (安史之乱), we propose a fine-grained sentence-level workflow centered on **emotion embeddings**: we use continuous hidden-state vectors, run automatic clustering, and then consolidate labels for interpretation. On the same 3,198 emotional sentences, one-word outputs show only 50.3% A/B exact agreement, while embedding-based clustering remains stable and well distributed (normalized entropy 0.989; 20/20 active clusters). On 7,195 labeled sentences, a char-based baseline reaches 0.446 micro-F1 and 0.395 macro-F1. This multi-stage label-construction path supports historically grounded findings, including the emotional turning point around 762, and also reveals layered patterns that are less visible in coarse setups. These results suggest that stable representation is a prerequisite for turning computational outputs into credible evidence for humanities interpretation.

1 Introduction

Large language models have made many digital-humanities tasks easier, but historical literary analysis still has a stubborn problem: fluent output is not the same as reliable evidence. This is especially true for Classical Chinese poetry, where emotion is compressed into short lines and tightly bound to historical context.

This study examines emotional change in Tang poetry around the An Lushan Rebellion. Existing workflows often rely on discrete emotion words generated by a model. In our corpus, semantically similar lines can receive different one-word labels after small prompt edits, which weakens any historical claim built on those labels. Compared with many earlier studies that stay at coarse polarity (positive/negative) or broad sentiment bins, this setting asks for finer emotional granularity and stronger reproducibility.

Our solution is to treat continuous representations as the primary evidence. Instead of clustering generated words, we build sentence-level emotion embeddings from model hidden states, use clustering to discover candidate emotional structure, and then consolidate labels for downstream prediction and interpretation. This keeps the computational pipeline reproducible while leaving room for close reading.

The main contributions are:

- We propose a fine-grained Tang-poetry emotion analysis framework that moves beyond coarse polarity and explicitly addresses prompt-level instability in LLM-generated emotion words through embedding-based representation.
- We show that the resulting quantitative patterns are historically grounded: the year-level and poet-level trends align with textual evidence and historical records, including the 762 turning point.
- We report humanities findings that are difficult to obtain with coarse labels alone, illustrating how digital methods and close reading can jointly reveal new structure in the corpus.

Two commitments shape the humanities side of this paper. First, we care about *interpretative validity*, not only model scores: if we claim a turning point, readers should be able to trace it to sentence-level evidence, label distributions, and specific lines. Second, we treat model-induced categories as working tools, not fixed truth. Our multi-stage label process follows that logic: data-driven

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 58–68
July 4, 2026 ©2026 Association for Computational Linguistics

discovery, then philological consolidation, then an explicit interpretive layer. The experiments are organized in the same order: establish learnability, test representation stability, run split and mapping diagnostics, and then move to year/poet interpretation with direct count checks. We keep the main LLM fixed and perturb only the prompt templates, because the claim under test here is prompt sensitivity rather than base-model ranking; broader cross-model comparison is an important next step, but it is not necessary for the current argument.

We position this work at the intersection of literary NLP and emotion analysis (Bamman et al., 2014; Buechel and Hahn, 2017).

2 Related Work

Computational literary studies have shown that NLP can quantify stylistic and thematic patterns while still supporting interpretative reading. In literary NLP, modeling character and narrative signals at scale has demonstrated that probabilistic or discriminative methods can expose latent literary structure (Bamman et al., 2014).

Emotion analysis has also moved from coarse polarity to richer representations, including multi-dimensional and perspective-aware annotations (Buechel and Hahn, 2017). For Chinese Tang-poetry analysis, prior work includes hierarchical text classification for poetry corpora (Xiao, 2006), classical machine-learning classification on Tang-poem categories (Hu and Zhu, 2015), and emotion-term extraction in digital-humanities settings (Zhang et al., 2021). The Dalian emotion ontology lexicon remains a widely used lexical resource (Institute of Information Retrieval, Dalian University of Technology, 2007). For historical Chinese poetry, this motivates two design choices in our work: (1) preserve fine-grained labels instead of collapsing directly to broad sentiment; (2) explicitly separate computational representation from interpretative mapping.

For humanities-oriented NLP work, a score alone is not enough. Outputs should remain stable under small perturbations and should be usable as evidence in an argument about texts. Following this expectation, we evaluate the method on two axes: computational adequacy (performance and robustness) and interpretative adequacy (historical plausibility, traceability to textual evidence, and transparent category design). In short, our gap is threefold: prior Tang-poetry emotion analysis is often coarse-grained, LLM-based pipelines are still relatively few, and prompt-sensitive word outputs make historical interpretation unstable.

3 Task Definition and Data

3.1 Data Scope and Canonical Source

Our study addresses two linked questions: how to reduce prompt sensitivity at sentence level, and how to use stabilized signals to analyze year/poet emotion distributions in a historically meaningful way. We first build robust sentence representations, then aggregate to year/poet patterns in the 754–764 observation frame, with each poem processed as a sequence of sentence-level units with year and author metadata. The cleaned source corpus contains **616 poem records** by **47 poets**, yielding **9,965 sentence instances** before final manual alignment. The final annotated table used for modeling expands this to **10,122 rows**, because some concatenated lines were manually split during annotation so that sentence and label counts could be aligned at the line level.

The workflow has two stages: (1) sentence extraction and multi-label annotation; (2) embedding construction, clustering, and year/poet aggregation.

To avoid denominator drift, all sentence totals and label statistics use one finalized annotated table as the canonical source. For per-sentence label-count statistics, duplicate labels inside the same sentence are deduplicated while preserving first-occurrence order. All year-level ratios in the paper are normalized within each year, so the interpretation is about emotional composition rather than raw poem totals.

Label construction and interpretive mapping. The 18-label inventory is derived from, not independent of, the 20-cluster step. We first run embedding-based unsupervised clustering and, using the elbow criterion, select $k = 20$ as an initial partition. We then manually merge clusters with unclear semantic boundaries to form 18 annotation labels for supervised modeling. In short, $k = 20$ serves exploratory discovery, while 18 labels serve downstream prediction and interpretation. Primary modeling is performed on this finalized 18-label inventory. For historical interpretation, we additionally map the 18 labels to 7 macro-categories (Sorrow, Disgust, Joy, Anger, Fear, Praise, Confusion), following the category framework of the Dalian University of Technology emotion ontology lexi-

Metric	Value / Note
Cleaned source corpus	616 poem records; 47 poets; 9,965 sentence instances
Total sentence rows	**10,122**; canonical annotated file
Rows with non-empty labels	**7,195**; used in supervised comparison
Rows without labels	**2,927**; retained for descriptive statistics
Label scheme	**18 labels, 0–3 per sentence**; obtained by manual consolidation of 20 initial clusters
Inter-annotator agreement	**Two annotators**, 200-sentence sample; **0.970 / 0.796** (OA / Cohen's κ)
Clustering setup	Elbow-selected $k = 20$, $N = 3,198$ embedding clusters for initial label discovery

Table 1: Main dataset and analysis statistics.

con (Institute of Information Retrieval, Dalian University of Technology, 2007). This mapping is a derived analysis layer, not a replacement for the original annotation protocol. Accordingly, the 18-label inventory remains the canonical annotation space, while the 7-class view is used only to make macro trends easier to read.

4 Method

For readers from non-technical backgrounds, the method can be read as three steps: (1) represent each sentence with a stable embedding instead of a single generated word; (2) build fine-grained labels through "automatic grouping + manual consolidation"; (3) aggregate labels by year and poet, then check whether the patterns match historical records and close reading.

4.1 Label Construction and Prediction Tasks

Given embedding vectors $\mathbf{e}_i$, we first perform K-means for a range of k values and use the elbow curve of within-cluster sum of squares to choose $k = 20$ as the automatic initial partition. These 20 clusters are then manually inspected and consolidated where semantic boundaries are not sufficiently clear, producing the 18-label scheme used in the supervised task. Therefore, there is no one-to-one requirement between the 20 clusters and the 18 labels: the former are exploratory units, the latter are annotation labels.

18-label classification task. Each sentence can have multiple labels (up to three in our annotation). We therefore use a standard multi-label setup: each label is predicted independently, and errors are averaged across all 18 labels with binary cross-entropy.

We benchmark majority-label, TF-IDF word/char features with OvR Linear SVM, and TF-IDF char with OvR Logistic Regression.

Mapping to 7 macro-emotions. For interpretative and error-analysis views, we define a deterministic mapping $g : \mathcal{L}_{18} \to \mathcal{L}_7$. The target 7-category space follows the Dalian University of Technology emotion ontology framework (Institute of Information Retrieval, Dalian University of Technology, 2007). Operationally, this means: if any fine-grained label of a sentence belongs to a given macro category, that macro category is marked as present for the sentence. This mapped form is used for descriptive aggregation and interpretative statistics.

For the mapped 7-class classifier ablation (Table 5), we use a diagnostic single-label projection: take the first valid fine-grained label among {`label1,label2,label3`}, then apply g to obtain one 7-class target. This diagnostic setting does not replace the primary 18-label multi-label task.

4.2 Representation Stability and Aggregation

Our core innovation is to treat continuous hidden states as the primary semantic signal. Given sentence x_i, we apply a zero-shot prompt to obtain a one-word emotion output (auxiliary cue), while extracting the final-position hidden vector from the last transformer layer as embedding $\mathbf{e}_i \in \mathbb{R}^d$. We then cluster these vectors with K-means to discover candidate emotion structure before manual consolidation.

To quantify distributional coverage across clusters, we report normalized entropy:

$$H_{\text{norm}} = -\frac{1}{\log K} \sum_{j=1}^{K} p_j \log p_j, \qquad (1)$$

where p_j is cluster proportion.

Inter-annotator agreement. Inter-annotator agreement is measured with Cohen's κ: κ compares observed agreement against chance-level agreement, so it is stricter than raw accuracy.

Year/poet aggregation. For year t and category c, ratio is computed as:

$$p_{c,t} = \frac{n_{c,t}}{\sum_{c'} n_{c',t}}, \qquad (2)$$

and first-order change as

$$\Delta p_{c,t} = p_{c,t} - p_{c,t-1}. \tag{3}$$

These summaries support year-wise and poet-wise interpretation around the event window.

5 Experimental Design

Our empirical design serves three connected aims: to show that fine-grained sentence-level emotion analysis is feasible, to compare the stability of embeddings and direct word outputs, and to examine whether the resulting patterns can sustain historical interpretation.

5.1 Study Design

For clarity, we organize the analyses in five parts:

- **Predictive baseline** (Table 2): evaluate the 18-label task and compare basic feature/model choices.
- **Representation stability** (Table 3): compare one-word prompt outputs with continuous embeddings.
- **Split comparison** (Table 4): compare grouped split and random split.
- **Mapped-label view** (Table 5): inspect behavior in a simplified 7-class view.
- **Historical interpretation** (Tables 6–8): aggregate canonical annotated counts by year and poet, then validate with sentence-level cases.

5.2 Data Splits and Metrics

The main supervised analysis uses the 18-label multi-label setting on all rows with non-empty labels ($N = 7{,}195$). For the 7-class simplified view, each sentence is mapped from the first valid fine-grained label. For the stability comparison, we use a fixed subset of emotional sentences ($N = 3{,}198$) shared by prompt A/B comparison and embedding-cluster profiling ($k = 20$).

Main supervised comparison uses group split by full poem (80/20), repeated over 3 seeds. This reduces train-test overlap from same-poem style and diction. We report Micro-F1, Macro-F1, and exact-match subset accuracy for the 18-label task. For the mapped 7-class diagnostic, we report Macro-F1 and accuracy.

Model	Micro-F1	Macro-F1	Subset Acc.
Majority-label baseline	0.151 ± 0.005	0.015 ± 0.000	0.013 ± 0.002
Word TF-IDF (1–2) + OvR SVM	0.076 ± 0.003	0.022 ± 0.001	0.000 ± 0.000
Char TF-IDF (1–4) + OvR SVM	0.446 ± 0.003	0.395 ± 0.007	$\mathbf{0.256 \pm 0.005}$
Char TF-IDF (1–3) + OvR LR	$\mathbf{0.448 \pm 0.007}$	$\mathbf{0.402 \pm 0.012}$	0.214 ± 0.009

Table 2: 18-label multi-label results (group split by poem, 3 seeds).

Setting	Consistency	Norm. H	Support
Word output A	–	0.807	542 unique words
Word output B	–	0.764	431 unique words
A/B exact match	0.503	–	3,198 pairs
Embedding clustering ($k = 20$)	N/A (continuous)	**0.989**	**20**/**20** active clusters

Table 3: Prompt sensitivity of single-word outputs vs. embedding coverage.

For representation comparison, we use two deterministic prompt templates (A/B) on the same 3,198 emotional sentences with DeepSeek-V3.2 (T=0). The generated words are used only for stability comparison; historical interpretation is based on canonical labels and embedding-derived structure.

6 Quantitative Results

We report results in the same order as the argument of the paper. We first establish that the fine-grained task is learnable, then show why embedding-based representation is more stable than direct word outputs, and finally examine whether the recovered temporal/poet patterns are historically meaningful and humanistically informative.

6.1 Predictive Performance and Representation Stability

Character features substantially outperform word n-grams, consistent with Classical Chinese's weak explicit token boundaries in short lines. SVM is best on strict subset accuracy, while LR slightly leads on Micro/Macro-F1. These results provide a stable baseline for the later checks.

Representation stability (words vs. embeddings). Even under deterministic decoding, A/B mismatch is 0.497, indicating non-trivial prompt sensitivity for discrete one-word outputs. Embedding clusters remain fully populated (cluster size 102–257) with high normalized entropy, supporting better semantic coverage. This experiment isolates representation stability; downstream historical trend tables are computed from canonical annotated labels rather than generated words. In practical terms, Table 3 explains a core design decision of this paper: embeddings are used to discover

Setting	Micro-F1
Group split by poem	0.446 ± 0.003
Random sentence split	$\mathbf{0.466 \pm 0.002}$

Table 4: Leakage sanity check with Char TF-IDF (1–4) + OvR SVM.

Model (7-class mapped)	Macro-F1	Accuracy
Word TF-IDF (1–3), SVM (bal.)	0.105 ± 0.004	0.583 ± 0.031
Char TF-IDF (1–3), SVM (bal.)	0.448 ± 0.028	0.653 ± 0.025
Char TF-IDF (1–4), SVM (bal.)	$\mathbf{0.449 \pm 0.026}$	0.652 ± 0.024
Char TF-IDF (1–4), SVM (unbal.)	0.433 ± 0.023	$\mathbf{0.683 \pm 0.014}$

Table 5: Mapped 7-class single-label diagnostic ablation.

structure, while consolidated labels are used for interpretation.

Split robustness check. The random split inflates performance, confirming that grouped split is the safer protocol for main reporting.

7-class diagnostic. Balanced training improves macro fairness across rare classes, while unbalanced training favors overall accuracy. This 7-class view is only a supplementary lens and does not replace the main 18-label analysis. Visual summaries of Tables 6–8 are provided in Appendix **??**.

6.2 Temporal Trends and Poet-Level Patterns

All ratios in this subsection are computed from canonical annotated counts; no classifier predictions are used here.

Note: Melancholy = 忧伤; Grief = 悲愁; Longing = 思念; Heroic Aspiration = 壮思; Romantic Affection = 爱恋.

The temporal pattern is coherent rather than random: Sorrow remains dominant and reaches its highest level in 762, while Joy shows a short recovery in 760 and then declines. Anger rises locally around 757 but does not persist, and secondary labels show a similar structure: Heroic Aspiration is relatively stronger in early war years, then weakens, whereas Melancholy intensifies by 762. These shifts align with a transition from mobilizing rhetoric to sustained grief-oriented expression.

6.3 Error Analysis and Validation

In the mapped 7-class view (Table 5), Macro-F1 is 0.4464 and accuracy is 0.6441. Performance is strongest on high-support classes (especially Sorrow) and weaker on rare classes. Most confusions appear in lines with mixed affective cues, especially because the single-label projection compresses secondary emotions. This pattern reflects the simplification of the diagnostic setting and class imbalance, rather than a breakdown of the historical trend signals.

Year	Sorrow	Disgust	Joy	Anger	Fear	Praise	Confusion
754	0.6659	0.0230	0.1590	0.0760	0.0161	0.0288	0.0311
755	0.6269	0.0206	0.1624	0.1161	0.0154	0.0226	0.0360
756	0.6018	0.0222	0.1519	0.1264	0.0287	0.0320	0.0369
757	0.5664	0.0288	0.1616	0.1592	0.0168	0.0344	0.0328
758	0.6297	0.0167	0.1862	0.0900	0.0136	0.0303	0.0335
759	0.6463	0.0199	0.1347	0.1168	0.0223	0.0291	0.0310
760	0.6388	0.0136	0.2097	0.0641	0.0078	0.0408	0.0252
761	0.6276	0.0444	0.1616	0.0887	0.0158	0.0333	0.0285
762	0.7024	0.0312	0.1239	0.0722	0.0166	0.0244	0.0293
763	0.6653	0.0116	0.1274	0.1126	0.0189	0.0242	0.0400
764	0.6403	0.0292	0.1278	0.1111	0.0278	0.0389	0.0250

Table 6: Year-wise primary emotion ratios (754–764).

Year	Melan.	Grief	Longing	Heroic Asp.	Romantic Aff.
757	0.1256	0.1528	0.0696	0.0608	0.0104
760	0.1456	0.0971	0.1029	0.0447	0.0155
761	0.1474	0.1252	0.0856	0.0380	0.0063
762	0.1824	0.1307	0.0985	0.0361	0.0049
764	0.1417	0.1486	0.0750	0.0375	0.0083

Table 7: Selected-year secondary emotion ratios around key turning points.

Interpretative counts are directly validated in the canonical table: Heroic Aspiration in 757 appears 76 times; Melancholy in 762 appears 187 times; Romantic Affection in 762 appears 5 times. Representative lines include 时将白羽挥, 吞声不许哭, and 偏得浑家怜.

7 Humanities Interpretation and Discussion

This section turns quantitative patterns into literary-historical interpretation without claiming deterministic causality.

7.1 762 as a Literary Turning Point

The yearly trajectory is not a simple one-way decline. What we observe is a reweighting of emotional expression under prolonged historical pressure. The Sorrow surge in 762 aligns with intensified war-aftershock memory and social dislocation in poetic expression. To make the historical anchor explicit, our original Chinese manuscript cites *Zizhi Tongjian* with the key chronology statement (Sima, 1956):

> 四月玄宗崩……五月肃宗崩。

Xuanzong died in the fourth lunar month and Suzong in the fifth, concentrating dynastic loss and succession stress within a short span. This chronology offers concrete historical context for the emo-

Poet	Sent.	Melancholy	Grief	Longing	Heroic Aspiration
Du Fu	3738	0.1527	0.1709	0.0654	0.0441
Li Bai	1783	0.1063	0.1175	0.0595	0.0898
Liu Changqing	819	0.1783	0.0783	0.1123	0.0245
Cen Shen	747	0.1113	0.0799	0.0870	0.0656
Dugu Ji	276	0.1242	0.0752	0.1536	0.0523
Yuan Jie	272	0.0983	0.1838	0.0342	0.0085
Gao Shi	229	0.1018	0.1228	0.0632	0.0912
Qian Qi	200	0.1602	0.1068	0.1068	0.0534

Table 8: Top-8 poets by sentence count and selected secondary emotion ratios.

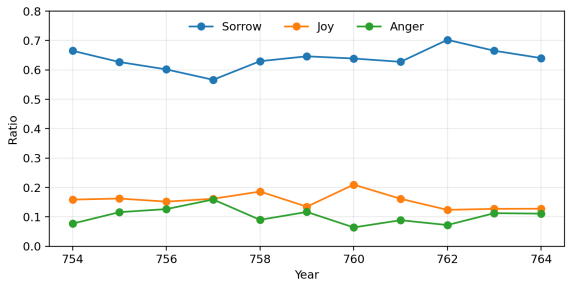


Figure 1: Year-wise trend of three core primary emotions (Sorrow, Joy, Anger).

tional downturn in our data and for the shift toward the restrained tone later described as 气骨顿衰. The table-level correspondence is direct. In Table 6, Sorrow peaks in 762 (0.7024), and Joy in 762 (0.1239) stays below its 760 peak (0.2097). In Table 7, Melancholy peaks in 762 (0.1824), while Romantic Affection reaches its minimum in 762 (0.0049). Our sentence-level count check further reports Melancholy $=$ 187 and Romantic Affection $=$ 5 in 762. At the same time, Joy and Anger do not disappear, which reminds us that crisis-era poetic emotion is mixed, not monolithic.

Our main turning-point reading is centered on **762**, a node that has been less explicitly foregrounded in prior An Lushan Rebellion emotion studies. At this point, Sorrow reaches its yearly maximum, Melancholy also peaks, and Romantic Affection falls to its minimum, producing the clearest structural shift in the corpus. By comparison, 757 is best read as the beginning of the decline in Heroic Aspiration rather than the full-system turning point: the category still appears relatively strong in the early war years, but then drops and does not recover to pre-war levels. This keeps the narrative consistent with both the annual counts and the original manuscript: 757 marks the fading of High-Tang confidence, while 762 marks the broader emotional restructuring toward grief-oriented and inward-looking expression.

Heroic Aspiration (壮思) is most visible around 755–757 and then declines, echoing the rise and fading of what Chinese literary history calls 盛唐气象. In Tang context, this category often overlaps with frontier rhetoric, loyalty discourse, and moral self-positioning. Its decline in later years, together with rising Melancholy, is consistent with a shift from mobilizing rhetoric to reflective and mournful expression.

From a literary-historical perspective, this pattern is compatible with the transition from expansive High-Tang voice to the more constrained and self-reflective tone associated with later mid-Tang poetics. Our contribution here is not to replace traditional periodization, but to provide a sentence-level quantitative trajectory that makes this transition empirically inspectable.

7.2 Poet-Level Variation and Relational Emotion Patterns

Poet-level heterogeneity is essential for humanities interpretation. Du Fu's high Melancholy+Grief profile supports readings of historically grounded poetic witness. Li Bai and Gao Shi retain relatively stronger Heroic Aspiration, matching traditions of expansive voice and frontier vigor. Thus, aggregated temporal change should be read together with poet-specific writing styles.

Another finding that echoes the original thesis concerns 思念 (Longing). In Table 8, Longing is relatively high for poets whose surviving lines in this window are less centered on direct battlefield narration (e.g., Dugu Ji and Liu Changqing). This does not imply that these poets were detached from historical crisis. Rather, it suggests that within the same war-era corpus, different social locations and communication contexts (farewell poems, correspondence, local circles) can sustain different emotional registers.

So the corpus is not emotionally homogeneous, even under severe macro-political turbulence. The coexistence of high Sorrow at year level and locally stronger Longing in some poet clusters supports a layered reading: "war pressure" structures the overall field, while regional life-worlds and poet networks still modulate the emotional surface of individual lines. This is precisely where computational aggregation and close reading become complementary. This type of cross-scale contrast is one place where the digital pipeline contributes genuinely new humanities value, because it links macro trend and micro poetic context within one auditable analysis chain.

7.3 Genre Boundaries and Methodological Implications

Romantic Affection remains low, with a clear minimum in 762. This does not imply absence of private emotion in Tang poetry; rather, within this wartime-centered corpus slice, public history, displacement, and political anxiety dominate the available space of expression. This reminds us that corpus framing matters when we make cultural claims.

For DH readers, the point is not that embeddings are "more advanced" than word outputs. The real issue is reliability of the evidence chain. If one-word outputs are prompt-unstable, downstream yearly distributions can mix true temporal signal with avoidable generation variance. By using embeddings for structure discovery and canonical labels for aggregation, we reduce this variance source and make claims easier to inspect and challenge. In that sense, representation choice directly affects humanities validity.

8 Limitations and Ethics

Interpretive Scope and Causal Boundary. We report structured correlations and trend alignments, not deterministic historical causality. Computational patterns should complement, not replace, close reading.

Data and Label Limitations. Supervised models are trained on pipeline-generated silver labels. A larger fully human-annotated benchmark is needed for stronger external validity and significance testing.

Category Construction Subjectivity. The 20-cluster discovery step is data-driven, but the merge to 18 labels includes human judgment. Another research team may make slightly different merge decisions at semantic boundaries.

Scope of Historical Generalization. Our observations are tied to the current corpus construction and year window (754–764). They should be interpreted as evidence for this textual slice, not as a full account of all Tang emotional history.

Reproducibility. All reported numbers are linked to executable outputs and logs. To support double-blind review, implementation details are not included in the main submission and can be released after acceptance.

Ethics. The study uses historical literary texts and focuses on scholarly analysis. No personal sensitive data is involved.

9 Conclusion

We analyzed emotional change in Tang poetry around the An Lushan Rebellion with a sentence-level, evidence-traceable pipeline. The paper makes three linked points. First, compared with coarse polarity-style analysis, a fine-grained embedding-centered framework better fits Tang-poetry emotion research and is more robust to LLM generation instability. Second, the resulting temporal and poet-level patterns are consistent with historical evidence, supporting the method's validity for humanities interpretation. Third, the framework surfaces layered findings (e.g., macro sorrow growth together with local longing stability) that are hard to observe in coarse setups. More broadly, the study shows a practical path for bringing digital methods into humanities work without flattening literary complexity: keep the evidence chain transparent, keep categories interpretable, and keep close reading at the center of final interpretation.

References

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Renfen Hu and Yuchen Zhu. 2015. Automatic topic classification for tang poems. *Journal of Peking University (Natural Science Edition)*, 51(2):262–268. [in Chinese].

Institute of Information Retrieval, Dalian University of Technology. 2007. Dalian university of technology emotion ontology lexicon. urlhttps://ir.dlut.edu.cn/zyxz/qgbtk.htm. [in Chinese].

Guang Sima. 1956. *Zizhi Tongjian*. Zhonghua Book Company, Beijing. [in Chinese].

Xue Xiao. 2006. Hierarchical classification of chinese text and its application to tang poetry classification. Master's thesis, Chongqing University. [in Chinese].

Wei Zhang, Hao Wang, Sanhong Deng, and Baolong Zhang. 2021. Emotion-term extraction and application for classical chinese poetry in digital humanities. *Journal of Library Science in China*, 47(4):113–131. [in Chinese].

Item	Detail
Corpus provenance	Project corpus derived from `anshi_poems.csv` and `anshi_annotated.xlsx`; yearly aggregation is computed from `anshi_analysis.xlsx`.
Coverage	616 poem records, 47 poets, 11 annual bins (754–764).
Sentence instances	9,965 sentence instances before final alignment; 10,122 annotated rows after manual line splitting.
Labels	18 fine-grained labels, 0–3 labels per sentence.
Annotation check	200-sentence sample by two annotators; OA 0.970, Cohen's $\kappa = 0.796$.

Table 9: Corpus provenance and annotation facts.

Label	Macro
羁旅	Sorrow
伤逝	Sorrow
偃蹇	Sorrow
悲愁	Sorrow
忧伤	Sorrow
思念	Sorrow
孤寂	Sorrow
讥讽	Disgust
喜悦	Joy
爱恋	Joy
壮思	Joy
淡泊	Joy
宴息	Joy
慷慨	Anger
愤恨	Anger
惊恐	Fear
赞美	Praise
迷茫	Confusion

Table 10: The 18 labels and their 7-class mapping.

A Data and Annotation Details

The 18-label inventory is the canonical annotation space; the 7-class view is only a derived reading layer for macro trends.

B Prompts

Prompt A (full text).
Source: `scripts/run_api_word_stability.py`

```
仅使用一个中文词语概括给定诗句的主要情
感。
不要解释，不要标点，不要多个词。
诗句：{line}
答案：
```

Prompt B (full text).
Source: `scripts/run_api_word_stability.py`

```
你是古典诗歌情绪标注助手。
请对下面诗句输出一个最贴切的中文情感词
（只许一个词）。
禁止解释、禁止句子、禁止多个词。
诗句：{line}
情感词：
```

18-label annotation prompt (full text).
Source: `scripts/annotate.py`

```
你是一位精通中国古典诗词的情感分析专家，
请为诗句标注 ** 最多 3 个 ** 最贴切的 ** 情
感标签 **。
# 可选的情感标签（共 18 个，必须严格使用
以下名称）：
羁旅、伤逝、偃蹇、悲愁、忧伤、思念、孤
寂、讥讽、喜悦、
爱恋、壮思、淡泊、宴息、慷慨、愤恨、惊
恐、赞美、迷茫。
# 标签定义（请严格遵守）：
- 羁旅：羁旅漂泊之苦（远行、风霜、路途艰
辛）。例：“归雁来时数附书”。
- 伤逝：年华老去的悲伤，涉及黑发、白发
等。例：“不堪玄鬓影”。
- 偃蹇：怀才不遇，壮志难酬。例：“无因见明
主”。
- 悲愁：深重人生苦难，如亲友去世、战乱、
贫穷、疾病。
例：“弟兄无一人”。
- 忧伤：内心郁结，忧思难解。例：“暮天摇落
伤怀抱”。
- 思念：思念恋人、故人（须在世）或故乡，
且归期无望。
例：“别后相思复何益”。
- 孤寂：寂寞孤单，天地独对。例：“已忍伶俜
十年事”。
- 讥讽：对人或现象的讽刺、鄙夷、厌恶。
例：“轻薄为文哂未休”。
- 喜悦：内心欢喜，轻松明朗。例：“却喜晒谷
天晴”。
- 爱恋：夫妻或恋人之间的美好感情。例：“或
恐是同乡”。
- 壮思：豪迈洒脱，人生得意。例：“冲天香阵
透长安”。
- 淡泊：超然物外，不慕名利（属积极心境）。
例：“悠然见南山”。
- 宴息：聚会饮酒、朋友欢聚之乐。例：“莫使
金樽空对月”。
- 慷慨：悲壮激昂，忧国忧民。例：“慨然抚长
剑”。
- 愤恨：强烈愤怒，痛斥不公。例：“朱门酒肉
臭”。
- 惊恐：对危险、变故的害怕与焦虑。例：“恐
惊平昔颜”。
- 赞美：对亲情、友情、爱情等真挚情谊的歌
颂。例：“天涯若比邻”。
- 迷茫：前路不明，人生困惑（重点在“不知
所措”）。例：“更欲东奔何处所”。
# 重要规则：
1. 只输出属于上述 18 个标签的内容，不得自
创。
2. 若诗句为纯写景、纯记事、语义不清 -> 返
回空列表 []。
3. 但若诗句隐含可合理推断的情绪（如“独
坐”“寒雨”“孤舟”），
请优先选择最贴近的 1 个标签，而非返回空
列表。
4. 最多标注 3 个标签，按相关性排序。
5. “泪”“愁”等字出现!= 自动打标签，需看整
体语境。
6. “独坐”“独钓”：超然 -> 淡泊；孤独无助 ->
孤寂。
# 输出要求：
- 仅输出合法 JSON，格式：{" 情感标签":[...]}
- 不要任何解释、注释、markdown 或额外字
段。
## 示例（Few-shot）：
```

诗句：” 白日依山尽”
{” 情感标签”:[]}
诗句：” 乡音无改鬓毛衰”
{” 情感标签”:[” 思念”,” 伤逝”]}
诗句：” 莫使金樽空对月”
{” 情感标签”:[” 宴息”]}
诗句：” 朱门酒肉臭，路有冻死骨”
{” 情感标签”:[” 愤恨”,” 悲愁”]}
诗句：” 独坐幽篁里，弹琴复长啸”
{” 情感标签”:[” 淡泊”]}
请标注以下诗句：
诗句：”{poem_line}”
输出（仅 JSON）：

Line-selection prompt (full text).

Source: `scripts/emopoem_get.py`

根据以下诗歌内容，分别选出这首诗中直接表达情感的诗歌单句。
诗歌内容：
{content}
请完成：
找出所有情感倾向明显外露的诗句，每次严格输出一个诗歌单句。
以逗号/句号/问号/感叹号为分界为一个诗歌单句。
不可以输出如“不睹皇居壮, 安知天子尊。”这样包含两句的诗。
请严格使用以下 JSON 格式输出（支持数组），不能在这一格式以外输出任何其他内容：
[
{”poem”:” 诗句 1”},
{”poem”:” 诗句 2”}
]

	129		125		127		153
0	得成比目何辞死	**1**	暴骨无全躯	**2**	金吾不禁夜	**3**	此地别燕丹
	153		182		146		242
4	四海遂为家	**5**	怅然临古城	**6**	玉漏莫相催	**7**	年年属数奇
	155		126		102		199
8	长作巢由也不辞	**9**	惟闻素棘与黄泉	**10**	还惊九逝魂	**11**	行来行去尽哀怜
	120		137		186		185
12	深怀国士恩	**13**	辛苦事旌麾	**14**	相逢秋月满	**15**	慷慨志犹存
	257		132		219		123
16	独有西山将	**17**	既伤千里目	**18**	生憎帐额绣孤鸾	**19**	徙倚欲何依

Table 11: Compact inventory of the 20 embedding clusters (size and one representative line each).

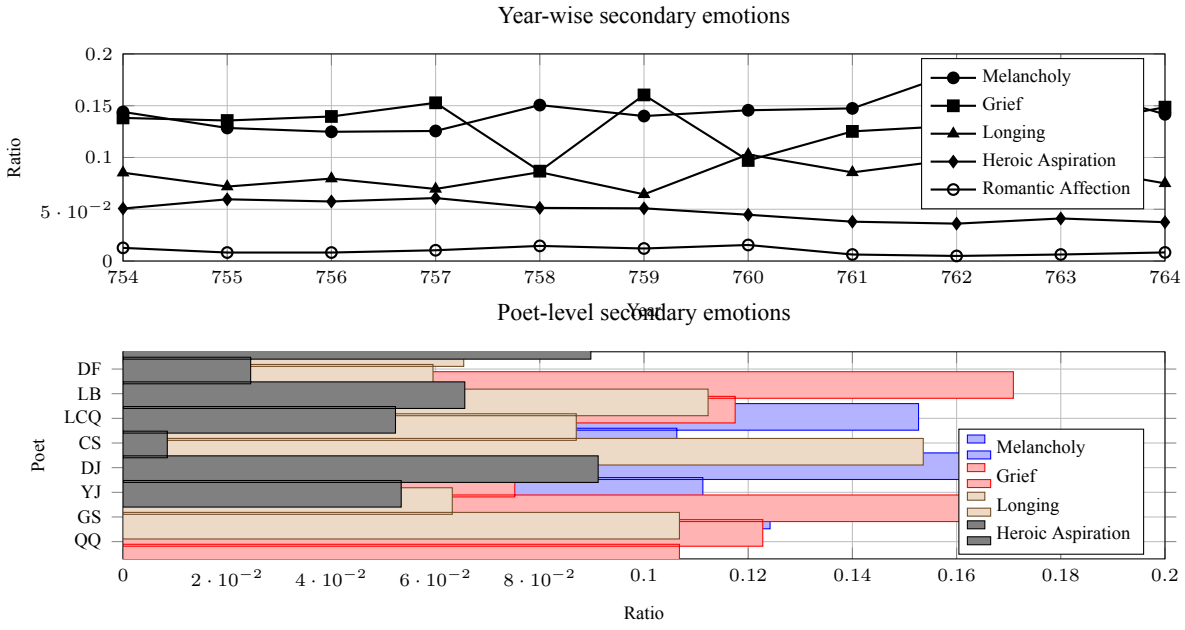


Figure 2: Visual summaries for the secondary-emotion tables in the main text.

Measuring Embedding Sensitivity to Authorial Style in French: Comparing Literary Texts with Language Model Rewritings

Benjamin Icard[1] **Lila Sainero**[1] **Alice Breton**[1]
Evangelia Zve[1,2] **Jean-Gabriel Ganascia**[1]

[1] LIP6, Sorbonne University, CNRS, France
[2] Infopro Digital, France

Abstract

Large language models (LLMs) can convincingly imitate human writing styles, yet it remains unclear how much stylistic information is encoded in embeddings from any language model and retained after LLM rewriting. We investigate these questions in French, using a controlled literary dataset to quantify the effect of stylistic variation via changes in embedding dispersion. We observe that embeddings reliably capture authorial stylistic features and that these signals persist after rewriting, while also exhibiting LLM-specific patterns. These analytical results offer promising directions for authorship imitation detection in the era of language models.

1 Introduction

Computational stylometry has long been central to digital humanities, using natural language processing (NLP) for authorship attribution and stylistic comparison (Stamatatos, 2009; Koppel et al., 2011). This has involved quantifying inter-textual distance and developing feature-based methods to support author verification (Savoy, 2012a,b; Cafiero and Camps, 2019).

In recent years, embedding methods, particularly Transformer-based contextual models (Vaswani et al., 2017; Devlin et al., 2019), have improved literary authorship attribution through richer textual representations of authorial profiles (Terreau et al., 2021; Kim et al., 2025). Yet embedding-based analysis often centers on semantic tasks, especially topic modeling (Bianchi et al., 2021) and content similarity (Rockmore et al., 2025), rather than style in embeddings, aside from a few studies (Wegmann et al., 2022; Icard et al., 2025). Assessing more systematically how embeddings capture authorial style would improve the explainability of embedding representations and strengthen authorship characterization and attribution in computational stylometry.

The increasing capabilities of large language models (LLMs) to imitate human-authored styles make this issue more urgent. Recent work on style transfer shows that LLMs can reproduce salient stylistic features in literary settings (Mikros, 2025), but also that they still struggle to imitate more implicit writing styles in extra-literary settings (Wang et al., 2025). Because LLMs are language models, this reinforces the need to assess the sensitivity of embedding vectors to style, and the extent to which LLMs preserve or alter that sensitivity during style transfer (Huang et al., 2025).

This paper reports a controlled experiment on French literary texts evaluating embedding sensitivity to target stylistic features in human-authored texts and their LLM imitations. We compile a dataset of French excerpts from Tufféry (Tufféry, 2000), Proust (Proust, 1913), Céline (Céline, 1932), and Yourcenar (Yourcenar, 1951), together with GPT-, Mistral-, and Gemini-based stylistic imitations under fixed-topic conditions. We encode all texts with thirteen embedding models and quantify aggregated stylistic effects on embedding dispersion across human authors and imitated texts.

Section 2 reviews prior work on embedding sensitivity to writing style, with particular attention to the human–LLM distinction. Section 3 introduces our French literary dataset and the evaluation of style transfer under LLM generation. Section 4 presents the vectorization of the dataset with thirteen embedding models, a structural evaluation with clustering, and results on how writing style affect embedding dispersion in human-authored texts and LLM imitations. Section 5 examines LLM-specific effects on embedding sensitivity, bringing explainability to the style transfer evaluation used to validate the dataset. Finally, Section 6 concludes and outlines directions for future work.

All reproducibility materials and results are available at: `https://github.com/sma-libra/style-embedding-sensitivity`

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 69–82
July 4, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Feature-Based Stylometry. Computational stylometry classically relies on surface features such as word frequencies, character n-grams (Cavnar and Trenkle, 1994; Ríos-Toledo et al., 2022) and punctuation patterns (Faye et al., 2024), often represented with TF-IDF term weighting (Salton and Buckley, 1988; Bui et al., 2011), to capture authorial voice (Verma and Srinivasan, 2019; Herrmann et al., 2021; Mani, 2022).

Style Embeddings. More recent work examines whether embedding-based representations also encode feature-based stylistic information (Liu et al., 2024). Terreau et al. (2021) propose an author-verification framework that tests whether embedding spaces encode stylistic features rather than mainly semantic content. Using this framework, they show that specialized Doc2Vec-based author embeddings (Le and Mikolov, 2014; Ganesh et al., 2016; Maharjan et al., 2019) are often more semantically driven, while simpler pretrained sentence encoders such as USE-DAN (Cer et al., 2018) and SBERT (Reimers and Gurevych, 2019) can perform better on stylistic feature families.

Measuring Sensitivity. A related line of work addresses style-semantics conflation more directly through topic-controlled or style-sensitive embedding representations. More specifically, Wegmann et al. (2022) show that content-controlled training improves style-topic separation in BERT-based representations. Adjacent work by Chen et al. (2023) shows that style-sensitive encoders can support efficient detection of stylistic shifts in multi-author documents. In parallel, Patel et al. (2023) introduce the style embedding model LISA whose dimensions correspond to stylistic features. Most directly related to our approach, Icard et al. (2025) use Queneau's fixed-topic variations to measure embedding sensitivity to style under rewriting by a single LLM, rather than focusing specifically on author-defining features and their preservation.

LLM Style Transfer. Recent research on LLMs' representation and high-fidelity transfer of literary style is motivated by their demonstrated capacity to imitate authorial writing. In literary contexts, Mikros (2025) show that GPT-4o reproduces salient authorial stylistic features under thematic control, while Sarfati et al. (2025) and Hicke and Mimno (2025) show that literary style is reflected in LLM internal representations. In parallel, Huang et al. (2024) study LLMs for authorship analysis, while Horvitz et al. (2024) propose TinyStyler, a lightweight approach to few-shot style transfer conditioned on author embeddings.

Taken together, these works have examined embedding sensitivity to style in smaller language models and, more recently, in LLMs. However, little work has investigated whether embeddings are *reliably* sensitive to the defining stylistic features of specific authors, and to what extent this sensitivity remains measurable *after* LLM rewriting. We investigate this question in French using stylistically marked and diverse literary texts.

3 Dataset

3.1 Text Materials

We assembled a dataset of 1,248 French literary texts that combines human-authored originals with LLM-generated imitations that reproduce the human authors' styles under fixed-topic conditions.

3.1.1 Reference Corpus

We first compiled a *reference corpus* of 384 French original literary texts, separated in two groups according to topic and style variation.

The first group, TUFFERY_REF, consists of 96 texts extracted from Stéphane Tufféry's *Le style mode d'emploi* (Tufféry, 2000).[1] We use this collection because, like Raymond Queneau's *Exercices de style* (Queneau, 1947) used in Icard et al. (2025), it keeps a single topic across all texts: the story of a bus journey in Paris. But unlike Queneau's variations, however, Tufféry structures stylistic variation more systematically, including explicit pastiches of major French authors such as Balzac, Flaubert, and Hugo. TUFFERY_REF therefore provides a fixed-topic, style-diverse baseline for stylometric comparison.

The second group, STYLE_REF, comprises 288 texts of length comparable to TUFFERY_REF, evenly distributed across three subclasses:

- PROUST_REF: 96 texts from Proust's *Du côté de chez Swann*, the first volume of *À la recherche du temps perdu* (Proust, 1913);
- CELINE_REF: 96 texts from Céline's *Voyage au bout de la nuit* (Céline, 1932);
- YOURCENAR_REF: 96 texts from Yourcenar's *Mémoires d'Hadrien* (Yourcenar, 1951).

[1] We excluded 3 of the 99 texts from Tufféry's *Le style mode d'emploi* when constructing TUFFERY_REF because they departed from the bus-journey topic.

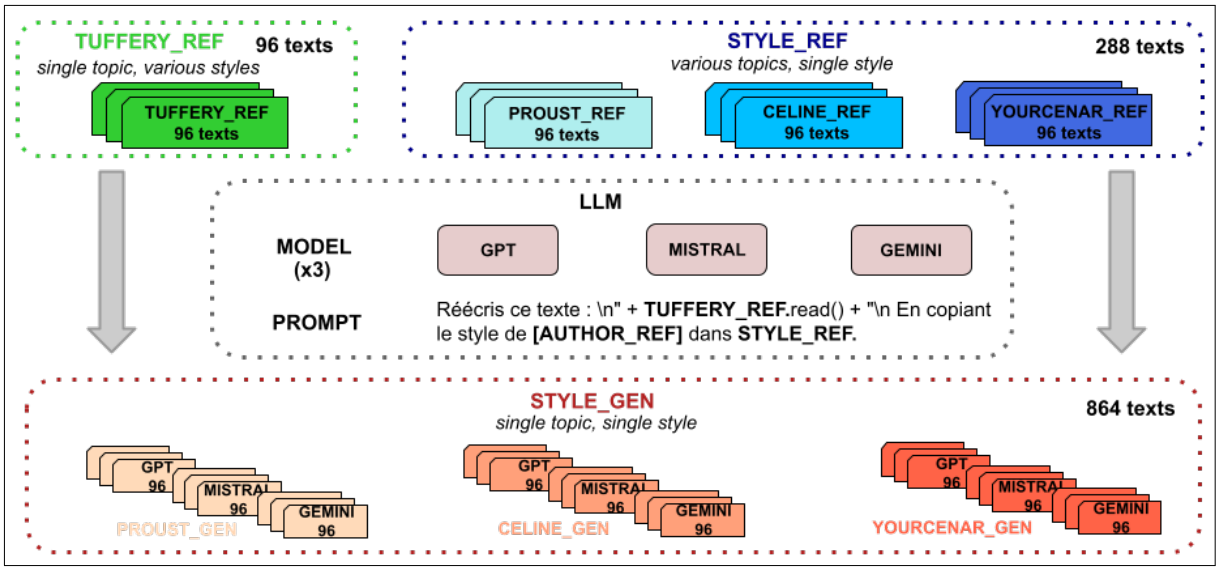


Figure 1: Textual corpora and generation scheme used to construct the STYLOGEN embedding dataset. The reference corpora are TUFFERY_REF (single topic, various styles) and STYLE_REF (various topics, single style per author). Three LLMs generate STYLE_GEN by rewriting Tufféry texts in the styles of the authors in STYLE_REF, using the French prompt shown.

The authors were selected to maximize stylistic contrast across four central dimensions in French literature (Magri-Mourgues, 2006; Brunet, 2014; Hough, 2016): structural complexity, morphosyntactic richness, lexical diversity, and referential anchoring. Proust exemplifies structural complexity (long sentences, high average word length) and morphosyntactic richness (diverse POS patterns), with recurrent proper-name allusions (named entities) (Brunet, 1982). Céline foregrounds a spoken register (short sentences and words, high pronoun/verb share) alongside lexical diversity and nonstandard forms (higher entropy, distinctive character distributions) (Zábojniková, 2006; Styhre, 2011; de Sacy, 2025). Yourcenar sustains balanced prose densely anchored in historical and geographic reference (named-entity density) (Colvin, 2005; Broche, 2022).

3.2 Generated Corpus

We constructed the generated corpus, STYLE_GEN, by prompting LLMs to rewrite the 96 texts from TUFFERY_REF in the writing styles of PROUST_REF, CELINE_REF, and YOURCENAR_REF, respectively. To ensure both diversity of behavior and architectural variety, we employed three LLMs for each target style: `GPT-4o`, `mistral-large-2411`, and `gemini-1.5-flash`. For simplicity's sake, we now refer to these LLM versions as GPT, MISTRAL, and GEMINI.

The group STYLE_GEN comprises 864 texts evenly distributed across three subclasses:

- PROUST_GEN: 288 texts produced by prompting each of the three generative models to rewrite Tufféry's texts in the style of Proust's *Du côté de chez Swann*;
- CELINE_GEN: 288 texts produced by prompting each of the three generative models to rewrite Tufféry's texts in the style of Céline's *Voyage au bout de la nuit*;
- YOURCENAR_GEN: 288 texts produced by prompting each of the three generative models to rewrite Tufféry's texts in the style of Yourcenar's *Mémoires d'Hadrien*.

Figure 1 presents an overview of the pipeline followed to assemble our textual corpora, including the French prompt used. The English translation of the prompt is: *"Rewrite this text: \n" +* **TUFFERY_REF***.read() + "\n By copying the style of* [**AUTHOR_REF**] *into* **STYLE_REF***."*

3.3 Style Transfer Evaluation

We assessed style transfer from STYLE_REF to STYLE_GEN as a function of the LLM used as the generator. Specifically, we evaluated the extent to which texts in PROUST_GEN, CELINE_GEN, and YOURCENAR_GEN were

Validation Set		PROUST_REF	CELINE_REF	YOURCENAR_REF	STYLE_REF (held-out 20%)
Corpus-Level	Macro-F1	—	—	—	0.965
	Accuracy	0.947	0.947	1.000	**0.966**

Test Set		PROUST_GEN	CELINE_GEN	YOURCENAR_GEN	STYLE_GEN (100%)
Corpus-Level	Macro-F1	—	—	—	0.669
	Transfer Accuracy	0.601	0.722	0.670	**0.664**
Per-Class	*when* GPT *is used*	0.490	0.542	0.646	0.559
	when MISTRAL *is used*	0.844	0.812	0.625	0.760
	when GEMINI *is used*	0.469	0.812	0.740	0.674

Table 1: Style transfer results with TF-IDF character 3-5-grams + `LinearSVC`. We report corpus-level performances on STYLE_REF (held-out 20%) and STYLE_GEN (100%), and per-class transfer accuracy for each target author label. We indicate corpus-level accuracy in bold, and the highest per-class transfer accuracy across LLMs in blue.

classified as PROUST_REF, CELINE_REF, and YOURCENAR_REF, respectively.

Character n-Gram Validator. For this task, we used a validator that represents each document with TF-IDF-weighted character 3-5-grams and classifies it with a linear support vector classifier (Cortes and Vapnik, 1995), implemented with `LinearSVC` from scikit-learn, into one of three author-style labels: PROUST_REF, CELINE_REF, YOURCENAR_REF. We chose this validator because character n-grams are strong surface-level features of authorial style (Kešelj et al., 2003; Stamatatos, 2009), and because it is embedding-independent, avoiding confounds in our later sensitivity analyses.[2]

Methodologically, we split STYLE_REF into train/validation sets and use STYLE_GEN as the test set:

— **Train/Validation:** using STYLE_REF only (i.e., PROUST_REF, CELINE_REF, and YOURCENAR_REF) split into 80% for train and 20% for validation (stratified by author, with seed = 42);

— **Test:** using 100% of STYLE_GEN (i.e., PROUST_GEN, CELINE_GEN, and YOURCENAR_GEN), evaluated without adaptation to measure style transfer from human to LLM-generated text.

Table 1 reports the validator's performances on the held-out STYLE_REF corpus (20%) and its *transfer* accuracy on the STYLE_GEN corpus (100%), broken down by imitated author class and by LLM used as a generator.

[2]Appendix A.1 (Table 3) reports additional style transfer results on the corpus using another linear `LinearSVC` classifier, this time with function-word frequencies as a classical stylometric baseline for authorial style (Burrows, 2002).

At the corpus level, the validator achieves near-ceiling accuracy on held-out human texts from STYLE_REF (Accuracy: 0.947 to 1.00). When applied to STYLE_GEN, transfer accuracy drops substantially, but remains well above the three-way chance level (1/3) (Accuracy: 0.664).

Transfer accuracy on generated texts varies with the LLM used for imitation. PROUST_GEN is best recognized when generated by MISTRAL (0.844), CELINE_GEN is best recognized when generated by MISTRAL or GEMINI (0.812 in both cases), and YOURCENAR_GEN is best recognized when generated by GEMINI (0.740). By contrast, GPT yields the lowest accuracies across the three generated corpora. The confusion matrix complementing these results is provided in Appendix A.2 (Figure 5).

Beyond style transfer validation from STYLE_REF to STYLE_GEN across LLM generators, we now analyze embedding representations of the same full textual corpora to compare embedding sensitivity to stylistic features before and after LLM rewriting.

4 Embedding Sensitivity to Writing Style

4.1 Dataset Vectorization

Embedding Model Selection. The reference corpora TUFFERY_REF and STYLE_REF, along with the generated corpus STYLE_GEN, were embedded using thirteen models (Table 2), with their embedding dimensionalities reported. Selection criteria were architectural and dimensional diversity, computational efficiency, and strong performance on the Massive Text Embedding Benchmark (MTEB; November 2025 version)[3] (Muennighoff et al., 2022). The resulting embedding dataset for all thirteen model is available in our data repository.

[3]https://huggingface.co/spaces/mteb/leaderboard

Clustering Validation. To assess embedding space separability across the three corpora (TUFFERY_REF, STYLE_REF, STYLE_GEN), we run k-means clustering (Hartigan and Wong, 1979) on the full-dimensional embeddings of the thirteen models, setting $k = 3$.[4] Clustering quality is evaluated using purity (Manning, 2008), an external score in $[0, 1]$ computed by comparing the induced clusters to the three ground-truth corpus labels (Soni and Dwivedi, 2024). Table 2 reports, for each embedding model, the purity score computed in the model's full-dimensional space (FullD).

Embedding Model	FullD	Purity
`xlm-roberta-large`	1024	0.7654
`multilingual-e5-large`	1024	0.6836
`mistral-embed`	1024	0.6809
`e5-base-v2`	768	0.6802
`distilbert-base-uncased`	768	0.6701
`text-embedding-3-small`	1536	0.6663
`text-embedding-004`	768	0.6539
`solon-embeddings-large-0.1`	1024	0.6524
`voyage-2`	1024	0.6493
`sentence-camembert-base`	768	0.6408
`all-roberta-large-v1`	1024	0.6242
`paraphrase-multilingual-mpnet-base-v2`	768	0.6080
`all-MiniLM-L12-v2`	384	0.5721

Table 2: Embedding dimensionality and k-means cluster purity ($k = 3$) for each model in full-dimensional space (FullD), computed with respect to the three corpus labels (TUFFERY_REF, STYLE_REF, STYLE_GEN). Models are ordered by purity scores.

Recovery of the three corpus groups is reasonable across the thirteen embedding models, with mean purity (0.6575) and a close median (0.6539). At the level of individual models, `xlm-roberta-large` scores highest (0.7654) and `all-MiniLM-L12-v2` lowest (0.5721), with intermediate models spread smoothly across this range.

UMAP Reduction. Beyond the consistent purity scores across models, Table 2 shows substantial variation in embedding dimensionality (from 384 to 1536). To enable comparisons across model-specific full-dimensional spaces (FullD), we apply Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). For each target dimensionality, we repeat the projection thirty times with different random seeds to account for UMAP's stochasticity.

Focus on 2D UMAP. We tested UMAP reductions, specifically 2D, 3D, and 10D, to find the optimal sufficient dimension with respect to numerical fidelity to the FullD purity scores. To quantify alignment, we computed the mean absolute error (MAE) to FullD, as well as the maximum absolute error (MaxAE).

Among the UMAP reductions, we observed that 2D performed best, with the lowest MAE (0.025 for 2D, vs 0.036 for 3D and 0.029 for 10D) but also the smallest MaxAE (0.056 for 2D, vs 0.090 for 3D and 0.066 for 10D), making it the reduction best aligned with FullD purity. To visualize corpus-group structure, Figure 2 shows the 2D UMAP projection of the `xlm-roberta-large` embedding model, which, as in FullD, obtains the best clustering purity score (0.7627).

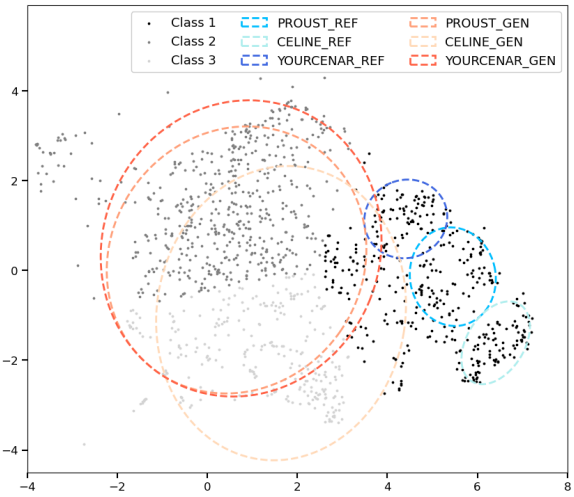


Figure 2: 2D UMAP projection of `xlm-roberta-large` embeddings on the dataset. Points indicate the three k-means clusters (Class 1-3), while dashed ellipses (visual guides, not k-means clusters) indicate label-based coverage regions for the human and generated corpora, drawn so that exactly 80% of the corpus lie inside the ellipse zone.

The dashed ellipses in Figure 2 suggest possible stylistic effects on the embedding representations. For PROUST_REF, CELINE_REF, and YOURCENAR_REF (right side of the figure), the clear separation between ellipses is consistent with differences in both topic and style across the human corpora, whereas the strong within-ellipse cohesion cannot be explained by topic, since topic varies within each corpus, and may instead reflect the homogenizing effect of shared authorial style. For PROUST_GEN, CELINE_GEN, and YOURCENAR_GEN (left side of the figure), the substantial overlap is consistent with intended topic alignment on TUFFERY_REF, but the overlap remains only partial, especially for CELINE_GEN,

[4] Using *scikit-learn* `KMeans` with default parameters and `random_state=0`.

suggesting that embeddings may retain residual stylistic differences after LLM rewriting.

These observations motivate measuring whether embeddings capture stylistic features beyond topic under fixed-topic rewriting. We now examine embedding sensitivity to author-characteristic features in French, both before and after LLM imitation.

4.2 Sensitivity Evaluation Metrics

Following Icard et al. (2025), we compute dispersion-based metrics on an aggregated 2D UMAP reduction derived from the thirteen embedding models listed in Table 2.

Embedding Dispersion. For the j-th UMAP iteration, we define $d_X^{(i,j)}$ as the Euclidean distance of the i-th embedding vector from the centroid $c_X^{(j)}$ of class X as follows:

$$d_X^{(i,j)} = \|v_X^{(i,j)} - c_X^{(j)}\| \quad (1)$$

where $v_X^{(i,j)}$ is the i-th embedding vector of class X in the j-th iteration and $\|\cdot\|$ is the Euclidean norm.

To capture the spatial dispersion of embeddings in the UMAP target space, we calculate the mean Euclidean distance from the centroid of each class X across all 30 UMAP iterations and all embeddings N in X, written $\bar{d}_X$:

$$\bar{d}_X = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{30}\sum_{j=1}^{30} d_X^{(i,j)}\right) \quad (2)$$

Target Stylistic Features. To analyze embedding sensitivity to stylistic variation in human texts and LLM rewritings, we must retain features that reflect the author-characteristic dimensions associated with Proust, Céline, and Yourcenar, as described in Section 3.1.1.

To this end, we drew on the stylometric framework proposed by Terreau et al. (2021), which is based on a comprehensive inventory of eight stylistic features families,[5] from which we retained a subset of five (*structural features*, *part-of-speech tags*, *indexes of lexical diversity*, *letter frequencies*, and *named entities*) most relevant to authorial style along the core dimensions previously considered:

- **Structural:** mean values of selected *structural features*, specifically word length and sentence length, normalized by text length;
- **Morphosyntax:** frequency of *part-of-speech tags* (nouns, verbs, adjectives) to capture grammatical diversity patterns;
- **Lexicon:** *lexical diversity* (using Shannon Entropy, see Shannon 1948) and *letter*-level patterns (character unigram and capitalization frequencies);
- **Referentiality:** density of *named entities* (e.g., persons, locations, organizations) per sentence, as a proxy for referential content.

For convenience, we write $f_X^s(i)$ to denote the average frequency value f of the stylistic feature s measured on the i-th document of class X.

Dispersion-Style Correlations. In the following analyzes, TUFFERY_REF serves as the reference class to study interactions between embedding dispersion and writing style on the corpus.

We denote by $\Delta d(\text{TUFFERY_REF}, Y)$ the difference in embedding dispersion between TUFFERY_REF and another class Y (human or generated, here), and by $\Delta f^s(\text{TUFFERY_REF}, Y)$ the difference in frequency of stylistic feature s between the two classes. More formally:

$$\Delta d(\text{TUFFERY_REF}, Y) = d_{\text{TUFFERY_REF}}(i) - d_Y(j) \quad (3)$$

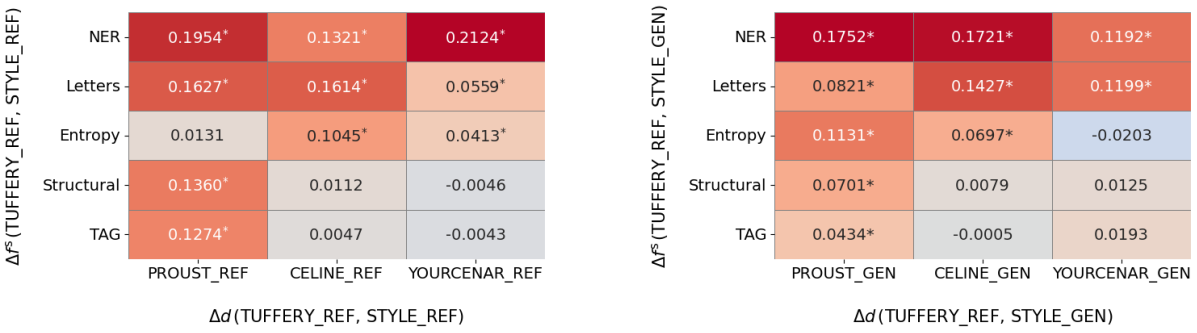
$$\Delta f^s(\text{TUFFERY_REF}, Y) = f^s_{\text{TUFFERY_REF}}(i) - f^s_Y(j) \quad (4)$$

where the comparison class Y is either a reference class $Y \in$ STYLE_REF or a generated class $Y \in$ STYLE_GEN, i is the i-th vector of TUFFERY_REF, and j is the j-th vector of class Y.

To evaluate embedding sensitivity to writing style, whether human (STYLE_REF) or model-generated (STYLE_GEN), we compute Pearson correlations, written r, between Δd and Δf^s for each stylistic feature s, relative to TUFFERY_REF.

Interpretation. We interpret the Δd–Δf^s correlations as distribution-level associations between dispersion shifts and stylometric shifts across corpora, rather than as causal estimates of rewriting effects (i.e., the indices i and j need not refer to paired source–rewrite items). In particular, for STYLE_GEN, the topic is controlled by construction via rewriting of TUFFERY_REF, so remaining variation is expected to reflect stylistic transfer and generator-specific rewriting behaviour specifically.

[5] The complete list and description of the eight feature families are available at: https://github.com/EnzoFleur/style_embedding_evaluation/.



(a) (b)

Figure 3: Pearson correlations (r) between embedding dispersion shifts in 2D UMAP reduction (Δd) and stylistic feature shifts (Δf^s) for each author-labeled corpus, comparing TUFFERY_REF with (a) the three human-authored corpora in STYLE_REF, and (b) the three style-imitated corpora in STYLE_GEN. Asterisks indicate $p < 0.01$ after Bonferroni correction.

4.3 Embedding Sensitivity to Authorial Style

To assess embedding sensitivity to human authorial style, we computed the 2D UMAP Δd–Δf^s correlations for comparisons between TUFFERY_REF and each author corpus PROUST_REF, CELINE_REF, and YOURCENAR_REF. Figure 3a reports the sensitivity correlations obtained for each author corpus, broken down by stylistic feature family. For transparency, we report the FullD correlations in Appendix A.3 (Figure 6a).[6]

For each human author, we observe moderate-to-weak yet significant correlations between embedding-dispersion shifts and stylistic features. PROUST_REF shows broad sensitivity across NER ($r = 0.195^*$), Letters ($r = 0.163^*$), Structural features ($r = 0.136^*$), and TAG ($r = 0.127^*$). CELINE_REF concentrates on Letters ($r = 0.161^*$), NER ($r = 0.132^*$), and Entropy ($r = 0.105^*$). Finally, YOURCENAR_REF is dominated by sensitivity to NER ($r = 0.212^*$) with minimal Letters and Entropy contributions. Overall, these correlations align with the expected authorial stylistic profiles of Proust (syntactic and structural complexity, morphosyntactic richness, proper-names allusions), Céline (spoken register with lexical unpredictability and nonstandard forms), and Yourcenar (strong referential density) described in subsection 3.1.1.

However, topic variation between STYLE_REF and TUFFERY_REF may confound these results. To isolate style from topic, we now examine STYLE_GEN, where topic is aligned with TUFFERY_REF while style varies through generative rewriting.

[6] The full set of raw Pearson correlations for 3D and 10D UMAP, in addition to 2D UMAP and FullD, is available in our GitHub repository.

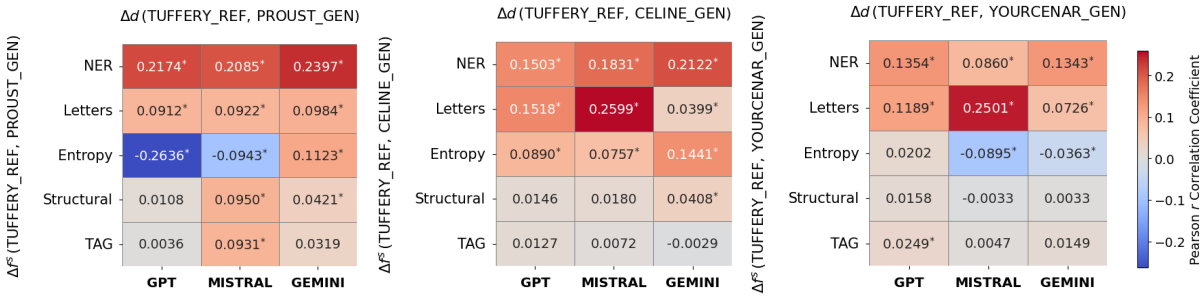
4.4 Sensitivity After LLM Rewriting

To assess embedding sensitivity in the context of LLM imitations, we computed the 2D UMAP Δd–Δf^s correlations for comparisons between TUFFERY_REF and each imitated-author corpus PROUST_GEN, CELINE_GEN, and YOURCENAR_GEN. Figure 3b reports the sensitivity correlations for each imitated-author corpus, broken down by stylistic feature family. For transparency again, we report the FullD sensitivity correlations in Appendix A.3 (Figure 6b).[7]

We observe that PROUST_GEN shows highest embedding sensitivity to NER ($r = 0.175^*$) and Entropy ($r = 0.113^*$), with weaker effects for Letters and Structural. CELINE_GEN concentrates on NER ($r = 0.172^*$) and Letters ($r = 0.143^*$). YOURCENAR_GEN emphasizes Letters ($r = 0.120^*$) and NER ($r = 0.119^*$).

Overall, STYLE_GEN exhibits moderate stylistic fidelity to the main authorial stylistic dimensions observed in STYLE_REF, but these dimensions are attenuated and reweighted. PROUST_GEN maintains the broad PROUST_REF stylistic profile in compressed form, with NER only slightly reduced and stronger reductions in Letters, Structural, and TAG. CELINE_GEN shows the highest stylistic fidelity among the generated classes, preserving spoken-style register of CELINE_REF (persistent Letters and weaker Entropy) while overweighting referentiality (NER). YOURCENAR_GEN retains the NER-based referential anchoring of YOURCENAR_REF, but more weakly, with partial dilution and a shift toward Letters.

[7] As before, the full set of raw Pearson correlations for all target dimensions is available in our GitHub repository.



(a) (b) (c)

Figure 4: Pearson correlations (r) between embedding dispersion shifts in 2D UMAP reduction (Δd) and stylistic feature shifts (Δf^s) per LLM imitator, comparing TUFFERY_REF with (a) PROUST_GEN, (b) CELINE_GEN, and (c) YOURCENAR_GEN. Asterisks indicate $p < 0.01$, with Bonferroni correction.

5 Discussion

LLM Stylistic Fidelity. Figure 4 reports LLM-level sensitivity for the three STYLE_GEN corpora and reveals author-specific preservation patterns hidden in the pooled results (Figure 3).

Figure 4a shows that MISTRAL achieves the highest stylistic fidelity for PROUST_GEN, preserving Proust's broad profile as measured in PROUST_REF. (NER, Letters, Structural, TAG). By contrast, Figure 4b reveals that GPT achieves higher stylistic fidelity for CELINE_GEN, best preserving Céline's core spoken-style pattern as observed in CELINE_REF (NER, Letters, Entropy). Finally, Figure 4c shows that GPT yields greater style fidelity to Yourcenar's defining features in case of YOURCENAR_GEN (NER, Letters).

Validation vs Stylistic Fidelity. These results on LLMs' fidelity to human style only partially align with the style transfer validation results reported in subsection 3.3. When evaluating style transfer from STYLE_REF to STYLE_GEN (Table 1), we obtained the highest transfer accuracy for Proust using MISTRAL, for Céline using MISTRAL and GEMINI, and for Yourcenar using GEMINI. However, when assessing stylistic fidelity to the target author profile by LLM (Figure 4), we find that while MISTRAL still best preserves Proust's broad stylistic profile, GPT shows higher stylistic fidelity for Céline than both MISTRAL and GEMINI, and higher stylistic fidelity for Yourcenar than GEMINI.

Explaining Metric Mismatch. The divergence stems from the two evaluations emphasizing different, though overlapping, stylistic features. Stylistic fidelity measures embeddings preservation of the broad author stylistic profile, while the 3-5-gram-based validator measures surface features, most directly Letters and indirectly NER and TAG.

For Proust, rankings align because MISTRAL best preserves the Letters-level features that dominate the validator, as well as the target author profile (NER, Letters, Structural, TAG). For Céline, MISTRAL and GEMINI obtain higher style transfer accuracy, consistent with their emphasizing Letters-level surface markers that are highly validator-salient, while GPT better preserves Céline's author-defining pattern (NER, Letters, Entropy). For Yourcenar, GEMINI similarly achieves higher validator accuracy by matching Letters-level features, whereas GPT better preserves Yourcenar's defining combination of NER and Letters.

6 Conclusion

Our study reveals moderate-to-weak embedding sensitivity to authorial style in French. Comparing literary texts from three authors with their LLM imitations, we find that embeddings broadly reflect author-characteristic stylistic patterns, although the preservation of specific features varies across authors and LLMs. By analyzing stylistic shifts in embedding space, our approach complements style transfer evaluation by tracing how specific features are preserved or altered by generative rewriting.

These findings remain preliminary and require replication across additional languages, broader author sets, and alternative stylistic features. Further analyses should also move beyond aggregated representations to examine more directly how individual embedding architectures encode stylistic features. More broadly, our results suggest that deviations from the human baseline after generative rewriting may help identify LLM-based imitation of authorial style through stylistic transformation patterns. We leave these questions for future work.

Limitations

This study has four main limitations. First, the analyses rely primarily on 2D UMAP because embedding dimensionality varies across models in FullD and because 2D yields the best alignment with both authorial style preservation and FullD purity scores. However, the distortions of authorial features observed in these dimensions require further investigation.

Second, the inventory of stylistic features we consider (NER, Letters, Entropy, Structural, POS Tags) is tailored to the current author sample rather than exhaustive, which may limit coverage of authorial stylistic signals beyond surface features.

Third, the stylistic correlations observed are consistently significant but moderate in magnitude and sometimes limited, calling for broader validation to ensure robustness.

Finally, our study focuses on French literary texts, so generalization to other languages and to non-literary materials in which style is also salient (e.g., correspondence, personal journals, speeches, periodicals), is needed.

Ethical Considerations

This work follows the principles of open science, AI transparency, and sustainability, with a strong emphasis on reproducibility and public access to results (lawful use under the EU text-and-data mining exception, Directive 2019/790/EC, Art. 3;[8] in the U.S. context, comparable research use would fall under transformative fair use[9]).

To support open science while respecting copyright regarding the text materials, we release only the 864 LLM-generated rewritings of Tufféry's texts *in-the-style-of* the three other authors (Proust, Céline, Yourcenar). These constitute transformative stylistic imitations consistent with fair use principles. We do not distribute the 384 original literary texts due to copyright restrictions. We do, however, release the full set of 16,224 vector embeddings (= 1,248 texts $\times$ 13 embedding models) (research-only license; reconstruction or re-identification attempts prohibited). These embeddings are provided for research use only, and we do not evaluate embedding inversion risk in this work.

To advance AI transparency, our GitHub repository releases all code, the 864 LLM-generated rewritings of Tufféry's texts, full set of 16,224 embeddings, and analytical results, in particular raw unadjusted Pearson correlations, with documentation for reproducibility.

To promote sustainability, we used the MTEB leaderboard (Hugging Face) to combine smaller and larger open-source pretrained models in a way that helps limit carbon emissions.

Acknowledgements

We thank Stéphane Tufféry for authorizing us to use the material from his book *Le style mode d'emploi* in our study, as well as two anonymous reviewers for helpful comments and feedback. This work was supported by the programs THEMIS (grant agreements n°DOS022279400 and n°DOS022279500) and TRUSTEDNEWS (ANR-25-ASM2-0003). EZ acknowledges Infopro Digital for supporting her PhD research, alongside her work.

Declaration of Contribution

BI conceptualized the research problem and designed the experiment. LS and AB managed the data collection and generation processes. LS was responsible for coding and testing the selected generation models, while BI managed the style evaluation models. BI analyzed and discussed the results, with LS and JGG. BI wrote the paper, which was read and revised collaboratively by all authors. Correspondence: benjamin.icard@lip6.fr.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Laurent Broche. 2022. Yourcenar et *Mémoires d'Hadrien* parmi les historiens. *Anabases*, 36:131–154.

Étienne Brunet. 1982. Le style de proust dans la recherche du temps perdu. etude quantitative. In *VII International Symposium of the Association for Literary and Linguistic Computing*, volume 3, pages 51–76. Giardini Editori.

[8] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L0790

[9] https://www.law.cornell.edu/uscode/text/17/107

Étienne Brunet. 2014. La lexicométrie française : naissance, évolution et perspectives. *Revue de l'Université de Moncton*, 45(1–2):13–33.

Quang Anh Bui, Muriel Visani, Sophea Prum, and Jean-Marc Ogier. 2011. Writer identification using tf-idf for cursive handwritten word recognition. In *2011 International Conference on Document Analysis and Recognition*, pages 844–848. IEEE.

John Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.

Florian Cafiero and Jean-Baptiste Camps. 2019. Why molière most likely did write his plays. *Science Advances*, 5(11):eaax5489.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV, USA.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and 1 others. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Haoyang Chen, Zhongyuan Han, Zengyao Li, and Yong Han. 2023. A writing style embedding based on contrastive learning for multi-author writing style analysis. In *CLEF 2023 Working Notes*. CEUR Workshop Proceedings.

Michael E. Colvin. 2005. *Baroque Fictions: Revisioning the Classical in Marguerite Yourcenar*, volume 271 of *Faux Titre*. Brill / Rodopi, Leiden.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Louis-Ferdinand Céline. 1932. *Voyage au bout de la nuit*. Denoël et Steele.

Antoine Silvestre de Sacy. 2025. Hypersegmentation du discours chez louis-ferdinand céline. *Corpus*, (27).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, and Paul Égré. 2024. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 62–72, Malta.

J Ganesh, Soumyajit Ganguly, Manish Gupta, Vasudeva Varma, and Vikram Pudi. 2016. Author2vec: Learning author representations by combining content and link information. In *WWW (Companion volume)*, pages 49–50.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

J Berenike Herrmann, Arthur M Jacobs, and Andrew Piper. 2021. Computational stylistics. *Handbook of Empirical Literary Studies*, pages 451–486.

Rebecca M. M. Hicke and David Mimno. 2025. Looking for the inner music: Probing llms' understanding of literary style. *Computational Humanities Research*, 1:e3.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390.

Carole Hough, editor. 2016. *The Oxford Handbook of Names and Naming*. Oxford Handbooks in Linguistics. Oxford University Press, Oxford.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Benjamin Icard, Evangelia Zve, Lila Sainero, Alice Breton, and Jean-Gabriel Ganascia. 2025. Embedding style beyond topics: Analyzing dispersion effects across different language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3511–3522, Abu Dhabi, UAE. Association for Computational Linguistics.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264.

Junghwan Kim, Haotian Zhang, and David Jurgens. 2025. Leveraging multilingual training for authorship representation: Enhancing generalization across languages and domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34867–34892, Suzhou, China. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Chang Liu, Zhongyuan Han, Haoyang Chen, and Qingbiao Hu. 2024. Team liuc0757 at pan: A writing style embedding method based on contrastive learning for multi-author writing style analysis. *Working Notes of CLEF*.

Véronique Magri-Mourgues. 2006. Corpus et stylistique. *Corpus*, (5). Special issue.

Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González, and Thamar Solorio. 2019. Jointly learning author and annotated character n-gram embeddings: A case study in literary text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*.

Inderjeet Mani. 2022. *Computational modeling of narrative*. Springer Nature.

Christopher D Manning. 2008. *Introduction to information retrieval*. Cambridge University Press.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

George Mikros. 2025. Beyond the surface: stylometric analysis of gpt-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities*, 40(2):587–600.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Marcel Proust. 1913. *Du côté de chez Swann*, volume 1, chapter 1. Gallimard.

Raymond Queneau. 1947. Exercices de style: Edition gallimard. *Collection Folio*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Germán Ríos-Toledo, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. Detection of changes in literary writing style using n-grams as style markers and supervised machine learning. *Plos one*, 17(7):e0267590.

Daniel Rockmore, Jiayi Chen, Mohammad Javad Latifi Jebelli, Allen Riddell, and Harrison Stropkay. 2025. On the literary landscapes of vector embeddings. *Computational Humanities Research*, 1:e18.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Raphaël Sarfati, Haley Moller, Toni J. B. Liu, Nicolas Boulle, and Christopher Earls. 2025. What's in a prompt? language models encode literary style in prompt embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24059–24068, Suzhou, China. Association for Computational Linguistics.

Jacques Savoy. 2012a. Authorship attribution: A comparative study of three text corpora and three languages. *Journal of Quantitative Linguistics*, 19(2):132–161.

Jacques Savoy. 2012b. Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems*, 30(2):12:1–12:30.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Urvashi Soni and Sunita Dwivedi. 2024. Clutching of clustering validation criteria. *International Journal of Future Computer and Communication*, 13(1).

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Alexander Styhre. 2011. Céline and the aesthetics of hyperbole: Style, points, parataxis and other literary devices. *ephemera: theory & politics in organization*, 11(3):258–270.

Enzo Terreau, Antoine Gourru, and Julien Velcin. 2021. Writing style author embedding evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 84–93. Association for Computational Linguistics.

Stéphane Tufféry. 2000. Le style mode d'emploi. *Paris: Cylibris*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Gaurav Verma and Balaji Vasan Srinivasan. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*.

Zhengxiang Wang, Nafis Irtiza Tripto, Solha Park, Zhenzhen Li, and Jiawei Zhou. 2025. Catch me if you can? not yet: LLMs still struggle to imitate the implicit writing styles of everyday authors. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10040–10055, Suzhou, China. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268. Association for Computational Linguistics.

Marguerite Yourcenar. 1951. *Mémoires d'Hadrien*. Plon.

Hviezdoslava Zábojniková. 2006. Louis-ferdinand céline et l'oral populaire. *Verbum: Analecta Neolatina*, 8(1):117–125.

Validation Set		PROUST_REF	CELINE_REF	YOURCENAR_REF	STYLE_REF (held-out 20%)
Corpus-Level	Macro-F1	—	—	—	0.825
	Accuracy	0.737	0.789	0.950	**0.828**

Test Set		PROUST_GEN	CELINE_GEN	YOURCENAR_GEN	STYLE_GEN (100%)
Corpus-Level	Macro-F1	—	—	—	0.528
	Transfer Accuracy	0.403	0.438	0.760	**0.534**
Per-Class	*when* GPT *is used*	0.260	0.385	0.854	0.500
	when MISTRAL *is used*	0.604	0.646	0.625	0.625
	when GEMINI *is used*	0.344	0.281	0.802	0.476

Table 3: Style transfer results with Function words Frequencies + `LinearSVC`. We report corpus-level performances on STYLE_REF (held-out 20%) and STYLE_GEN (100%), and per-class transfer accuracy for each target author label. We indicate corpus-level accuracy in bold, and the highest per-class transfer accuracy across LLMs in blue.

A Appendix

A.1 Style Transfer Evaluation Using Function Words

Table 3 reports style transfer performances obtained on the corpus with the validator using function word frequencies in combination to `LinearSVC`.

At the corpus level, this validator yields weaker but overall consistent corpus-level results compared to 3-5-gram-based validator, with lower performance on both STYLE_REF (0.828 vs. 0.966 accuracy; 0.825 vs. 0.965 Macro-F1) and STYLE_GEN (0.534 vs. 0.664 transfer accuracy; 0.528 vs. 0.669 Macro-F1), while preserving the same contrast between validation on reference texts and weaker transfer on generated texts.

Across reference authors (PROUST_REF, CELINE_REF, YOURCENAR_REF), both validators agree in assigning the highest accuracy to Yourcenar (0.950 vs. 1.000), although the function-word-based model is notably less accurate for Proust and Céline.

Across imitated author (PROUST_GEN, CELINE_GEN, YOURCENAR_GEN) and LLM combinations, the picture is partly stable and partly shifted: MISTRAL remains the best model for PROUST, and also remains best for CELINE, though without the earlier tie with GEMINI. For YOURCENAR, however, the best system shifts from GEMINI under the character *n*-gram validator to GPT under the function-word-based validator, while GEMINI remains strong.

A.2 Confusion Matrix for the 3-5-Gram-Based Validator

Figure 5 complements the aggregate performance results given in Table 1 for the 3-5-gram-based validator, showing the class-specific error structure on STYLE_GEN. The dominant errors are mutual confusions between PROUST_GEN and YOURCENAR_GEN (107 PROUST_GEN instances predicted as YOURCENAR_REF, and 73 YOURCENAR_GEN instances predicted as PROUST_REF), whereas misclassifications into CELINE_REF are comparatively rare for the other two styles (8 and 22).

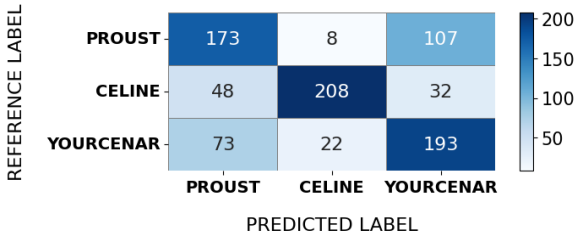
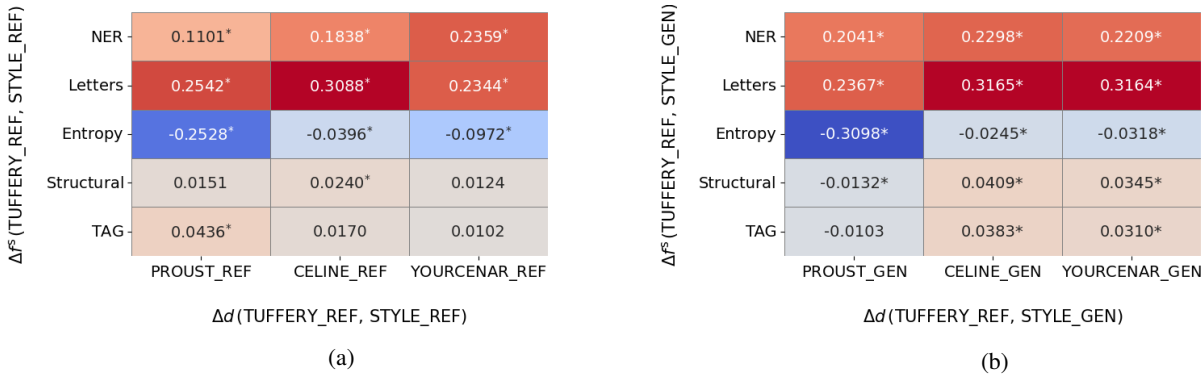


Figure 5: Confusion matrix for the STYLE_GEN evaluation of the TF-IDF character 3-5-gram + `LinearSVC` classifier trained on STYLE_REF, shown by imitated authorial style. Rows correspond to true author labels and columns to predicted labels.

A.3 FullD Sensitivity Correlations to Style

Figure 6 reports sensitivity correlations between changes in stylistic features and changes in FullD embedding dispersion across authors, comparing TUFFERY_REF with both human-written corpora (STYLE_REF) and style-imitated (STYLE_GEN) corpora.

Compared to 2D UMAP (Figure 3a), Figure 6a shows that many feature-wise correlations remain significant in FullD and some are actually stronger than in 2D UMAP, but the resulting profiles are less faithful to the authorial characteristics of the human authors. For PROUST_REF, the expected broad profile becomes distorted: Letters dominates, while Structural is nearly absent and NER is reduced. For CELINE_REF, the expected concentration on Letters, NER, and Entropy is weakened by a reversal



(a) (b)

Figure 6: Pearson correlations (r) between embedding dispersion shifts in FullD (Δd) and stylistic feature shifts (Δf^s) for each author-labeled corpus, comparing TUFFERY_REF with (a) the three human-authored corpora in STYLE_REF, and (b) the three style-imitated corpora in STYLE_GEN. Asterisks indicate $p < 0.01$ after Bonferroni correction.

of Entropy and the appearance of Structural. For YOURCENAR_REF, the expected referential dominance of NER becomes less distinctive because Letters rises to a comparable level and Entropy turns negative. Overall, compared with 2D UMAP, the FullD correlations differentiate the three human authors less clearly because they overweight shared features, especially Letters and, to a lesser extent, NER, while weakening the author-specific balance of stylistic features.

Compared to 2D UMAP (Figure 3b), Figure 6b shows that these distortions become even stronger after LLM rewriting. In PROUST_GEN, the profile narrows to NER and Letters, while Entropy becomes strongly negative and Structural and TAG largely disappear. In CELINE_GEN, Letters and NER remain prominent, but Entropy again turns negative and Structural and TAG become more visible than expected for Céline. In YOURCENAR_GEN, the profile becomes flatter and less specific than in 2D, with strong Letters and NER accompanied by positive Structural and TAG and negative Entropy.

FullD therefore preserves authorial characteristic features less well than 2D UMAP, both for human-authored texts and after LLM rewriting, because the more selective feature-level preservation observed in 2D is replaced by a more generic pattern shared across corpora. Our GitHub repository shows similar distortions with the 3D and 10D UMAP reductions.

Prompting the Past: Linguistic Transformations and Cultural Accuracy in AI-Generated Image Reconstructions for Multivocal Cultural Heritage

Ravini Wimalasuriya[1] **Lea Krause**[2] **Gert-Jan Burgers**[1]
[1]Faculty of Humanities [2]Faculty of Science
Vrije Universiteit Amsterdam, The Netherlands
raviniwim@gmail.com {l.krause, g.l.m.burgers}@vu.nl

Abstract

This research explores the intersection of cultural heritage and Generative AI (Gen-AI), examining AI-generated historical image reconstructions as a potential tool for visualising multiple perspectives in heritage interpretation. In critical heritage studies, the concept of multivocality or polyvocality advocates for representing diverse, often underrepresented, perspectives in how heritage is understood and communicated. We evaluated three prominent AI image generation models across three heritage test cases. A total of 13 user prompts generated 39 images, which underwent both linguistic analysis of intermediate prompt transformations and systematic visual assessment by heritage experts for historical accuracy and cultural sensitivity. The findings revealed both strengths and limitations of the models. While the models produced visually compelling outputs and, in some cases, meaningfully distinct depictions across perspectives, they also exhibited representation imbalances, neutralisation and amplification tendencies, inconsistencies in human portrayal, and misinterpretations introduced during the linguistic transformation of user inputs. Based on these findings, we propose initial guidelines for structured prompt construction that target the specific failure patterns identified. The research suggests that generative AI could serve as a supplementary tool, not a definitive historical source, for exploring multivocal heritage interpretation, particularly in museum and visitor engagement contexts, provided it is used critically and in conjunction with expert input.

1 Introduction

The emergence of critical heritage studies in the late 20th century challenged traditional heritage discourses that predominantly reflected state and institutional authority. This shift opened pathways for focusing on the diversity lens of the factual past, the diverse perspectives, which were often

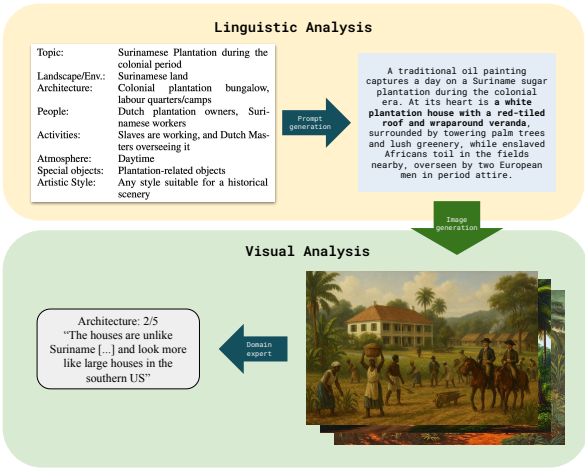


Figure 1: Overview of the research workflow. User-provided keywords across eight cultural criteria (left) are transformed by DALL-E 3 into an expanded model-generated prompt (top right), which is then used to generate an image (right). The resulting image undergoes two parallel analyses: a linguistic analysis examining how the intermediate prompt transformation altered the user's input, and a visual analysis in which a heritage domain expert rates the image's historical accuracy across the same eight criteria on a 1–5 scale. The example shown depicts the Dutch visitor's perspective for the Surinamese plantation test case.

marginalised or silenced due to power dynamics, and the contemporary socio-economic forces that shaped the appropriation of heritage and its interpretation. This context supported the emergence of the concept 'Multivocality' or 'Polyvocality', allowing previously oppressed or overlooked voices to be heard in heritage interpretation.

The existence of multiple voices for heritage sites, objects, and histories is often shaped by nationality, ethnicity, class, gender, education, culture and other forms of human identity (Derrida, 1994; Kojan, 2008; Deumert, 2018; Smith, 2020). Accordingly, when knowledge is co-created with stakeholders and communities, rather than solely with heritage experts, it fosters dialogue and un-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 83–95
July 4, 2026 ©2026 Association for Computational Linguistics

derstanding among those with different relationships to the past (Franke et al., 2024). Moreover, truthful heritage interpretation requires critically acknowledging these diverse and at times conflicting perspectives. Hence, heritage professionals and scholars increasingly explore innovative methods and tools to pursue more inclusive, diverse, and reflective approaches to facilitate multivocality.

This research investigates the potential of generative AI to produce historical image reconstructions that represent multiple cultural perspectives in heritage contexts (see Figure 1). The study makes three contributions: *(1)* It provides a linguistic analysis of the intermediate prompt transformation process, i.e. how AI models rewrite user-provided inputs before generating images, and identifies how these transformations affect culturally specific content. *(2)* It presents a systematic, expert-assessed comparison of three leading image generation models across three historically complex and contested heritage test cases. *(3)* Drawing on the findings, it proposes practical guidelines for structured prompt construction aimed at improving the historical accuracy and cultural sensitivity of AI-generated heritage imagery.

2 Literature Review and Related Work

2.1 Multivocality / Polyvocality

Heritage institutions are increasingly incorporating previously silenced or marginalised voices into historical narratives. Examples include London's Migration Museum, The Hague's Humanity House (refugee stories), and the Smithsonian's Slavery and Freedom exhibition. Common strategies include participatory or co-curative exhibitions, oral history initiatives, and collaborative research with underrepresented communities.

However, implementing multivocality is complex. Smith (2020) notes that even well-intentioned multivocal projects can yield mixed outcomes, as parties must accept coexisting, at times contradictory truths. Zheng (2023) highlights that amplifying marginalised voices remains harder in practice than in theory, partly because audiences are conditioned to accept singular narratives (Atalay, 2008; Barnabas, 2016). Authorised Heritage Discourse (AHD), highlighted by Smith (2006), further points out the resistance to these alternative viewpoints.

However, at the same time, scholars also underline the need to critically examine multivocality, as not all voices carry equal validity. While welcoming diverse perspectives, scholars advocate for the critical evaluation of each voice to ensure truthful heritage interpretation and to prevent misuse, whether for malicious intentions or unintended misinterpretations (Atalay, 2008; Kojan, 2008; Wylie, 2008).

2.2 Natural Language Processing

Research in NLP and computer vision increasingly examines how textual prompts shape the outputs of text-to-image (T2I) models. Oppenlaender (2024) established that prompt structure significantly affects generation quality. More broadly, using one language model to rewrite or optimise prompts for another model, or for a downstream generation model, has become common practice in NLP pipelines, from automated prompt engineering (Hao et al., 2023) to chain-of-thought prompt rewriting for improved image-text alignment (Wang et al., 2025). DALL-E 3 implements this approach by default: user prompts are automatically transformed before image generation. Our study examines this intermediate transformation as a site of potential meaning change, analysing the linguistic operations through which culturally specific user inputs are reshaped, an aspect that, to our knowledge, has not been systematically studied for heritage-related content.

Separately, a growing body of work addresses bias in T2I models. Wan et al. (2024) survey bias across gender, skin tone, and geo-cultural dimensions, finding the latter remains under-explored. Zhang et al. (2024) demonstrate that T2I models systematically underrepresent or stereotype cultures from less-represented regions in training data. These findings align with our observations of Western-centric defaults across the three heritage test cases. Additionally, T2I models exhibit well-documented compositional failures, difficulty in correctly binding attributes to objects in complex scenes (Huang et al.; Zarei et al., 2025). Heritage reconstruction prompts are inherently compositional, requiring models to associate specific architectural styles, attire, and activities with distinct cultural groups within a single scene, making this limitation particularly consequential in our domain.

2.3 Generative AI in Cultural Heritage

AI and generative AI are increasingly applied in cultural heritage for restoration, documentation, digital reconstruction, and public engagement. AI

image generation models have been used to reconstruct destroyed or incomplete heritage sites and to support immersive AR/VR storytelling (Altaweel et al., 2024; Arzomand et al., 2024; Moral-Andrés et al., 2024; Kutlu and Şimşek, 2025).

However, gen-AI historical image generation faces notable limitations. Models frequently produce visually plausible but historically inaccurate details, reflecting biases in training data that are disproportionately drawn from Western contemporary sources (Rane, 2023; Foka and Griffin, 2024; Spennemann, 2024; Sukkar et al., 2024; He et al., 2025; Liu et al., 2025). The closed-source nature of leading models further complicates scholarly verification of how inputs are processed and outputs are generated.

While substantial research exists on both multivocality in heritage and AI image generation separately, their intersection—using gen-AI to represent multiple perspectives through historical image reconstructions—remains largely unexplored. He et al. (2025) offer the closest prior work, examining how individuals used Stable Diffusion to create personal visual narratives about cultural heritage sites based on their own memories. While they did not investigate the representation of diverse perspectives attributed to the heritage sites themselves, their findings highlight that users' familiarity with a site significantly affects output accuracy, alongside the impact of cultural biases in model training data.

Our research extends this work in three specific ways. First, it incorporates a linguistic analysis of the intermediate prompt transformation process, how AI models rewrite user inputs before generating images, an aspect that remains under-examined in the literature (see Section 2.2). Second, it introduces expert verification by heritage professionals to assess the historical accuracy of depicted details. Third, it systematically compares model performance across multiple culturally contested heritage scenarios, rather than a single site or tradition.

3 Methodology

Given the interdisciplinary nature of this research, we adopted a Mixed Methods approach with a Deductive orientation, while combining Experimental Strategy with Case Study Strategy. This allowed assessing the behaviour of different generative AI models across varied heritage scenarios.

Accordingly, we selected heritage test cases with multiple, often conflicting perspectives, including religious, colonial, conflict, slavery, and multilayered heritage, to evaluate generative AI's ability to represent diverse perspectives. Furthermore, each case incorporated a random perspective to examine how AI responds to users with limited subject familiarity, informed by He et al.'s (2025) findings on the impact of users' familiarity or prior knowledge of the particular heritage context.

The case studies are:

(i) **A Surinamese colonial plantation** — This case study examines the Surinamese plantations during the Dutch colonial period, a contested history of forced labour and cultural identity (3 perspectives: Random, Dutch, Surinamese).

(ii) **Religious beliefs associated with Sri Pada Mountain in Sri Lanka** — This case study focuses on Sri Pada Mountain in Sri Lanka, a religious heritage site uniquely sacred to Buddhist, Hindu, Islamic, and Christian communities (6 perspectives: Buddhist, Hindu, Islamic, Christian-I, Christian-II, Random).

(iii) **1640 Dutch siege of the Portuguese-held Galle Fort in Sri Lanka** — This case study addresses the Dutch VOC siege of Portuguese-held Galle Fort in Sri Lanka in 1640, a turning point in colonial history, not only for Sri Lanka but also for the Indian Ocean region, changing colonial rule from the Portuguese to Dutch (4 perspectives: Dutch, Portuguese, Sri Lankan, Random).

3.1 Image Generation Set-up

Using zero-shot text-to-image generation, we prompted the following three GenAI models with a total of 13 perspectives (from the three case studies), resulting in 39 images.

- DALL-E 3 (Betker et al., 2023)
- Stable Diffusion V3.5 (Esser et al., 2024)
- Midjourney V.6 (Midjourney, Inc., 2024)

Informed by Oppenlaender (2024), we constructed the 13 user inputs for the following eight keyword categories which we derived from the prompt engineering guidelines of the selected models:

(i) Topic
(ii) Landscape/Environment
(iii) Architecture
(iv) People (including attire)
(v) Activities
(vi) Atmosphere
(vii) Special Objects (if any)
(viii) Artistic Style

These identified keyword categories serve dual purposes: (i) guiding effective prompt construction, and (ii) functioning as cultural parameters/criteria for evaluating the historical accuracy of the generated images.

To identify relevant keywords for each perspective, we conducted semi-structured discussions with 7 members of the general public, each lasting approximately 20–40 minutes. We selected participants through purposive sampling to represent the cultural perspectives relevant to each test case: individuals with Dutch and Surinamese backgrounds for Test Case 1, practitioners of the relevant religious traditions for Test Case 2, and individuals with Dutch, Portuguese, and Sri Lankan heritage connections for Test Case 3. When a participant lacked knowledge for a given keyword category, we recorded the phrase "I don't know" as the keyword input. This deliberate inclusion simulates visitors with limited prior knowledge, informed by He et al. (2025), who found that user familiarity with a heritage site significantly affects output accuracy. Each random perspective was obtained from a participant with no prior familiarity with the specific heritage site or its cultural context, simulating a general visitor with limited subject knowledge. Throughout this study, the term 'visitor' refers to an individual representing a particular cultural perspective, simulating the range of people who might engage with a heritage site, whether as members of an associated community, a historically connected nationality, or as a general member of the public.

The user keywords were first input into DALL-E 3, which automatically rewrites user inputs into expanded prompts before generating images (see Table 2). As discussed in Section 2.3, this intermediate prompt rewriting is common practice in current T2I pipelines. It also represents a key site of potential meaning change and constitutes one of the central objects of analysis in this study. Section 3.2 examines how this transformation reshapes user intent across all 13 perspectives.

To compare image generation performance on an equal basis, we used the DALL-E-expanded prompts as identical inputs for Stable Diffusion V3.5 and Midjourney V.6. This two-stage design serves two purposes. First, it enables analysis of how LLM-based prompt rewriting affects culturally specific inputs. Second, it allows a controlled comparison of three image generation models interpreting the same enriched prompt. This means Stable Diffusion and Midjourney were not tested on raw user keywords. Their scores reflect image generation capability given an already-transformed prompt.

3.2 Analytical Procedures

This user prompt-driven AI image generation was followed by analytical and evaluation procedures that combined qualitative and quantitative techniques, analysing two key aspects:

1. **Linguistic Analysis** — examining the intermediate conversion of user-given keywords into model-generated prompts
2. **Visual Analysis** — evaluating the historical accuracy of the AI-generated images

We identified the following eleven linguistic operations through iterative analysis of the 13 keyword-prompt generation pairs, drawing on established categories from paraphrase typology (Vila et al., 2014) and adapted to the specific context of LLM-based prompt rewriting for image generation:

(i) Synthesis/Generation
(ii) Specification
(iii) Expansion
(iv) Enrichment
(v) Clarification/Disambiguation
(vi) Merging
(vii) Substitution
(viii) Omission
(ix) Rephrase
(x) Visual mapping
(xi) Style-context matching

To assess the accuracy of historical details depicted in the generated images, three heritage experts evaluated the 39 image outputs using a custom-designed evaluation form. One domain-specific expert with relevant specialisation (in history, art history, or heritage interpretation respec-

tively) assessed each test case. The evaluation combined quantitative and qualitative assessments: each image was rated across the 8 cultural criteria (see Section 3), using a scale of 1 (no resemblance to the historically expected depiction) to 5 (highly accurate and culturally appropriate representation), with space provided for textual remarks. The qualitative feedback was particularly valuable for interpreting lower scores. It is important to note that the experts were not asked to evaluate the perspective conveyed by the images, but solely to assess the plausibility of visual details across the 8 cultural criteria. As each test case was evaluated by a single expert, inter-rater reliability could not be calculated; this limitation is discussed in Section 6.

Following expert feedback, we cross-analysed all 39 images using three methods:

(a) Single perspective across models

(b) Multiple perspectives within one model

(c) Multiple test cases within one model

4 Data Analysis & Results

The data analysis across three test cases suggested notable variations in how AI models performed depending on the heritage contexts.

4.1 Intermediate Linguistic Transformations between User-Given Keywords and Model-Generated Prompts

In Test Case 1 (Surinamese Plantations), linguistic operations such as *Specification*, *Enrichment*, and *Expansion* have been consistently applied for six keyword categories, except in *Atmosphere* and *Artistic Style*, across Random, Dutch, and Surinamese visitor perspectives, to make prompts more comprehensive for supporting image generation (see Table 1). It is noteworthy that requested architectural references have been omitted or reinterpreted depending on the visitor's standpoint (i.e. accommodating 'Dutch colonial architecture' in all prompts, and omitting 'Surinamese architectural elements', except in the Surinamese prompt). Notably, the model retained the input 'Surinamese workers' only in the Surinamese visitor prompt, but substituted it with 'Enslaved Africans, labourers' for the other two perspectives.

The linguistic operations *Enrichment*, *Expansion*, and *Specification* have been used across all eight criteria in all six perspectives for Test Case 2 (Sri Pada Mountain religious heritage site), producing poetic language and descriptively detailed descriptions. However, except in the Buddhist and Random perspectives, inaccuracies have emerged in all prompts, as the model has incorrectly implied the pre-existence of a footprint or introduced architectural elements, supposedly based on documented history rather than the user's keyword input.

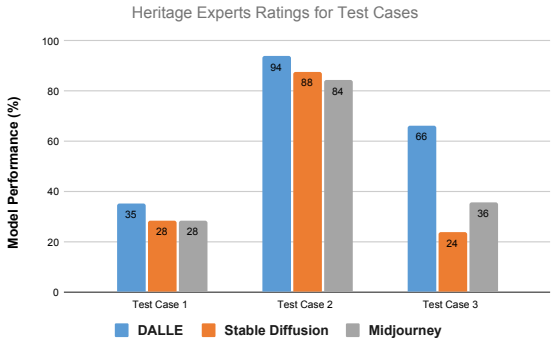


Figure 2: Cumulative heritage expert ratings (and percentages of maximum possible scores) for AI-generated images across three test cases and three models. Scores aggregate one expert's ratings per test case across all perspectives and eight cultural criteria, each rated on a 1–5 Likert scale.

Test Case 3 (Dutch siege of Galle Fort) with four perspectives (Dutch, Portuguese, Sri Lankan, and Random), had an inverse relationship between input details and output quality. While the model extensively expanded the Random visitor's brief input into the most detailed and contextually accurate model prompt, the Dutch and Portuguese visitors' comprehensive inputs were simplified, with commander names omitted and specific historical details substituted by vague generalisations.

Across all three test cases, these intermediate linguistic transformations have had both positive and negative impacts. Positive effects are refining vague inputs into coherent and visually descriptive prompts, while negative effects include inaccuracies, cultural misinterpretations, and the modification of user intents, which directly impacted the resulting historical image reconstructions.

4.2 Comparative Image Analysis

Heritage expert ratings are recorded as cumulative scores for the AI-generated image depictions of each test case (see Figure 2). DALL-E consistently outperformed Stable Diffusion and Midjourney, with the strongest results in Test Case 2 and

Dutch Visitor's Perspective			
Criterion	**User Input Keyword(s)**	**Prompt Phrase(s)/ Expansions**	**Linguistic Operation(s) Observed by the Author**
1) Topic	Surinamese Plantation during the colonial period	a day on a Suriname sugar plantation during the colonial era	Specification
2) Landscape/ Environment	Surinamese land	surrounded by towering palm trees and lush greenery	Expansion Enrichment
3) Architecture	Colonial plantation bungalow, labour quarters/ camps	a white plantation house with a red-tiled roof and wraparound veranda	Specification Enrichment Omission
4) People	Dutch plantation owners, Surinamese workers	enslaved Africans… two European men in period attire	Substitution Specification
5) Activities	Slaves are working, and Dutch Masters overseeing it	toil in the fields… overseen by…	Rephrase
6) Atmosphere	Daytime	…captures a day	*(no significant observation)*
7) Special Objects (if any)	Plantation-related objects	baskets, wheelbarrow, agricultural tools	Generation
8) Artistic Style	Any style suitable for a historical scenery	A traditional oil painting	Style-context matching

Table 1: Linguistic transformations of the Dutch visitor's perspective in Case Study 1 (Surinamese Plantations), showing user-provided keywords, DALL-E 3's rewritten prompt phrases, and the linguistic operations identified for each criterion. Notable transformations include the substitution of 'Surinamese workers' with 'enslaved Africans,' the omission of user-requested labour quarters, and the generation of specific objects not present in the original input.

the weakest across all models in Test Case 1.

In Test Case 1, involving Surinamese colonial plantations, the cumulative scores of all perspectives given by heritage experts for each model were: DALL-E 42, Stable Diffusion 34, and Midjourney 34, out of a possible 120. The score differences were minimal, and no model scored higher than 3 on any individual cultural criterion. These widespread inaccuracies across all visitor perspectives reveal that all three models struggled in representation (see Figure 3).

Test Case 2, which focused on the plurality of religious beliefs of the Sri Pada Mountain heritage site, surprisingly revealed strong results. Despite initial expectations that South Asian religious settings might be challenging to represent for predominantly Western-trained models, all three performed reasonably well in representing diverse pilgrim perspectives. DALL-E led with 225 points, followed by Stable Diffusion with 210, and Midjourney with 202, out of 240. However, some Stable Diffusion outputs resembled Far East Asian rather than South Asian landscapes.

Test Case 3, depicting the Dutch capturing Portuguese-held Galle Fort in 1640, revealed significant performance gaps. DALL-E scored 106, Midjourney 57, and Stable Diffusion 38, out of a possible 160, with each model scoring roughly half that of the preceding one. While DALL-E produced the most historically plausible visualisations, all models still displayed weaknesses in historical accuracy across the four visitor perspectives.

4.3 Overall Observations

We observed several patterns across the three test cases. The linguistic analysis revealed that when users provided vague and ambiguous text inputs (i.e. "I don't know", "…maybe?"), especially in random visitor perspectives, the model (DALL-E) filled the gaps with information that was at times accurate and sometimes not. In the Surinamese test case, we noticed an imbalance in representation of architectural elements, as the colonial bungalows were consistently and prominently depicted, while the user-requested structures, like labour houses, were largely ignored, appearing only in two of the nine images generated. Further we observed imbalances in the Galle Fort test case, with the depiction of Portuguese and Dutch forces, where images have frequently under-represented the Dutch forces regardless of the visitor perspective, especially in Stable Diffusion and Midjourney. However, linguistic transformations that occurred in prompts (see Section 3.2) appear to have had an impact on

this imbalance.

The analysis of the linguistic transformation process further revealed notable model tendencies: neutralisation and amplification, observed in certain contexts. In the Surinamese test case, the model tended to soften or neutralise references to colonial violence (in user inputs) without explicit instructions, and expert feedback confirmed that the resulting images have often been romanticised, lacking any reflection of the historical hardship inherent to the setting. One possible explanation is that the model's safety or content moderation mechanisms may influence how it handles references to historical violence, though this cannot be confirmed given the closed-source nature of these models.

Conversely, in the Sri Pada Mountain test case, the linguistic transformation has amplified the spiritual and emotional language, deepening the sacred tone of the images. However, this same linguistic transformation process has introduced a critical error in the Hindu, Islamic, Christian-I and Christian-II pilgrim perspectives, by misrepresenting those less-documented religious beliefs (the act of imprinting a footprint as a pre-existing one on the rock summit), leading to inaccurate image outputs.

Furthermore, Test Case 3 revealed a surprising inverse relationship between input details and output quality, where the random visitor's brief input yielded the most detailed and contextually accurate prompt, while the model simplified the comprehensive Dutch and Portuguese visitor inputs into brief, vague, generalised prompts.

Across the three test cases, the models performed better on the spiritual heritage scenario (Test Case 2) than on the historical event reconstructions (Test Cases 1 and 3). However, this likely reflects differences in representational complexity rather than an inherent category distinction—Test Case 2 involved fewer compositional demands per image frame. Furthermore, notable differences in how human figures are portrayed were evident across the models. Stable Diffusion produced vague depictions, Midjourney frequently avoided showing faces, and DALL-E consistently rendered highly detailed, expressive faces in the foreground. Similarly, the models differed in their use of colour and style. Despite prompts requesting traditional oil painting styles and historical scenes, Stable Diffusion images consistently used bright, high contrast colours, which experts remark as unrealistic and less effective for historical scene depictions.

Overall, errors, omissions, and misinterpretations of user inputs during the intermediate linguistic transformation process were observed across all perspectives and test cases. In addition, the varied nature of the heritage test cases required domain-specific experts (one per case, with expertise in domains including history, art history, and heritage interpretation) to validate the historical accuracy of the image depictions. However, it also introduced varying levels of criticality and interpretive subjectivity, which should be kept in mind when interpreting the results.

5 Discussion

5.1 Potential of Gen-AI in Fostering Meaningful Dialogue on Multivocality in Heritage Interpretation

Despite the limitations identified, the findings suggest that Gen-AI may offer a useful starting point for engaging audiences with the complexities of heritage interpretation, particularly when used alongside expert guidance. Across the three test cases, while models often struggled with historical accuracy, they demonstrated an ability to visualise multiple perspectives within a single heritage scenario, providing an initial basis for exploring how diverse perspectives on heritage might be visualised.

For instance, in the Sri Pada Mountain case, AI models captured distinct spiritual atmospheres for Buddhist, Hindu, Islamic, Christian pilgrims (and Random visitors), reflecting the site's layered, overlapping beliefs. Even in the more challenging Surinamese and Galle Fort contexts, generated visuals reflected diverse standpoints, local, colonial, and neutral, illustrating how generative AI can serve as a starting point for critical conversations about how complex histories are viewed and remembered.

The models tended to simplify or neutralise sensitive histories in the Surinamese plantation and Galle Fort cases, while amplifying spiritual and emotional language in the Sri Pada Mountain case. These tendencies likely reflect the composition of training data, which is predominantly drawn from digitised Western sources and dominant historical narratives. In heritage contexts such as the Surinamese and Galle Fort cases, the perspectives of colonised and enslaved populations were historically excluded from official documentation by the very authorities who produced it. The available training data for these contexts is therefore both

scarce and skewed toward the dominant perspective, meaning the models have fewer and less diverse references to draw on. AI-generated heritage imagery consequently risks compounding existing representational gaps, amplifying well-documented perspectives while further marginalising those that were suppressed at the source. Recognising these limitations is essential, as gen-AI alone cannot resolve historical bias or representation gaps.

When such imagery is considered for use in cultural institutions, additional considerations arise beyond historical accuracy. This study consulted individuals with Surinamese heritage to inform prompt construction; however, the deployment of AI-generated imagery depicting enslaved populations in museum or visitor engagement settings would require broader consultation with descendant communities. Not only on what is depicted, but on the conditions under which such imagery is displayed, contextualised, and circulated. As several generated images in this study tended toward romanticised depictions of plantation life, institutions adopting these tools should establish review processes that ensure such imagery is accompanied by appropriate contextualisation and does not inadvertently aestheticise historical suffering.

Nevertheless, its capacity to produce multivocal visual narratives offers heritage professionals, educators, and communities a participatory tool that, used critically and alongside expert input, can foster richer and more inclusive dialogue around complex and contested heritage. To support this critical use, we propose practical guidelines derived from the failure patterns documented above.

5.2 Guidelines for Structured Prompt Construction in Heritage Image Reconstruction

The analysis across three test cases and three models revealed recurring failure modes in AI-generated heritage imagery: omission of culturally specific elements, substitution of user terminology with generalised alternatives, romanticisation of sensitive historical contexts, and imbalanced representation of different groups within a scene. Based on these observations, we propose the following guidelines for constructing prompts that mitigate these specific issues.

We recommend describing the intended scene using structured keywords across eight dimensions, rather than in paragraph form, as this reduces the likelihood of elements being merged or omitted during the intermediate prompt transformation:

(a) **Topic** — the event or scenario, including explicit time period and geographic specificity, to prevent the model defaulting to generic historical settings.

(b) **Landscape/Environment** — geography, lighting, season, flora and fauna, structured as foreground, middle-ground, and background where possible, to reduce the model's reliance on stereotypical landscape compositions.

(c) **Architecture** — structures, materials, and culturally distinctive designs. Our analysis showed that colonial-dominant structures were consistently prioritised while user-requested alternatives (e.g. labour quarters) were omitted. Listing all intended structures with equal specificity can counteract this tendency.

(d) **People (including attire)** — clothing, ethnicity, and social roles, using precise terminology. The substitution of 'Surinamese workers' with 'enslaved Africans' in some perspectives but not others demonstrates that user-provided terms are not always preserved. Explicit, consistent terminology across this dimension is critical.

(e) **Activities** — specific actions and interactions, described with sufficient detail to prevent neutralisation. Vague descriptors like 'working in the fields' were consistently softened; specifying the nature and conditions of labour can resist romanticisation.

(f) **Atmosphere/Ambience** — mood and emotional tone. This dimension proved particularly important for counteracting the model's tendency to aestheticise difficult histories. Explicit descriptors such as 'reflecting hardship and forced labour' rather than neutral terms like 'daytime' can guide the model away from picturesque defaults.

(g) **Special Objects** — tools, machines, furnishings, transportation. When left unspecified, the model generated plausible but historically inaccurate objects. Providing specific items grounded in the historical context reduces this risk.

(h) **Artistic Style** — preferred style aligned with the historical context. Our findings showed that style-context matching was one of the more reliable transformations, but specifying style explicitly remains important to avoid anachronistic visual treatments.

We further recommend the following process:

1. Define the image's intended purpose and the perspective it should represent.

2. Construct keywords across all eight dimensions, ensuring equal specificity for all cultural groups and elements in the scene.

3. Generate the prompt using a suitable LLM, then critically review the expanded prompt before image generation — checking specifically for omissions, substitutions, and neutralisation of sensitive content.

4. Generate images across multiple models to compare outputs, as our results showed significant variation in how models interpreted identical prompts.

5. Seek expert review, particularly for professional or educational use in cultural institutions.

These guidelines do not guarantee historically accurate outputs, but they address the specific failure patterns documented in this study and provide a structured basis for iterative improvement.

6 Conclusion and Future Work

This research investigated whether generative AI image models can produce historical reconstructions that reflect multiple cultural perspectives. It examined both the linguistic transformations applied to user prompts and the visual accuracy of the resulting outputs across three heritage test cases and three T2I models. Three findings emerge from this exploratory study. First, intermediate prompt rewriting, a largely unstudied stage in T2I pipelines, substantially affects output accuracy, introducing both helpful enrichments and culturally inaccurate modifications. Second, model performance varied across heritage contexts: all three models performed reasonably on the spiritual heritage scenario but struggled with the representational complexity of colonial and military scenarios. Third, recurring patterns of representation imbalance, content neutralisation, and Western-centric defaults appeared across test cases, consistent with the broader literature on cultural bias in T2I systems. Based on these findings, we proposed practical guidelines for structured prompt construction in cultural heritage contexts. Future work should expand to additional models and heritage contexts, include multiple expert evaluators per test case to enable inter-rater reliability assessment, and test whether structured prompts measurably improve accuracy through iterative application. Generative AI shows early promise as a supplementary tool for exploring multivocal heritage interpretation, but its use requires critical evaluation and expert oversight.

Limitations

The scope of this research was limited to three AI models and three heritage test cases; broadening to a wider range of heritage scenarios and models would yield deeper insights into model behaviour. Each test case was evaluated by a single heritage expert, precluding inter-rater reliability measurement, and engaging multiple experts per case is recommended to reduce the impact of varying interpretive thresholds. The experimental design used DALL-E 3's expanded prompts as inputs for Stable Diffusion and Midjourney, meaning those models were not tested under their own pipeline conditions, which limits direct comparability. The semi-structured discussions involved seven purposively sampled participants, a small sample that may not fully represent the breadth of cultural perspectives relevant to each case. Furthermore, the study assessed historical accuracy of visual details but did not evaluate whether end-users perceived the images as representing meaningfully distinct perspectives. While individuals with Surinamese heritage were consulted for keyword generation, this study did not assess the broader ethical considerations that would arise from deploying AI-generated imagery of enslaved populations in cultural institutions. Future work should explore how museums and heritage sites can establish appropriate ethical frameworks, including descendant community consultation, contextualisation protocols, and display guidelines, before integrating such imagery into public-facing interpretation. Finally, the closed-source nature of the models tested limits our ability to make causal claims about why specific transformations or outputs occurred.

Acknowledgments

We would like to thank the individuals representing the general public who participated in brief, informal discussions, which helped us gather keywords for the user prompts. We are also grateful to the heritage experts who generously shared their knowledge and perspectives through the expert evaluation. This work was partially funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105, HAICu).

References

Mark Altaweel, Adel Khelifi, and Mohammad Hashir Zafar. 2024. Using Generative AI for Reconstructing Cultural Artifacts: Examples Using Roman Coins. *Journal of Computer Applications in Archaeology*, 7(1):301–315.

Kawsar Arzomand, Michael Rustell, and Tatiana Kalganova. 2024. From ruins to reconstruction: Harnessing text-to-image AI for restoring historical architectures. *Challenge Journal of Structural Mechanics*, 10(2):69.

Sonya Atalay. 2008. *Multivocality and Indigenous Archaeologies*, pages 29–44. Springer New York, New York, NY.

Shanade Bianca Barnabas. 2016. Heritage-making and the dilemma of multivocality in south africa: a case of wildebeest kuil. *International Journal of Heritage Studies*, 22(9):690–701.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. Technical report. Available at `https://cdn.openai.com/papers/dall-e-3.pdf`.

Jacques Derrida. 1994. *Specters of Marx: The State of the Debt, the Work of Mourning, and the New International*, 1st edition. Routledge, New York.

Ana Deumert. 2018. The multivocality of heritage – moments, encounters and mobilities. In Angela Creese and Adrian Blackledge, editors, *The Routledge Handbook of Language and Superdiversity*, 1 edition, pages 149–164. Routledge.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning (ICML)*.

Anna Foka and Gabriele Griffin. 2024. AI, Cultural Heritage, and Bias: Some Key Queries That Arise from the Use of GenAI. *Heritage*, 7(11):6125–6136.

Isabel F. Franke, Stefania D. Conte, Claudia A. Libbi, Victor De Boer, and Tilo Hartmann. 2024. A Polyvocal Approach to Virtual Heritage: An Immersive Case Study. *Journal of Computing and Cultural Heritage*, 17(4):1–23.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. *arXiv preprint*.

Zhiting He, Jiayi Su, Li Chen, Tianqi Wang, and Ray LC. 2025. "I Recall the Past": Exploring How People Collaborate with Generative AI to Create Cultural Heritage Narratives.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. *NeurIPS*.

David Kojan. 2008. Paths of Power and Politics: Historical Narratives at the Bolivian Site of Tiwanaku. In Junko Habu, Clare Fawcett, and John M. Matsunaga, editors, *Evaluating Multiple Narratives: Beyond Nationalist, Colonialist, Imperialist Archaeologies*, pages 69–85. Springer New York, New York, NY.

İzzettin Kutlu and Deryanur Şimşek. 2025. Artificial intelligence (AI) assisted digital reconstruction of historical buildings. In *4th International Civil Engineering & Architecture Conference Volume 2: Architecture*. Golden Light Publishing.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and Improving Cultural Understanding of Vision-Language Models for over 100 Countries. *arXiv preprint*.

Midjourney, Inc. 2024. Midjourney. `https://www.midjourney.com`. Version 6. Accessed: [insert date].

Fernando Moral-Andrés, Elena Merino-Gómez, Pedro Reviriego, and Fabrizio Lombardi. 2024. Can Artificial Intelligence Reconstruct Ancient Mosaics? *Studies in Conservation*, 69(5):313–326.

Jonas Oppenlaender. 2024. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, 43(15):3763–3776.

Nitin Rane. 2023. Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Arts and Humanities. *SSRN Electronic Journal*.

Claire Smith, editor. 2020. *Encyclopedia of Global Archaeology*. Springer International Publishing, Cham.

Laurajane Smith. 2006. *Uses of Heritage*, 1st edition. Routledge: Taylor & Francis Group, London.

Dirk H. R. Spennemann. 2024. Generative Artificial Intelligence, Human Agency and the Future of Cultural Heritage. *Heritage*, 7(7):3597–3609.

Ahmad W. Sukkar, Mohamed W. Fareed, Moohammed Wasim Yahia, Salem Buhashima Abdalla, Iman Ibrahim, and Khaldoun Abdul Karim Senjab. 2024. Analytical evaluation of midjourney architectural virtual lab: Defining major current limits in ai-generated representations of islamic architectural heritage. *Buildings*, 14(3).

Marta Vila, M. Antònia Martí, and Horacio Rodríguez. 2014. Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(1):205–218.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of Bias in Text-to-Image Generation: Definition, Evaluation, and Mitigation. *arXiv preprint*.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, QiXun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. 2025. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*.

Alison Wylie. 2008. *The Integrity of Narratives: Deliberative Practice, Pluralism, and Multivocality*, pages 201–212. Springer New York, New York, NY.

Arman Zarei, Keivan Rezaei, Samyadeep Basu, Mehrdad Saberi, Mazda Moayeri, Priyatham Kattakinda, and Soheil Feizi. 2025. Improving Compositional Attribute Binding in Text-to-Image Generative Models via Enhanced Text Embeddings. *arXiv preprint*.

Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, Honolulu, HI, USA. ACM.

Nana Zheng. 2023. *Coming to Grips with Diverse Voices in Participatory Heritage Initiatives*. PhD thesis, Vrije Universiteit Amsterdam.

A Appendix

Table 2: Case Study 1 – Surinamese Plantations: User input keywords and AI-generated prompts across three visitor perspectives.

	User Input Keywords		AI-generated Prompt (DALL-E 3)
Random Visitor	Topic:	Surinamese Plantation during the colonial period	A colonial-era oil painting showcases a day on a Surinamese plantation, with enslaved African workers toiling under the watch of Dutch overseers. In the foreground, laborers engage in agricultural tasks, while a grand plantation house with Dutch architectural elements dominates the background, surrounded by lush tropical vegetation. The piece employs earthy tones, rich greens, and subtle splashes of color to create a balanced composition, focusing on the contrast between the workers and their overseers.
	Landscape/Env.:	Surinamese land, maybe flat land I guess	
	Architecture:	Colonial architecture, Surinamese architecture	
	People:	Surinamese people, Dutch people	
	Activities:	Surinamese people working in the land, Dutch overseeing it	
	Atmosphere:	Daytime	
	Special objects:	I don't know	
	Artistic Style:	Any style suitable for a historical scenery	
Dutch Visitor	Topic:	Surinamese Plantation during the colonial period	A traditional oil painting captures a day on a Suriname sugar plantation during the colonial era. At its heart is a white plantation house with a red-tiled roof and wraparound veranda, surrounded by towering palm trees and lush greenery, while enslaved Africans toil in the fields nearby, overseen by two European men in period attire.
	Landscape/Env.:	Surinamese land	
	Architecture:	Colonial plantation bungalow, labour quarters/camps	
	People:	Dutch plantation owners, Surinamese workers	
	Activities:	Slaves are working, and Dutch Masters overseeing it	
	Atmosphere:	Daytime	
	Special objects:	Plantation-related objects	
	Artistic Style:	Any style suitable for a historical scenery	
Surinamese Visitor	Topic:	Surinamese Plantation during the colonial period	A historical 2D reconstruction of a Surinamese plantation during the colonial period, set in the lush tropical Surinamese landscape. In the scene, a colonial-style plantation bungalow with Surinamese architectural elements stands prominently. Surinamese workers are seen laboring in the fields under the hot daytime sun, engaged in plantation activities such as harvesting and carrying loads. Dutch plantation owners in colonial attire are sternly overseeing the work. The environment features typical plantation crops, tools, and equipment of the time. The artistic style reflects a realistic yet painterly tone suitable for a historical scene, emphasizing the atmosphere and social dynamics of the colonial era.
	Landscape/Env.:	Surinamese land	
	Architecture:	Colonial plantation bungalow, Surinamese architecture	
	People:	Surinamese workers, Dutch Plantation owners	
	Activities:	My people (Surinamese) working hard, and Colonial Dutch plantation owners are tightly overseeing	
	Atmosphere:	Daytime	
	Special objects:	Plantation-related objects	
	Artistic Style:	Any style suitable for a historical scenery	

Random Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Dutch Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Surinamese Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Figure 3: Case Study 1 – Images generated of Surinamese Plantations during the colonial period for random, Dutch, and Surinamese perspective.

Temporal Text Classification with Large Language Models

Nishat Raihan
George Mason University
Fairfax, VA, USA
mraihan2@gmu.edu

Marcos Zampieri
George Mason University
Fairfax, VA, USA
mzampier@gmu.edu

Abstract

Languages change over time. Computational models can be trained to recognize such changes enabling them to estimate the publication date of texts. Despite recent advancements in Large Language Models (LLMs), their performance on automatic dating of texts, also known as Temporal Text Classification (TTC), has not been explored. This study provides the first systematic evaluation of leading proprietary (Claude 3.5, GPT-4o, Gemini 1.5) and open-source (LLaMA 3.2, Gemma 2, Mistral, Nemotron 4) LLMs on TTC using three historical corpora, two in English and one in Portuguese. We test zero-shot and few-shot prompting, and fine-tuning settings. Our results indicate that proprietary models perform well, especially with few-shot prompting. They also indicate that fine-tuning substantially improves open-source models but that they still fail to match the performance delivered by proprietary LLMs.

1 Introduction

Language variation is a well-researched topic in computational linguistics. It is manifested in different dimensions such as *diatopic*, related to place as in the case of dialects and language varieties (Joshi et al., 2025; Zampieri et al., 2020), and *diachronic* which addresses language change over time (Frermann and Lapata, 2016; Schlechtweg et al., 2020). Several papers have explored computational approaches to model diachronic variation in word embeddings (Hamilton et al., 2016), semantic modeling (Kulkarni et al., 2015), and related tasks.

Temporal Text Classification (TTC) (Dalli and Wilks, 2006; Ciobanu et al., 2013; Popescu and Strapparava, 2015), also referred as automatic date detection in texts (Jauhiainen, 2024) or automatic text dating (Ren and Peng, 2025), is among the most important tasks related to modeling diachronic language variation in texts. TTC consists of assigning temporal labels to diachronic collection of texts where the labels usually represent pre-defined time spans (e.g., 10 years, 50 years, a century). It is vital step in organizing and analyzing large corpora in digital humanities and information retrieval. While a couple studies have approached TTC as a document ranking problem (Niculae et al., 2014), the clear majority of studies have approached the task using supervised categorization with statistical machine learning classifiers trained on stylistic and n-gram based features (Szymanski and Lynch, 2015; Nunes et al., 2007).

The latest generation of LLMs (Hurst et al., 2024; Grattafiori et al., 2024) has displayed state-of-the-art performance in many NLP tasks including question answering (Fischer et al., 2024), machine translation (Kocmi et al., 2024), and text summarization (Zhang et al., 2024a). While LLMs show promising results in related areas like temporal reasoning (Zhang et al., 2024b), their performance on TTC has not yet been explored.

To address this important gap, we present the first evaluation of proprietary and open-source LLMs on TTC. We acquire three well-known historical corpora in two languages, namely Colonia (Portuguese) (Zampieri and Becker, 2013), CLMET (De Smet, 2005) and DTE (English) (Popescu and Strapparava, 2015). We analyze LLMs performance using different strategies such as zero-shot prompting, few-shot prompting, and task fine-tuning. Furthermore, we also examine the influence of input text length and the efficacy of fine-tuning for open-source models. Finally, we evaluate the models' data generation capabilities when generating synthetic data from different time periods.

The main contributions of this work are the following:

1. We present the first comprehensive evaluation of proprietary and open-source LLMs for TTC considering multiple paradigms (zero-shot, few-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 96–105
July 4, 2026 ©2026 Association for Computational Linguistics

Dataset	Lang	Time Span	Size	Classes
Colonia	PT	16th–20th C.	>5M Tokens	5 (100 year)
CLMET	EN	1710–1920	34M Words	3 (70 year)
DTE	EN	1700–2010	Varies by Split	∼4 Intervals*

Table 1: Overview of datasets. Size indicates approximate total corpus size. *Temporal classes for SemEval DTE depend on the specific sub-task configuration (e.g., 4 intervals like pre-1700, 18th C., 19th C., post-1900 were common).

shot, fine-tuning) and various text input lengths.

2. We explore the synthetic data generation capabilities of these models when generating synthetic data from various time periods.

3. We present new insights into the effectiveness of fine-tuning for enhancing open-source LLM performance on TTC tasks relative to proprietary models.

The research presented here provides important insights into current LLM capabilities and limitations for TTC and diachronic language change more broadly. This will, in turn, inform future model development and application in temporal text analysis. Finally, the inclusion of corpora in two languages allows us to understand current LLMs support to languages other than English.

2 Related Work

The first studies on TTC can be traced back to de Jong et al. (2005) and Dalli and Wilks (2006). Both papers describe computational models to automatically dating texts in diachronic collections of newspapers texts. de Jong et al. (2005) explored texts extracted from Dutch newspapers while Dalli and Wilks (2006) studied English news texts from the GigaWord corpus.

TTC increased in popularity with the SemEval-2015 Diachronic Text Evaluation (DTE) task (Popescu and Strapparava, 2015). Participants were tasked to train computational models to date English texts using temporal labels in three levels of granularity: fine (2-6 years), medium (6-12 years), and coarse (12-20). Approaches to DTE have relied on statistical classifiers trained on n-grams, grammatical, and stylistic features (Szymanski and Lynch, 2015), temporal ranking (Zampieri et al., 2015), and the temporal distribution of tokens on Google n-grams (Salaberri et al., 2015). Since DTE, several studies have approached TTC using primarily statistical classifiers combined with feature engineering (Boldsen and Wahlberg, 2021).

A couple recent studies have approach TTC using neural language models (Ren et al., 2023; Ren and Peng, 2025). However, to the best of our knowledge, the use of recently-introduced foundational LLMs (e.g., GPT, LLama) has not yet been approached for this task. Our work fills this gap by presenting the first evaluation of LLMs on TTC.

3 Datasets and Evaluation Setup

We evaluate LLM performance on three corpora, each presenting unique characteristics as summarized in Table 1.

Colonia (Zampieri and Becker, 2013) is a diachronic collection of Portuguese texts including full novels spanning the 16th to early 20th centuries, sourced from Brazil's `Domínio Público` digital library, the `GMHP` corpus, and the `Tycho Brahe Parsed Corpus of Historical Portuguese`. Texts are divided into five classes each corresponding to a century (1500s, 1600s, 1700s, 1800s, and 1900s).

CLMET The Corpus of Late Modern English Texts (CLMET) (De Smet, 2005) covers British English from 1710 to 1920, constructed from works in the public domain. Sourced from digital library `Project Gutenberg`, the University of Oxford's `Oxford Text Archive`, and Indiana University's `Victorian Women Writers Project`. The classification task involves assigning documents to one of three 70-year periods: 1710–1780, 1780–1850, or 1850–1920.

DTE The dataset from SemEval-2015 Task 7 (Diachronic Text Evaluation) (Popescu and Strapparava, 2015) is specifically compiled for temporal analysis. The collection is curated from the historical archive of the British weekly magazine - `The Spectator`.

Models We evaluate the performance of seven LLMs, detailed in Table 2.

Model	Type	Param.
Claude 3.5 Sonnet	Proprietary	-
GPT-4o	Proprietary	-
Gemini 1.5 Pro	Proprietary	-
LLaMA 3.1	Open-Source	8B
Gemma 2	Open-Source	9B
Mistral	Open-Source	7B
Nemotron	Open-Source	4B

Table 2: List of LLMs evaluated.

Input Lengths Testing with initial text segments of 10, 50, and 100 tokens to analyze the impact of context size. Longer documents are handled via chunking strategies where applicable. The reason for segmentation, in line with previous TTC studies (Zampieri et al., 2016), is to make texts comparable in length.

Evaluation Paradigms Assessing performance under **Zero-Shot** (prompting with task description only), **Few-Shot** (providing 5 representative examples), and, for open-source models, **Fine-tuning** (using LoRA (Hu et al., 2021) on dataset-specific training splits). We also explore fine-tuning with synthetic data.

Metric Reporting performance using the **Macro F1 score** to ensure a balanced evaluation across uneven temporal class distributions.

4 Results & Analysis

For each experiment, we present some key findings, trends, and discuss variation with potential justifications whenever appropriate. We consider Support Vector Machine (SVM) (Cortes and Vapnik, 1995) as the baseline in line with previous work (Zampieri et al., 2015; Popescu and Strapparava, 2015; Jauhiainen, 2024).

4.1 Prompting

We present the results for zero-shot and few-shot prompting in Table 3.

Key Findings Proprietary LLMs demonstrate strong temporal classification capabilities across all datasets. Performance (F1 score) consistently scales with input length (10 to 100 tokens), indicating effective use of richer lexical and stylistic signals in longer texts.

Trends Few-shot (5-shot) prompting significantly outperforms zero-shot evaluation in most cases, often achieving F1 scores above 0.90, especially with inputs of 100+ tokens. This highlights robust in-context learning, where models adapt efficiently from only five examples. GPT-4o frequently yields the highest scores, particularly in few-shot scenarios. These performance patterns hold across both Portuguese (Colonia) and English (CLMET, DTE) datasets, suggesting language-agnostic temporal reasoning.

Variation Few-shot learning offers limited benefit for very short inputs (e.g., DTE 10 tokens), possibly due to noise or ambiguity. Performance occasionally plateaus at the maximum token count (1000), potentially reflecting evaluation variance or long-context handling challenges. While zero-shot performance rankings vary, few-shot prompting more consistently amplifies the strengths of top models like GPT-4o.

4.2 Finetuning

Due to the large data set size and high experimental cost, we only consider the open-source models for the finetuning experiments. We present the results of the experiments with finetuning in Table 4.

Key Findings Open-source LLMs exhibit varied performance using 100-token inputs. Zero-shot F1 scores are modest, generally ranging from 0.26 to 0.36 across while few-shot (5-shot) prompting improves these scores, typically reaching the 0.44 to 0.53 range. Finetuning yields the most substantial gains, achieving F1 scores consistently between 0.61 and 0.77.

Trends We observe that Finetuning provides the highest performance, significantly surpassing prompting methods. Few-shot prompting generally offers a notable improvement over the baseline zero-shot approach. Among the evaluated models, Gemma 2 frequently achieves the highest scores, particularly in the few-shot and finetuned configurations. Mistral also performs strongly when finetuned, matching Gemma 2's top score on the Colonia dataset. A substantial performance gap exists between prompting methods (zero-shot, few-shot) and the finetuning approach across all tested models.

5 Analyzing Performance with Synthetic Data

With the goal of better understanding LLM generation capabilities, we carry out classification experi-

Dataset	SVM	Length	Zero-Shot/Few-Shot (5-Shot)						
			Claude	GPT4o	Gemini	LLaMA	Gemma	Mistral	Nemotron
Colonia	.51	10	.35/.45	.50/**.61**	.39/.50	.12/.11	.18/.25	.17/.24	.16/.23
	.65	50	**.72/.88**	**.68/.72**	.59/.64	.20/.35	.28/.38	.27/.37	.25/.34
	.69	100	.68/**.91**	**.71/.96**	**.77/.82**	.26/.22	.36/.51	.34/.49	.33/.45
CLMET	.70	10	.48/.58	.52/.48	.41/.40	.14/.18	.17/.26	.18/.23	.16/.25
	.77	50	.60/.62	.73/**.85**	.63/.67	.21/.27	.26/.39	.27/.34	.23/.37
	.78	100	**.82/.93**	.78/**.87**	.72/.78	.28/.35	.34/.53	.35/.46	.31/.49
DTE	.71	10	.40/.38	.46/.52	.43/.45	.16/.25	.16/.26	.13/.24	.15/.22
	.72	50	.62/**.85**	.71/**.78**	.57/**.73**	.24/.37	.22/.39	.20/.36	.22/.33
	.72	100	**.76/.74**	**.80/.91**	**.79/.86**	.32/.49	.31/.52	.26/.48	.29/.44

Table 3: Combined `Zero-Shot` and `Few-Shot (5-Shot)` F1 score comparison for seven models on Temporal Text Classification tasks. For the SVM column, TF-IDF vectors are used and the best results are reported.

Model	Colonia	CLMET	DTE	Δ
LLaMA 3.2	0.61	0.65	0.66	+0.29
Gemma 2	0.73	0.75	0.78	+0.23
Mistral	0.72	0.68	0.76	+0.24
Nemotron 4B	0.70	0.67	0.71	+0.23

Table 4: Macro-F_1 for finetuned models on authentic data. Δ is the mean gain over the non-finetuned 5-shot setting across the three datasets (fixed token length of 100).

ments with synthetic LLM-generated test data. We compare the results against the original test sets. Our assumption is that if the classification results using original and synthetic data are similar, this indicates that the models can generate data that best resembles a particular time period.

Curating Synthetic Test Sets We curate 3 test sets, similar to the original ones. We use the 3 proprietary models (GPT4o, Claude, and Gemini) in a few-shot (5) fashion with examples from the original dataset. Since these models are used for generation, we do not include them in the evaluation phase. We only evaluate the open-source models in all 3 experimental setting - zero-shot, few-shot, and finetuned and evaluate on the synthetic test split.

Trends Table 5 reveals that the overall pattern observed on the authentic splits largely persists on the synthetic test sets: scores rise from zero-shot to few-shot to fine-tuned for every model–dataset pair. Gemma 2 and Mistral remain the strongest overall (e.g., Gemma 2 reaches .83 on COLONIA and .84 on DTE), while LLaMA 3.2 shows the largest proportional jump after fine-tuning (e.g., $\Delta = +0.09$ on COLONIA). The Δ columns quantify how much each synthetic score diverges from its corresponding authentic score, and they indicate that gains are generally preserved rather than reversed: models that start strong in few-shot mode tend to stay strong after fine-tuning, even on the newly generated splits.

Variation Variations across datasets and regimes are informative. Zero-shot scores shift more than fine-tuned ones (e.g., Mistral on CLMET sees -0.11 at few-shot but $+0.11$ at fine-tuned), suggesting that parameter updates help stabilize performance when facing synthetic distributions. CLMET exhibits tighter spreads and smaller deltas overall, whereas COLONIA and DTE show broader swings, consistent with their wider temporal ranges and genre diversity. Notably, Nemotron 4B benefits markedly on DTE (fine-tuned $\Delta = +0.12$), showing that even smaller models can capitalize on synthetic supervision when the temporal signal is learnable. Collectively, these shifts imply that synthetic evaluation splits do not scramble the ranking; they accentuate where each model's temporal reasoning is already robust or fragile.

6 Conclusion

This paper presents, to our knowledge, the first focused investigation into the application of LLMs for TTC. Our results demonstrate that leading proprietary models possess strong TTC capabilities accessible via few-shot prompting, with performance scaling effectively with input context length. While open-source models show more limited success with prompting alone, they achieve substantial performance gains through finetuning, significantly closing the gap. Our analysis using synthetic data reveals that while models maintain their rankings,

Dataset	LLaMA	Δ	Gemma	Δ	Mistral	Δ	Nemotron	Δ
Colonia	.21/.31/.70	−.05/ + .09/ + .09	.47/.46/.83	+.11/ − .05/ + .10	.29/.58/.67	−.05/ + .09/ − .05	.24/.52/.62	−.09/ + .07/ − .08
CLMET	.35/.29/.74	+.07/ − .06/ + .09	.25/.62/.70	−.09/ + .09/ − .05	.39/.35/.79	+.04/ − .11/ + .11	.38/.44/.60	+.07/ − .05/ − .07
DTE	.27/.57/.55	−.05/ + .08/ − .11	.42/.41/.84	+.11/ − .11/ + .06	.30/.59/.70	+.04/ + .11/ − .06	.18/.53/.83	−.11/ + .09/ + .12

Table 5: Observed Macro-F_1 scores (*Zero-Shot/Few-Shot(5)/Finetuned*) for open-source models on synthetic test datasets. Adjacent Δ columns show the change from the original scores (same order).

the classification results vary, suggesting that models do not yet perfectly replicate the stylistic nuances of such texts.

Our evaluation shows LLMs have significant potential for TTC. However, results on synthetic test data reveal complex dynamics that require further investigation. Future work will investigate these issues using improved finetuning, prompting, and synthetic data generation techniques, while also extending TTC to more languages and finer temporal granularities.

Acknowledgments

We would like to thank the creators of the corpora explored in this work for making the data available to the research community. We further thank the anonymous NLP4DH reviewers for their insightful feedback.

Limitations

This study provides an initial exploration of LLMs for TTC, focusing on both proprietary (GPT-4o, Claude 3.5) and open-source models (LLaMA 3.2, Gemma 2, Mistral, Nemotron 4B). Our evaluation relies on three existing benchmarks (Colonia, CLMET, DTE) covering English and Portuguese; performance could differ on datasets with varying characteristics, languages, or temporal granularities. The prompting strategies were limited to zero-shot and 5-shot, and finetuning explored specific synthetic data quantities (10-1K). The observed non-monotonic behavior with synthetic data warrants deeper investigation into optimal generation techniques and potential overfitting, which is beyond the scope of this initial work. Further research could also explore parameter-efficient finetuning methods and more diverse prompting approaches. Finally, we further acknowledge that data contamination during model pre-training may have occurred.

Ethical Considerations

Our work adheres to ethical research practices, aligning with principles like those in the ACL Code of Ethics. Our experiments utilize established public datasets (Colonia, CLMET, DTE); users should be aware of any potential biases inherent in these datasets regarding text sources, time periods, or represented samples. The LLMs employed, both proprietary and open-source, may also reflect biases from their original training data, potentially impacting temporal predictions across different contexts. While we used synthetic data for finetuning explorations, we acknowledge that synthetically generated content can inadvertently introduce or amplify biases if not carefully curated.

References

Sidsel Boldsen and Fredrik Wahlberg. 2021. Survey and reproduction of computational approaches to dating of historical texts. In *Proceedings of NoDaLiDa*.

Alina Maria Ciobanu, Liviu P Dinu, Octavia-Maria Şulea, Anca Dinu, and Vlad Niculae. 2013. Temporal text classification for romanian novels set in the past. In *Proceedings of RANLP*.

Corinna Cortes and Vladimir N. Vapnik. 1995. Support-vector networks. *Machine Learning*.

Angelo Dalli and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of ARTE*.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Proceedings of AHC*.

Hendrik De Smet. 2005. The corpus of late modern english texts (clmet).

Kevin Fischer, Darren Fürst, Sebastian Steindl, Jakob Lindner, and Ulrich Schäfer. 2024. Question: How do large language models perform on the question answering tasks? answer. *arXiv preprint arXiv:2412.12893*.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, et al. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Aaron Hurst, Adam Lerer, Adam P Goucher, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tommi Jauhiainen. 2024. Data availability and evaluation reproducibility for automatic date detection in texts, a survey. In *Proceedings of DHNB*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, et al. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of WMT*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, et al. 2015. Statistically significant detection of linguistic change. In *Proceedings of WWW*.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of EACL*.

Sandro Nunes, Cristina Ribeiro, and Gilberto David. 2007. Using neighbors to date web documents. In *Proceedings of WIDM*.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of SemEval*.

Han Ren and Minna Peng. 2025. Tictac: Time-aware supervised fine-tuning for automatic text dating. In *Findings of the ACL (ACL 2025)*.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-aware language modeling for historical text dating. In *Findings of the ACL (EMNLP 2023)*.

Haritz Salaberri, Iker Salaberri, Olatz Arregi, and Beñat Zapirain. 2015. Ixagroupehudiac: A multiple approach system towards the diachronic evaluation of texts. In *Proceedings of SemEval*.

Dominik Schlechtweg, Barbara Mcgillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval*.

Terrence Szymanski and Gerard Lynch. 2015. Ucd: Diachronic text classification with character, word, and syntactic n-grams. In *Proceedings of SemEval*.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*.

Marcos Zampieri, Alina Maria Ciobanu, Vlad Niculae, et al. 2015. AMBRA: a ranking approach to temporal text classification. In *Proceedings of SemEval*.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: The case of portuguese. In *Proceedings of LREC*.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, et al. 2024a. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*.

Zhihan Zhang, Yixin Cao, Chenchen Ye, et al. 2024b. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of ACL*.

A Zero-Shot Prompts

In the zero-shot setting, models receive only a task description and the input text, with no labeled examples. The prompt instructs the model to classify a text snippet into one of the pre-defined temporal classes for the respective dataset. Below we present the zero-shot prompt template for each dataset, followed by a concrete example using an actual instance from the corpus.

A.1 Colonia (Portuguese, 5 Classes)

```
[System] You are a temporal text classifier.
Given a Portuguese text, classify it into the century when it was most likely written.

Classes: 1500s, 1600s, 1700s, 1800s, 1900s

Respond with ONLY the class label.

[User] Classify the following text:

"Desta maneira me achei, e o que mais senti, foi ver que a minha fazenda e o meu
trabalho de tantos annos se perdeo em hum momento; e que me era necessario, sem
remedio algum, tornar a cuidar em meyo de viver, e tratar do sustento da vida."

[Assistant] 1700s
```

Figure 1: Zero-shot prompt for the Colonia corpus. The example shown is a representative 18th-century Portuguese excerpt exhibiting characteristic orthographic features of the period such as *annos*, *hum*, and *meyo*.

A.2 CLMET (English, 3 Classes)

```
[System] You are a temporal text classifier.
Given an English text, classify it into the period when it was most likely written.

Classes: 1710-1780, 1780-1850, 1850-1920

Respond with ONLY the class label.

[User] Classify the following text:

"It is a truth universally acknowledged, that a single man in possession of a good
fortune must be in want of a wife. However little known the feelings or views of such
a man may be on his first entering a neighbourhood, this truth is so well fixed in
the minds of the surrounding families, that he is considered as the rightful property
of some one or other of their daughters."

[Assistant] 1780-1850
```

Figure 2: Zero-shot prompt for the CLMET corpus. The example is an excerpt from a novel included in the Period 2 (1780–1850) sub-corpus. The text reflects the formal prose style characteristic of British Romantic-era fiction.

B Few-Shot Prompts (5-Shot)

In the few-shot setting, we provide five labeled examples (one per class when possible) before presenting the test instance. The examples are randomly sampled from the training split of each dataset and are representative of each temporal class.

B.1 Colonia (Portuguese, 5 Classes)

```
[System] You are a temporal text classifier for Portuguese texts. Given a text,
classify it into the century when it was most likely written.
Classes: 1500s, 1600s, 1700s, 1800s, 1900s. Respond with ONLY the class label.

Here are some examples:

Example 1:
Text: "E assi como chegarom ao dito porto acharom hi gram multidom de gente da
terra que nunca vira navios nem gente branca, e chegarom a praya e comecarom de
olhar os navios com gram espanto"
Label: 1500s

Example 2:
Text: "Nenhuma cousa mais propria da soberba que buscar sempre desculpas, e
nunca confessar a culpa, e tanto maior he a soberba, quanto mais voluntaria
he a cegueira. A soberba he huma torre fundada no ar"
Label: 1600s

Example 3:
Text: "A vaidade dos homens he hum effeito natural, e por isso commum
a todos; nao ha quem a nao tenha em alguma cousa, e os que mais livres
se considerao della, sao talvez os mais vaidosos"
Label: 1700s

Example 4:
Text: "O aggregado foi despedindo os escravos e a fazenda; vendeo tres
molecoens e huma negrinha; desfez-se das melhores joias da mulher;
comeu os alugueis de tres predios; empenhou a prata"
Label: 1800s

Example 5:
Text: "A modinha brasileira e o lundu foram os generos que mais
contribuiram para a formacao da nossa musica popular. Ambos se
aclimataram entre nos desde os tempos coloniais"
Label: 1900s

Now classify the following text:
Text: "O commendador nao approvou a conducta da filha, e censurou-a
asperamente; ella porem nao cedeu, e tanto insistiu que afinal o pae
consentiu, posto que a contragosto"
```

Figure 3: Few-shot (5-shot) prompt for the Colonia corpus. Each example contains a Portuguese excerpt from one of the five centuries represented in the corpus. Example 1 reflects 16th-century exploration chronicles. Example 2 uses Baroque rhetorical style typical of 17th-century sermons. Example 3 mirrors Enlightenment-era philosophical prose. Example 4 depicts 19th-century social realism. Example 5 uses modern 20th-century Brazilian Portuguese.

B.2 DTE (English, Temporal Intervals)

```
[System] You are a temporal text classifier for English news snippets. Given a
text, classify it into the time period when it was most likely written.
Classes: 1700-1720, 1720-1740, 1740-1760, 1760-1780, 1780-1800, 1800-1820,
1820-1840, 1840-1860, 1860-1880, 1880-1900, 1900-1920, 1920-1940, 1940-1960,
1960-1980, 1980-2000, 2000-2020. Respond with ONLY the class label.

Here are some examples:

Example 1:
Text: "At the Court at St.James's, the 29th Day of March, Present, the King's
most excellent Majesty in Council. His Majesty's Declaration of War against
the French King."
Label: 1740-1760

Example 2:
Text: "Receipts at Chicago to-day. Wheats 206 cars; corn fill; oats, 181 cars.
Estimated receipts to-morrow. Wheat, 400 cars; corn, 85 cars; oats, 235 cars;
hogs, 16,000 head."
Label: 1880-1900

Example 3:
Text: "Red Blankets $1.98 a pair. White Blankets 69c a pair. Bed Comforts 69c
each. Heavy Knit Skirts 69c each."
Label: 1900-1920

Example 4:
Text: "By 1971 about one-third of Edison's electric output will be generated
with nuclear capacity"
Label: 1960-1980

Example 5:
Text: "Occasional selfies are acceptable, but uploading a new picture of
yourself every day is not necessary."
Label: 2000-2020

Now classify the following text:
Text: "We have cabled the English house from which we get it and expect a reply
to-morrow."
```

Figure 4: Few-shot (5-shot) prompt for the DTE corpus. All examples are representative of actual snippets from the SemEval-2015 Task 7 dataset (Popescu and Strapparava, 2015). Temporal cues include institutional references (Example 1), archaic spelling such as *to-day* and *to-morrow* (Examples 2 and test instance), period-appropriate pricing (Example 3), mentions of then-emerging technology (Example 4), and modern vocabulary such as *selfies* (Example 5).

C Fine-Tuning Input Format

For fine-tuning open-source models using LoRA (Hu et al., 2021), we format the training data as instruction-response pairs compatible with the Alpaca format. The system instruction remains consistent across all training examples, while the input text and expected output vary per instance. Table 6 presents the template along with an example from each dataset.

Field	Content
Instruction	Classify the following text into one of the given temporal classes: {class_list}. Respond with only the class label.
Input (Colonia)	"O aggregado foi despedindo os escravos e a fazenda; vendeo tres molecoens e huma negrinha; desfez-se das melhores joias da mulher; comeu os alugueis de tres predios; empenhou a prata"
Output	1800s
Input (CLMET)	"The village lay in a hollow, which was, perhaps, one reason for its dullness; and the church, standing on higher ground, overlooked it with a certain air of superiority"
Output	1850-1920
Input (DTE)	"We have cabled the English house from which we get it and expect a reply to-morrow."
Output	1880-1900

Table 6: Fine-tuning input format used for training open-source models via LoRA. Each training instance follows the instruction-input-output template. The {class_list} placeholder is replaced with the appropriate temporal classes for each dataset: 5 century-level classes for Colonia, 3 period classes for CLMET, and interval-based classes for DTE.

D Synthetic Data Generation Prompt

For generating synthetic test data (Section 4), we prompt proprietary models in a 5-shot fashion to generate text that resembles a given time period. The prompt instructs the model to produce a text snippet stylistically consistent with the target temporal class.

```
[System] You are an expert in historical linguistics. Generate a text passage that
is stylistically consistent with the specified time period. The passage should
reflect the vocabulary, syntax, and writing conventions typical of that era.

[User] Generate a Portuguese text passage (approximately 100 tokens) in the style
of the 1600s.

Here are some reference examples from that period:

[5 labeled examples from the 1600s class]

Generate a new text similar to these examples.
```

Figure 5: Prompt template for synthetic test data generation. Reference examples are drawn from the training split of each dataset. This template is adapted for each corpus by adjusting the language, temporal class, and reference examples accordingly.

Evaluating Latin and Ancient Greek Sentence Alignment through Parallel Sentence Mining

Sebastian Reichbauer[1] **and Shu Okabe**[1,2] **and Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology, Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
Corresponding authors: {sebastian.reichbauer,shu.okabe}@tum.de

Abstract

Cross-lingual detection of intertextuality and translation in Latin and Ancient Greek through computational approaches is of great interest for classical studies. While several systems exist for parallel sentence detection, based on general multilingual or specific models for Latin–Ancient Greek, they have not been compared against each other. Therefore, we present a synthetic benchmark to evaluate the performance of language models regarding cross-lingual Ancient Greek and Latin parallel sentence mining. We first compare six language models to encode sentences and then further improve the cross-lingual alignment through post-processing, fine-tuning, and knowledge distillation. We find that the whitening transformation in combination with knowledge distillation provides excellent results. Specifically, SPhilBERTa, a trilingual language model for Ancient Greek and Latin, benefits the most from the improvements and achieves a notable mining score of 97.6 on our benchmark.

1 Introduction

Identifying instances of semantic similarity across texts is a major task in Classics and Philology, often phrased as intertextuality detection. Manual spotting of such passages is laborious, however, and will almost certainly stay incomplete, since scholars cannot memorize the whole corpus of Latin and Greek[1] texts to detect such paraphrases. Therefore, tools emerged to facilitate this task by applying computational concepts to automatically detect paraphrase candidates (Coffee et al., 2013). The more modern approaches rely mostly on the use of embeddings generated by language models, which have already shown promising results (Riemenschneider and Frank, 2023b). Nevertheless, only a reliable representation of tokens or sentences in the embedding space allows meaningful comparison between text passages.

To obtain reliable cross-lingual representations in language models, substantial training data from parallel corpora is needed. However, such resources are scarce for Latin and Greek, compared to modern and better-resourced language pairs. For instance, as of April 2025, there are fewer than 50 parallel sentences listed for the language pair on OPUS (Tiedemann, 2012), and open-access corpora for Classical Latin and Greek remain limited. Moreover, if Latin can be considered as a mid-resource language in Natural Language Processing (cf. language taxonomy from Joshi et al. (2020)), *Ancient* Greek remains extremely low-resource, with little to no representation in common multilingual language models.

In this article, we hence focus on retrieving cross-lingual Greek–Latin *translations*, as a first step towards paraphrase detection. We explore mining approaches which require few to no direct Greek–Latin parallel sentences for training, as this language pair faces the same challenges as other low-resource languages. Following an established approach, we first generate a bitext mining dataset to evaluate the Greek–Latin cross-lingual alignment. We compare six sentence encoding approaches as a backend of a mining pipeline. Moreover, we apply two types of post-processing to the sentence representations, whitening (Huang et al., 2021) and cluster-based isotropy enhancement or CBIE (Rajaee and Pilehvar, 2021), to improve cross-lingual alignment without training. We also consider a standard fine-tuning with limited parallel sentences in Greek–Latin from the Post-Medieval period and two knowledge distillation approaches (Reimers and Gurevych, 2020), where English or Latin are considered as anchor languages.

Our study extends Riemenschneider and Frank (2023b), which considered the task of parallel sentence *matching*. There, 1,000 parallel sentences

[1]'Greek' refers to Ancient Greek in this article, while Modern Greek is indicated by 'Modern Greek'.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 106–120
July 4, 2026 ©2026 Association for Computational Linguistics

were shuffled and had to be paired back. Our mining benchmark is more challenging, as it features 2,000 translation pairs within a 23k corpus. Our contributions are the following:

(i) We create a synthetic benchmark to evaluate performance in cross-lingual Greek–Latin similarity detection.

(ii) We systematically compare six off-the-shelf language models for parallel sentence mining.

(iii) We study how to further improve mining quality with and without parallel sentences through post-processing, fine-tuning, and knowledge distillation.

(iv) We release the best-performing model and an open-source version of the benchmark.[2]

2 Related Works

Language Models for Greek and Latin A variety of monolingual language models for Greek and Latin exist, such as Latin BERT (Bamman and Burns, 2020) for Latin, while Krahn et al. (2023) provided BERT models for Greek, using knowledge distillation. Language models with explicit cross-lingual Greek–Latin training and understanding are much scarcer, however. Yousef et al. (2022) introduced Ugarit, a model based on XLM-R (Conneau et al., 2020), a multilingual language model, which was fine-tuned with both monolingual data and parallel sentences. Riemenschneider and Frank (2023a) introduced PhilBERTa, a trilingual language model for Greek, Latin and English. Based on this latter model, SPhilBERTa (Riemenschneider and Frank, 2023b) takes advantage of the knowledge distillation approach with English as an anchor language to create a language model for Greek–Latin intertextuality detection.

Parallel Sentence Mining To compare these methods, a suitable evaluation system must be designed to measure model performance in Semantic Textual Similarity (STS) tasks. The BUCC shared task (Zweigenbaum et al., 2017) introduced a synthetic mining corpus to evaluate the ability of models to detect parallel sentences for four high-resource language pairs. Following a similar protocol, Okabe et al. (2025) devised a mining benchmark for low-resource language pairs. A standard pipeline for mining relies on bilingual or multilingual sentence embeddings that are compared according to similarity metrics, such as the CSLS metric (Conneau et al., 2018). This score remains more robust and suited for mining than the cosine similarity, as it mitigates the hubness problem (Dinu et al., 2015).

Anisotropy One notable issue in multilingual language models remains the quality of cross-lingual alignment. Hämmerl et al. (2023) particularly explored the effects of anisotropic representations for STS tasks. Anisotropy thereby denotes the unbalanced use of the available vector space by embeddings, with some parts heavily occupied, while others remain largely empty. High anisotropy was found to induce significantly deteriorated performance for parallel sentence mining and matching tasks, explained by the poor cross-lingual alignment in the multilingual semantic space.

Thus, two approaches were considered to make the representations more isotropic without requiring parallel data. Whitening (Huang et al., 2021) lowers anisotropy by applying a linear transformation on the vector matrix to make the covariance matrix become the identity matrix. Another approach is CBIE (Rajaee and Pilehvar, 2021), which detects clusters in the embeddings, then identifies and removes the top x principal components in each cluster.

Knowledge Distillation Since direct parallel sentences between Greek and Latin are scarce, while English translations of Greek and Latin texts are more abundant, knowledge distillation (Reimers and Gurevych, 2020) can effectively extend the model's coverage to further languages. Distillation works by using a teacher model that creates high-quality embeddings in one language (e.g., English), and a student model that aims to represent parallel sentences (e.g., Greek–English or Latin–English) with embeddings closer to those of the teacher model. The student model can thus leverage the teacher model's robust language modelling without requiring extensive training data, and additionally keeps the additional capabilities from its initial multilingual embedding space. SPhilBERTa (Riemenschneider and Frank, 2023b) actually stems from distillation, where an English monolingual sentence encoder serves as teacher to extend a multilingual student model.

[2] https://github.com/TUM-NLP/latin-greek-mining

3 Benchmark Creation

We evaluate the quality of parallel sentence mining for Greek–Latin by creating a synthetic corpus. We follow the BUCC Shared Task approach (Zweigenbaum et al., 2017), where actual parallel sentences are injected into two monolingual corpora (one in each language). The data is split into a train partition that is used to find suitable mining parameters. Those are then applied to the test partition, which ultimately determines performance.

3.1 Datasets

Parallel Sentences We chose to focus on ancient Latin and Greek (i.e., not medieval or more recent) for parallel sentences to ensure measuring the performance of models in both language periods. This means that modern Latin translations (e.g., from the 19th century) of Greek texts are ruled out. Consequently, we only use 2,000 sentences from Boethius' translation in Latin of works of Aristotle on logic as parallel data, particularly *Analytica Priora* and *Topica*. Other works are not selected due to the limited online availability of texts or the high manual labour required to align them properly. The Bible has been specifically excluded because many models have already been trained on it, thereby increasing the risk of data contamination and reducing the meaningfulness of the results.

Monolingual Sentences Monolingual sentences should be as similar as possible semantically to increase the difficulty in detecting parallel sentences, while not creating accidental translation pairs. To achieve this, we chose to use authors largely from the same periods as Boethius and Aristotle, to match the language style, while keeping the topics different. Therefore, as shown in Table 1, we selected approximately 23,000 sentences in Greek and Latin from authors mostly from Late Antiquity (Latin) and the Classical Period (Greek). The topics include mostly historiography and theological texts, as well as philosophy, leaving out analytical works about logic. All text sources are detailed in Table 7 of Appendix A.

3.2 Preprocessing

Sentences should not be distinguishable with regard to formal features to prevent models from detecting parallel sentences due to their similar external appearance. Therefore, we preprocess our selected sentences as follows: The sentence length for monolingual as well as parallel sentences is bounded by a minimum of 10 and a maximum of 50 words. This provides a similar word length per sentence, as shown in Table 1. Additionally, due to the large amount of combinable characters in the Greek alphabet, we apply an NFC normalization on the data to ensure the same byte representations for the same characters. Latin sentences that cite Ancient Greek using the Greek alphabet are discarded to prevent recognition of the same alphabet as parallel. Excess whitespaces are removed, and Latin texts are normalized orthographically. This means specifically that spellings with 'j' are automatically replaced by 'i'. We also ensure manually that no texts using the orthographic variant of 'e' instead of 'ae' as casus endings (for genitive, for instance) are included.

	Latin		Greek	
	mono.	par.	mono.	par.
N_{sent}	22,727	2,000	21,641	2,000
$N_{words/sent}$	22.67	19.68	22.36	21.76

Table 1: Statistics of our synthetic benchmark datasets. N_{sent} indicates the number of sentences, while $N_{words/sent}$ denotes the average number of words per sentence. Parallel sentences (par.) are not counted in the monolingual sentence number (mono.).

4 Mining without Additional Data

We now detail our mining methodology, where no additional training data is needed by default.

4.1 Mining Pipeline

We use the parallel sentence mining pipeline suitable for BUCC-style corpora from Okabe and Fraser (2025) to evaluate the out-of-the-box performance of language models. Once sentences are converted into embeddings, we compute the CSLS metric (Conneau et al., 2018) for every combination of source and target sentence representation, as presented in Equation (1):

$$CSLS(x,y) = 2\cos(x,y) - \sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x,z)}{k} - \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y,z)}{k} \quad (1)$$

where $NN_k(x)$ denotes the k nearest neighbours of a vector x. We use this method, as it has been shown to achieve better results than the standard cosine similarity. We follow Okabe and Fraser

(2025) and set k to 20. Each source sentence is then paired with its closest target sentence based on their CSLS score. Since many sentences are not parallel and therefore do not have a corresponding sentence in the target corpus, we rely on a dynamic threshold, as defined by Hangya et al. (2018):

$$\theta = \text{mean}(S) + \lambda \times \sigma(S) \quad (2)$$

where S denotes the similarity scores and σ the standard deviation over S. If sentence pairs achieve a value greater than this threshold θ, they are accepted as parallel. λ serves as a hyperparameter here, which can be adjusted on the train split of the BUCC-style corpus, as described above, to maximise the sentence mining quality.

We evaluate the mining results according to the standard Precision, Recall, and F-score.

4.2 Base Benchmark

To assess the off-the-shelf performance of different models utilizing the described pipeline, we select specific models and conduct a baseline comparison on the created corpus. This provides comparability between the base versions of the considered models and underlines possible improvement paths.

4.2.1 Multilingual Language Models

We first consider multilingual language models to represent sentences, as they may have been trained on Greek or Latin. We include the following models in the benchmark:

Token-level models XLM-R (Conneau et al., 2020) is a token-level language model trained on 100 languages, including Latin (390 million tokens), but not Greek. Glot500 (Imani et al., 2023) extends XLM-R to support many low-resource languages and may therefore be well-suited for this task. It has been pre-trained on 1.2 million Latin and 380,000 Greek *monolingual* sentences.

Sentence encoder LaBSE (Feng et al., 2022), in contrast, is a state-of-the-art sentence encoder that was not pre-trained on Greek (only Modern Greek), but is trained to specifically set non-similar sentences apart in the embedding space. It has notably been pre-trained using English–Latin *parallel* sentences. It serves as a reliable reference point for the Greek–Latin mining performance of multilingual sentence encoders.

LLM-based embeddings Qwen3-embeddings (Team, 2025) is an LLM-based approach to obtain text embeddings. It consequently allows for examining the ability of LLMs for the mining task. To maintain comparability with the other models, we choose to use its rather smaller version, with 0.6 billion parameters and 1,024 dimensions, even though versions with up to 30 billion parameters exist. We denote this model Qwen below.

4.2.2 Ancient Language Specific Models

We choose Ugarit (Yousef et al., 2022) and SPhilBERTa (Riemenschneider and Frank, 2023b) as representatives of models that have been specifically trained for Greek–Latin semantic similarity detection. Ugarit is based on XLM-R and has been fine-tuned on 12,000,000 monolingual Ancient Greek tokens, 8,000 Greek–Latin, and 32,500 English–Greek parallel sentences. SPhilBERTa is based on PhilBERTa (Riemenschneider and Frank, 2023a) and fine-tuned via knowledge distillation using English–Latin and English–Greek parallel sentences, with English as the anchor language.

4.3 Isotropy-Enhancing Transformations

Anisotropy analysis We analyze the anisotropy of the best three models from the base benchmark. Following Hämmerl et al. (2023), we measure three properties: First, we examine the total anisotropy, defined as the mean cosine similarity over all considered vectors. Second, we assess the contribution of individual dimensions to the total anisotropy. It is defined by a partial cosine similarity, as shown in Equation (3).

$$CC(f_i) = \frac{1}{n} \sum_{\{x,y\} \in C} CC_i(f(x), f(y)) \quad (3)$$

where $CC_i(u, v)$ equals the contribution of the individual dimension i of two vectors u and v to the total cosine similarity and is defined by $CC_i = \frac{u_i v_i}{\|u\| \|v\|}$. Third, we count the number of outlier dimensions, i.e., those with values greater than 3 times the standard deviation of all dimensions. We apply these measures both before and after our transformations.

Whitening We apply ZCA whitening (Huang et al., 2021) to the three best models of the base benchmark. It is defined by Equation (4):

$$\hat{E} = (E - m)UD^{-\frac{1}{2}} \quad (4)$$

where $E \in \mathbb{R}^{n \times d}$ is the embedding matrix to be whitened, subtracted by the mean vector m. The covariance matrix $cov(E)$ is decomposed into $E = UDU^T$, with D as a diagonal matrix of eigenvalues and U as a matrix of eigenvectors.

CBIE Additionally, we apply CBIE (Rajaee and Pilehvar, 2021) to the same model representations to compare it to whitening, closely following their approach. First, we cluster the embeddings using k-means into 7 partitions. Then, we detect the top 12 principal components using PCA and zero-mean the clusters. Then, the identified principal components are removed for each cluster.

5 Results without Additional Data

5.1 Off-the-shelf Results

Model	Precision	Recall	F-score
XLM-R	0.2	0.1	0.1
Glot500	8.3	4.7	6.0
Qwen	44.9	35.3	39.6
SPhilBERTa	46.4	39.0	42.4
Ugarit	77.4	59.4	67.2
LaBSE	**79.9**	**64.6**	**71.4**

Table 2: Mining results on our synthetic benchmark for six sentence encoding approaches.

Table 2 displays the results of the six sentence encoding approaches on our mining benchmark.

Glot500 and XLM-R We can observe that Glot500 and XLM-R both show poor performance. This was expected for XLM-R, as it also performed poorly in Riemenschneider and Frank (2023b). Glot500 is based on XLM-R and might therefore yield poor results. However, the Ugarit model is based on XLM-R as well. Consequently, we see that the massively multilingual pre-training of Glot500 is of little help for our task, while Ugarit achieves far superior results thanks to its specific training for Greek–Latin similarity detection.

Qwen embeddings Qwen achieved a reasonable result of 39.6. Due to its lower parameter count (0.6 billion) compared to other Qwen versions with up to 30 billion parameters, LLMs with more parameters might perform better. However, 0.6 billion parameters are more than all the other language models we tested have, as LaBSE, the model with the second-most parameters, has only 0.5 billion.

Model	Base	Whitened	CBIE
LaBSE	0.20	$2.58 \cdot 10^{-5}$	$9.29 \cdot 10^{-6}$
SPhilBERTa	0.40	$3.05 \cdot 10^{-5}$	$2.93 \cdot 10^{-5}$
Ugarit	0.76	$2.42 \cdot 10^{-5}$	$1.10 \cdot 10^{-5}$

Table 3: Anisotropy of selected models out-of-the-box and after applying whitening and CBIE.

The embedding dimension is also higher for Qwen, compared to LaBSE (768).

SPhilBERTa and Ugarit SPhilBERTa and Ugarit show vastly different performance despite both being trained specifically for cross-lingual similarity detection in Greek and Latin. The main difference lies in the training method. SPhilBERTa relies on knowledge distillation, while Ugarit uses parallel sentences for training, inferring that fine-tuning with parallel sentences might be better suited for this task.

LaBSE LaBSE shows the best performance with 71.4 points, better than Ugarit or SPhilBERTa, although it has not been trained on Greek. This might be due to its training method, which separates dissimilar sentences on large corpora and has been proven valuable for sentence mining tasks, even for unseen languages.

5.2 Anisotropy Study

Table 3 presents the three anisotropy measures before and after CBIE and whitening transformations. We can observe that LaBSE shows the lowest anisotropy, while Ugarit exhibits the highest by far. The low anisotropy by LaBSE can be explained by its training method, which aims to move dissimilar sentences apart and thereby utilizes the embedding space better. Ugarit's high anisotropy is not surprising either, since Hämmerl et al. (2023) discovered that its base model, XLM-R, showed very high anisotropy, which seems to be inherited here. After applying whitening as well as CBIE, the anisotropy is reduced to almost 0. Both transformations, consequently, are effective in lowering the anisotropy of embedding spaces. Appendix B displays the full results in Table 8, and a visual illustration of the measured anisotropy is shown in Figure 2.

5.3 Whitening and CBIE

Table 4 shows the results of applying whitening and CBIE on the embeddings of the mining benchmark. As expected, LaBSE does not profit much

from these transformations, as it already had low anisotropy and better cross-lingual alignment. In contrast, SPhilBERTa shows a significant performance increase with both whitening and CBIE, more than expected from its moderate anisotropy, especially in comparison to Ugarit. Despite the similar performance of SPhilBERTa with both methods, it is clear that whitening is the superior choice for Greek–Latin similarity detection with these models. It yields better results than CBIE across all models. This is specifically pronounced when excluding SPhilBERTa. Ugarit with whitening emerges now as the best model on our benchmark, with a score of 86.5.

Model	Precision	Recall	F-score	Δ_F
Whitening				
LaBSE	71.3	68.9	70.0	– 1.4
Ugarit	**89.4**	**83.9**	**86.5**	+19.3
SPhilBERTa	84.0	68.4	75.4	+33.0
CBIE				
LaBSE	71.4	49.3	58.3	–13.1
Ugarit	80.9	57.3	67.1	– 0.1
SPhilBERTa	80.5	67.3	73.3	+30.9

Table 4: Mining results after applying whitening and CBIE. Δ_F indicates the difference with the base F-score.

6 Fine-tuning

We now explore a standard fine-tuning setting to leverage a small Greek–Latin parallel corpus.

6.1 Methodology

Dataset We use around 10,000 direct parallel sentences for fine-tuning. 6,612 sentences stem from the DFHG corpus,[3] which has already been used to fine-tune Ugarit. Another 3,587 sentences come from Thucydides' *Peloponnesian War* with a modern Latin translation. By applying a minimum word count per sentence of 3, the sentence number is reduced to 10,013.

Training Settings We employ the CachedMultipleNegativesRankingLoss as the objective function. Since the MultipleNegativesRankingLoss seemed to work quite well for LaBSE, we also use it here in its cached version to support large in-batch sizes despite limited VRAM. The in-batch size is indeed

[3] https://www.dfhg-project.org/

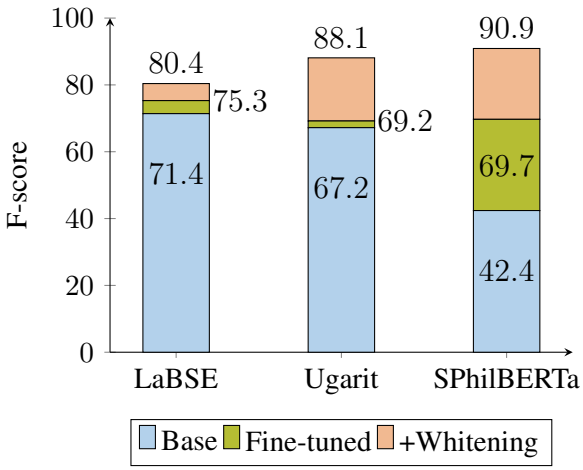


Figure 1: F-scores of the base benchmark (blue), after fine-tuning (green) and after applying whitening to the fine-tuned results (red).

crucial to the performance of the MultipleNegativesRankingLoss, since it works by pulling dissimilar sentences within a single batch apart. The batch size is set to 256, and we fine-tune with two epochs to account for the limited amount of training data. This training method is applied to the best three models of the base benchmark.

Whitening Since whitening improved the mining performance of the off-the-shelf models, we also examine whether this effect remains as pronounced *after* fine-tuning.

6.2 Mining Results

Fine-tuned Results Figure 1 presents the results with fine-tuning. One key takeaway is that all models improve at least slightly relative to the base performance. SPhilBERTa exhibits a significant score increase by 27.3. Ugarit's small improvement can be attributed to data redundancy, since the sentences from the DFHG dataset were already used for training Ugarit. The marginal improvement by LaBSE is surprising, however, since it has not been trained on Greek–Latin sentences. Nevertheless, since it uses a similar objective function, that positive effect might lead to other models profiting more than LaBSE, which has already experienced the effect of this loss function.

Results after Whitening After whitening, both Ugarit and SPhilBERTa perform much better, while LaBSE only increases its score by a moderate margin. This is in line with our expectations, as LaBSE showed the lowest anisotropy and therefore did not benefit much from isotropy-enhancing meth-

ods such as whitening. SPhilBERTa specifically proves to be the new best model with 91 points on our benchmark. We see that fine-tuning with parallel sentences yields positive results, especially when combined with whitening, which retains its effectiveness even in already fine-tuned models.

7 Knowledge Distillation

7.1 Methodology

As Riemenschneider and Frank (2023b) explored knowledge distillation to obtain a reliable Greek–Latin model, we also explore combining the strengths of two chosen models. Distillation aims to take advantage of a teacher model M with an advanced representation of a specific anchor language. The student model $\hat{M}$ is trained to mimic the embeddings of the teacher model. This is done by using parallel sentences between the target language and the anchor language with the loss function in Equation (5):

$$MSE(B) = \frac{1}{|B|} \sum_{j \in B} \Big[(M(s_j) - \hat{M}(s_j))^2 + (M(s_j) - \hat{M}(t_j))^2 \Big] \quad (5)$$

where $M(s)$ indicates the embedding of an anchor language sentence s in batch B from model M, while t denotes a sentence from the target language. The anchor language embeddings as well as the target language embeddings are shifted into the embedding space of the teacher model, thereby replicating the semantic space of the anchor language in the teacher model. Consequently, a more robust cross-lingual representation is achieved without the need for full retraining.

Experiment Details We examine both English and Latin as anchor languages. We chose English as it is a high-resource language with much data available, while Latin is one of the target languages. With the latter, we can also compare its performance with the English-mediated approach. We selected Latin over Greek due to the higher amount of data on which models have been trained.

As a teacher model, we use LaBSE due to its better English representation comparatively. It is indeed crucial to have a teacher model that has a reliable representation of the anchor language. Since its embeddings are mimicked by the student model, their quality largely influences the result. Additionally, we use the fine-tuned version of SPhilBERTa, which we call SPhiTune, as teacher model. This provides a reference point for models with stronger abilities in one of the target languages, here, Latin. We note that it also scored second in Section 6.2.

As student models, we rely on Glot500 and SPhilBERTa, the former for its broad multilingual understanding, the latter for its already specialized training on data relevant to the Greek–Latin pair.

All experiments with English serving as an anchor language are conducted with 1 epoch, while for Latin, we train with 2 epochs due to the limited amount of training data.

Given the promising results of whitening, we also examine whether it can further improve performance. We employ it both after distillation and for the teacher embeddings. Applying it to the teacher embeddings should lower the anisotropy and pass that effect to the student models. Appendix D presents further reproducibility details.

English-based Distillation For English as an anchor language, we use Greek–English and Latin–English parallel sentences for distillation, similarly to SPhilBERTa. Since the two language pairs feature relatively larger datasets, we rely on 380,000 Greek–English sentences from Krahn et al. (2023) and 100,000 Latin–English sentences.[4]

Latin-based Distillation When we consider Latin as an anchor language, we use the same 10,000 sentences as for fine-tuning (Section 6.1). The training data, therefore, is not comparable in size to using English as an anchor language, but covers both of our languages of study.

7.2 Mining Results with Distillation

English Table 5 presents selected results of knowledge distillation with English as an anchor language. The full results are displayed in Table 9 of Appendix C. We can observe that knowledge distillation through English is indeed an effective approach for better mining performance. LaBSE performs better as a teacher model than SPhiTune. This is not surprising, since the relevant property for teacher models is their representation of the anchor language, where LaBSE excels, while SPhiTune was not trained for English understanding.

Applying whitening afterwards leads to another notable performance increase. SPhilBERTa fine-tuned with knowledge distillation can profit sim-

[4] https://huggingface.co/datasets/grosenthal/latin_english_translation

Teacher model	Precision	Recall	F-score	Δ_F
Base Result				
LaBSE	**91.2**	**83.3**	**87.1**	+44.7
SPhiTune	68.6	61.5	64.9	+22.5
Whitening applied afterwards (to "Base Result")				
LaBSE	**98.0**	**97.1**	**97.6**	+55.2
SPhiTune	91.3	74.6	82.1	+39.7

Table 5: Results of knowledge distillation with **English** as an anchor language, before and after whitening. We compare two *teacher* models (LaBSE and SPhiTune) for the same student model (SPhilBERTa). Δ_F indicates the difference to the base benchmark score of the *student* model.

ilarly from whitening as its base version and receives the best result yet. The same experiment has been conducted with Glot500 as student model. However, it exhibited low performance. This is probably due to its worse representation of Greek and Latin, as already seen in the base benchmark. Hence, we left it out for this analysis and report it in Appendix C. We additionally report the results with whitening applied to the teacher embeddings there, due to its small influence on mining performance.

Teacher model	Precision	Recall	F-score	Δ_F
Base Result				
LaBSE	42.6	36.8	39.5	– 2.9
SPhiTune	**72.5**	**63.1**	**67.5**	+25.1
Whitening applied afterwards (to "Base Result")				
LaBSE	**92.0**	81.5	86.4	+44.0
SPhiTune	91.8	**89.0**	**90.4**	+48.0

Table 6: Results of knowledge distillation with **Latin** as an anchor language, before and after whitening. Δ_F indicates the difference to the base benchmark score of the *student* model (SPhilBERTa).

Latin We show selected results of knowledge distillation with Latin as an anchor language in Table 6. The full results are displayed in Table 10 of Appendix C. We note that Latin as anchor language leads to worse results in almost every case. This is in line with our expectations, however, because the sentence representations of LaBSE and SPhiTune turn out worse than those for English as a high-resource language, due to the limited amount of training data. Another factor is the difference in the size of the parallel datasets, with 10,000 sentences for Latin as anchor language and 480,000 for English.

We also observe that here SPhiTune seems to perform better than LaBSE as teacher model. This is due to the specific training SPhiTune received regarding the Latin language. This allows it to create better Latin sentence embeddings than LaBSE, which remains multilingual with no language-pair-specific pre-training. Here also, Glot500 shows worse scores than SPhilBERTa on our benchmark (cf. results in Table 10 of Appendix C). Since applying whitening to the teacher embeddings led to negligible differences, we also report those results in the Appendix. The smaller amount of available parallel sentences, combined with poorer language representation than English, makes Latin less suitable as an anchor language for distillation.

In short, knowledge distillation in combination with whitening appears to be a promising approach to improve cross-lingual alignment. Here, English seems to be a better anchor language, whereas using Latin is less effective in our case.

Both fine-tuning with parallel sentences and knowledge distillation do enhance the model performance on our synthetic benchmark. However, the results of knowledge distillation through English are much better than those of direct fine-tuning. This is not surprising, since the data available for knowledge distillation with English as anchor language is also much more abundant than Greek–Latin parallel sentences.

8 Conclusion

This article assessed Latin and Greek sentence alignment through parallel sentence mining and compared several enhancement methods based on differing data requirements. First, we created a synthetic benchmark to evaluate the performance of language models in cross-lingual Greek–Latin similarity detection. Using this benchmark, we analyzed a diverse set of six models regarding their suitability for mining. LaBSE, Ugarit, and SPhilBERTa achieved here the best off-the-shelf performance. We then compared post-processing methods without additional data to improve cross-lingual alignment, namely CBIE and whitening. We found that whitening consistently outperforms CBIE and strongly improves mining performance. Fine-tuning and knowledge distillation are both valuable methods for improving models to leverage additional available parallel data. In our setting, knowledge distillation with English as an anchor

language appeared as more efficient, mainly thanks to larger available datasets and better language representation. Finally, combining knowledge distillation with whitening leads to the best results. Specifically, SPhilBERTa with LaBSE as a teacher achieves a significantly high F-score of 97.6 on our synthetic benchmark.

Future work will extend the limited set of language models evaluated in this article. Additionally, our best mining approaches can be used to find additional parallel sentences from monolingual corpora. Another problem that remains to be tackled is to find a method to evaluate the performance of models for detecting actual semantic similarity as in paraphrases rather than translations.

Limitations

The main limitation of this study comes from the bias when selecting the datasets to create our synthetic benchmark. As suitable corpora are scarce, we focused on selected authors and time periods for both monolingual and parallel sentences. This choice, however, guarantees that the created benchmark is consistent and that no clear divergence can be found between the two types of sentences. As writer style, language period, and domain are all influential factors in mining, we report all the sources used for our benchmark creation.

Our benchmark also relies on a synthetic approach, instead of a true mining setting, with comparable corpora. Nevertheless, the methodology is established through the BUCC Shared Task and subsequent endeavours for bitext mining evaluation.

Acknowledgments

We thank the anonymous reviewers for their comments. This work was partly funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *Preprint*, arXiv:2009.10053.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2013. The tesserae project: intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28(2):221–228.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the Workshop Track at ICLR*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 7–13, Brussels. International Conference on Spoken Language Translation.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117,

Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kevin Krahn, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence embedding models for Ancient Greek using multilingual knowledge distillation. In *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shu Okabe and Alexander Fraser. 2025. Bilingual sentence mining for low-resource languages: a case study on upper and Lower Sorbian. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 11–19, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Shu Okabe, Katharina Hämmerl, and Alexander Fraser. 2025. Improving parallel sentence mining for low-resource and endangered languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–205, Vienna, Austria. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023a. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023b. Graecia capta ferum victorem cepit. detecting Latin allusions to Ancient Greek literature. In *Proceedings of the Ancient Language Processing Workshop*, pages 30–38, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022. An automatic model and gold standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Dataset sources details

Table 7 displays the detailed *monolingual* sources used to create our synthetic Latin and Ancient Greek benchmark.

B Anisotropy results

Figure 2 presents the t-SNE representations of Latin and Ancient Greek parallel sentences for three models: Ugarit, LaBSE, and SPhilBERTa. Moreover, Table 8 details our anisotropy study before and after transformations.

C Knowledge Distillation Results

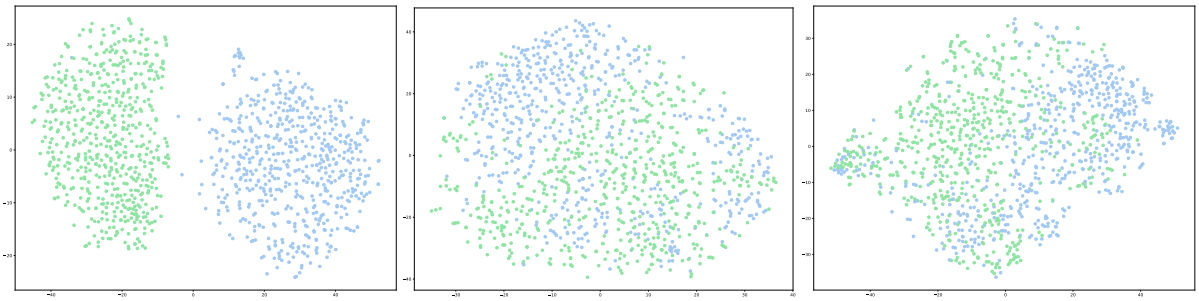
Tables 9 and 10 present the full results for our knowledge distillation experiments with English or Latin as an anchor language.

D Reproducibility

For training, we used one H100 GPU with 94GB HBM2 VRAM. Fine-tuning with Greek–Latin parallel sentences took only minutes to execute (all specifications for one model). It took a similar length of time to apply knowledge distillation with the same sentence pairs. For knowledge distillation with English as the anchor language, much more time was needed due to the larger dataset. The training took around 2 hours on the same GPU.

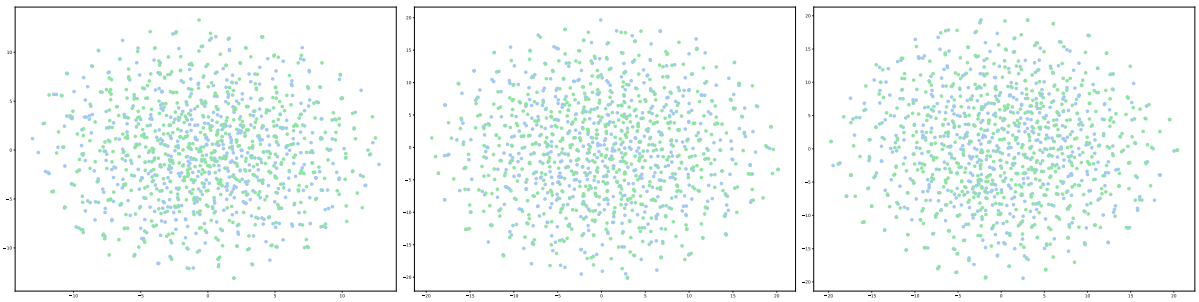
Author	Work	Domain	N_{sent}	$N_{words/sent}$
		Latin		
Ammianus Marcellinus	*Res gestae*	Historiography	3,621	26.38
Augustinus	*Confessiones*	Theology	2,413	22.33
	De civitate dei	Theology	9,898	
	De fide et symbolo	Theology	174	
Boethius	*De fide catholica*	Theology	55	22.30
	Liber de persona et duabus naturis	Theology	184	
	De trinitate	Theology	103	
Cassiodorus	*Variae*	Administrative Literature	4,483	20.99
	De anima	Theology	438	
Jordanes	*Getica*	Historiography	665	22.09
	Romana	Historiography	693	
Total			22,727	22.67
		Ancient Greek		
Xenophon	*Anabasis*	Historiography	2,311	20.19
	Apologia	Philosophy	88	
	Oikonomikos	Economy	661	
	Hellenika	Historiography	2,679	
	Cyropaedia	Biography	3,110	
	Memorabilia	Philosophy	1,274	
	Symposion	Philosophy	380	
Demosthenes	*Phillipika*	Speech	96	22.23
Polybios	*Historia*	Historiography	9,706	24.40
Isokrates	*Panathenaikos*	Speech	321	26.76
Galen	*Peri Physikon Dynameon*	Medicine	1,015	24.00
Total			21,641	22.36

Table 7: Summary of the most important properties of the **monolingual** data selected, Latin at the top, Ancient Greek at the bottom. $N_{words/sent}$ describes the average number of words per sentence. For comparison, 2,000 parallel sentences are used.



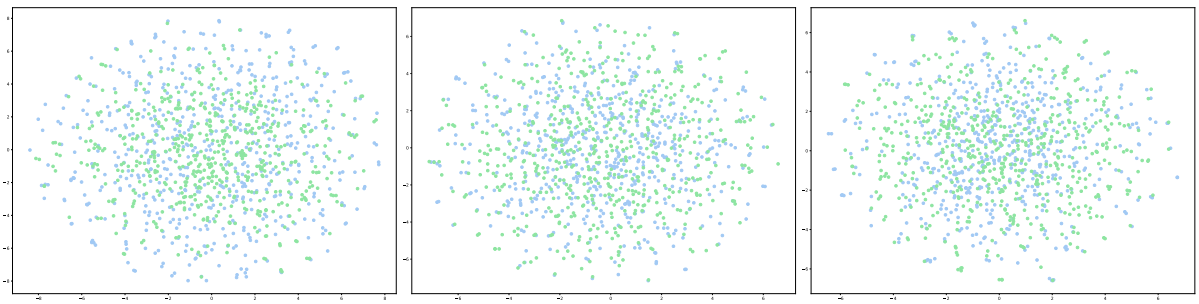
(a) Ugarit (b) LaBSE (c) SPhilBERTa

Untransformed Results



(d) Ugarit (e) LaBSE (f) SPhilBERTa

Transformed through whitening



(g) Ugarit (h) LaBSE (i) SPhilBERTa

Transformed through CBIE

Figure 2: Graphical representation of embeddings of Ancient Greek and Latin sentences from three models (Ugarit, LaBSE, and SPhilBERTa) using t-SNE, before and after applying whitening and CBIE. Each color represents a specific language.

Model	Anisotropy	Outlier Dimensions	Average Contribution of Dimension to Total Anisotropy
Base Results			
LaBSE	0.20	Dim 316 Dim 242 Dim 404 Dim 263	0.012 0.004 0.004 0.004
SPhilBERTa	0.40	Dim 461 Dim 513	0.023 0.007
Ugarit	0.76	Dim 588	0.380
Whitening			
LaBSE	$2.58 \cdot 10^{-5}$	Dim 216 Dim 591	$8.27 \cdot 10^{-7}$ $3.64 \cdot 10^{-7}$
SPhilBERTa	$3.05 \cdot 10^{-5}$	Dim 355	$4.85 \cdot 10^{-7}$
Ugarit	$2.42 \cdot 10^{-5}$	Dim 459 Dim 267 Dim 184 Dim 239 Dim 72	$1.74 \cdot 10^{-5}$ $1.57 \cdot 10^{-6}$ $1.40 \cdot 10^{-6}$ $9.78 \cdot 10^{-7}$ $7.96 \cdot 10^{-7}$
CBIE			
LaBSE	$9.29 \cdot 10^{-6}$	Dim 136	$1.33 \cdot 10^{-7}$
SPhilBERTa	$2.93 \cdot 10^{-5}$	Dim 535 Dim 555	$4.13 \cdot 10^{-7}$ $3.99 \cdot 10^{-7}$
Ugarit	$1.10 \cdot 10^{-5}$	Dim 588 Dim 459	$2.20 \cdot 10^{-6}$ $6.77 \cdot 10^{-7}$

Table 8: Total anisotropy of selected models combined with outlier dimensions and their respective contribution to the total anisotropy, out-of-the-box and after applying whitening and CBIE.

Model	Precision	Recall	F-score	Change to base benchmark
Base Result				
LaBSE-Glot500	80.8	74.6	77.6	+71.6
LaBSE-SPhilBERTa	**91.2**	**83.3**	**87.1**	+44.7
SPhiTune-Glot500	49.2	38.6	43.3	+37.3
SPhiTune-SPhilBERTa	68.6	61.5	64.9	+22.5
With Whitening on the Teacher Embeddings				
LaBSE-Glot500	**91.5**	73.6	81.6	+75.6
LaBSE-SPhilBERTa	89.6	**83.7**	**86.6**	+44.2
SPhiTune-Glot500	71.5	65.3	68.2	+62.2
SPhiTune-SPhilBERTa	85.9	78.6	82.1	+39.7
Whitening applied afterwards (to "Base Result")				
LaBSE-Glot500	95.1	93.9	94.5	+88.5
LaBSE-SPhilBERTa	**98.0**	**97.1**	**97.6**	+55.2
SPhiTune-Glot500	84.3	77.2	80.6	+74.6
SPhiTune-SPhilBERTa	91.3	74.6	82.1	+39.7
Whitening applied afterwards (to "With Whitening")				
LaBSE-Glot500	94.7	92.5	93.6	+87.6
LaBSE-SPhilBERTa	**98.5**	**96.5**	**97.5**	+55.1
SPhiTune-Glot500	89.5	76.5	82.5	+76.5
SPhiTune-SPhilBERTa	93.9	93.4	93.6	+51.2

Table 9: Results of knowledge distillation with **English** as anchor language, with comparison to base benchmark score of student model.

Model	Precision	Recall	F-score	Change to base benchmark
Base Result				
LaBSE-Glot500	0.9	0.9	0.9	-5.1
LaBSE-SPhilBERTa	42.6	36.8	39.5	-2.9
SPhiTune-Glot500	1.4	1.1	1.2	-4.8
SPhiTune-SPhilBERTa	**72.5**	**63.1**	**67.5**	+25.1
With Whitening (Teacher Embeddings)				
LaBSE-Glot500	1.8	2.1	2.0	-4.0
LaBSE-SPhilBERTa	48.0	37.7	42.3	-0.1
SPhiTune-Glot500	3.3	3.9	3.6	-2.4
SPhiTune-SPhilBERTa	**74.0**	**69.3**	**71.6**	+29.2
Whitening applied afterwards (to "Base Result")				
LaBSE-Glot500	7.2	3.8	5.0	-1.0
LaBSE-SPhilBERTa	**92.0**	81.5	86.4	+44
SPhiTune-Glot500	28.8	19.9	23.5	+17.5
SPhiTune-SPhilBERTa	91.8	**89.0**	**90.4**	+48.0
Whitening applied afterwards (to "With Whitening")				
LaBSE-Glot500	18.3	18.2	18.2	+12.2
LaBSE-SPhilBERTa	85.6	75.9	80.5	+38.1
SPhiTune-Glot500	26.2	23.4	24.7	+18.7
SPhiTune-SPhilBERTa	**90.7**	**90.5**	**90.6**	+48.2

Table 10: Results of knowledge distillation with **Latin** as anchor language, with comparison to base benchmark score of student model.

Modeling the *Dalet* Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis

Rachel Tal,[2] Cheyn Shmuel Shmidman,[2] Avi Shmidman[1, 2]
[1]Bar-Ilan University, Ramat Gan, Israel
[2]DICTA, Jerusalem, Israel
rachel.tal@mail.huji.ac.il
cheynshmuel@gmail.com avi.shmidman@biu.ac.il

Abstract

The Aramaic proclitic *dalet*, widely used in historical Hebrew texts, serves two distinct grammatical functions: as a subordinating conjunction and as a possessive preposition. Because these functions are orthographically identical and no annotated resources exist for this task, large-scale computational analysis of their usage has previously been infeasible. In this paper we introduce a new BERT model for historical Hebrew in which all prefixes are segmented and encoded as independent tokens. This representation allows the model to evaluate proclitics directly and provides a probe-based unsupervised method for determining the grammatical role of the *dalet* clitic using masked language modeling predictions. We evaluate the approach on a manually annotated dataset drawn from historical Hebrew literature spanning multiple regions and historical periods, achieving over an average F1 score of over 0.89. Applying the method to a corpus of more than 300 million words of historical Hebrew texts, we conduct large-scale stylistic analyses of the choice between the Aramaic *dalet* and available Hebrew alternatives. The results reveal geographic and diachronic trends and identify distinct stylistic clusters within the corpus. The prefix-segmented model and annotated dataset are released for unrestricted use.

1 Introduction

The Aramaic clitic *dalet* - commonly integrated within historical Hebrew texts - admits to two separate grammatical functions: it may function as a subordinator, or as a possessive preposition.

The double usage of this clitic is attested in original Aramaic writing, such as the Babylonian Talmud, as in (1), as well as within post-Talmudic historical Hebrew works, as in (2).

We begin with an example from the Babylonian Talmud, containing three consecutive instances of the *dalet* clitic (bolded):

(1) wkêwān **d**dînêh **d**nokrî
and.since that.the.law.of.him of.gentile
dāzêl bātar ārbā...
that.follows according.to the.guarantor...
'And since the law of the gentile is that he follows the guarantor...'

The first and last instances function as a subordinator, while the one in the middle functions as a preposition.

In post-Talmudic historical Hebrew, this clitic continues to appear with great frequency, in both functions. Consider for example the following passage from Chiddushei Ha-Ritva, the commentary on the Talmud by Rabbi Yom Tov ben Avraham of Seville (Ha-Ritva, around c. 1250-1330 , Spain). This passage contains an intensive jumble of six consecutive occurrences of *dalet* clitic:

(2) whû haddîn dmāṣê limpārēk
and.it.is the.same that.he.can refute
mēemṣāît šebbkat rišônâ **d**bābā
from.the.middle of.group first of.latter
dsêpā **d**iqṭannôt **d**tištrê
of.clause of.small.ones that.will.be.permitted
dnêmā **d**lā miqqaryā qṭannâ
that.it.is.said that.not it.is.called a.small.one
'And, in fact, he could have raised a difficulty from the "middle one of the first group" that is [included] in the latter clause of the small ones, that it should be permitted. For it is stated that it is not called a small one.'

The first three instances function as prepositions, while the subsequent three function as subordinators.

Crucially, the writers of historical Hebrew could alternatively use Hebrew markers to express these same grammatical functions. Regarding subordination, the writers could avail themselves of the Hebrew subordinator *še-*, rather than the Aramaic *dalet*. And regarding the use of the *dalet* as a

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 121–131
July 4, 2026 ©2026 Association for Computational Linguistics

preposition, where it marks possession, the writers could alternatively use the Hebrew prepositional possessive marker *šel*, or they could express the possessive connection via a Hebrew construct form. The key point is that the presence of the clitic *dalet* within historical Hebrew texts is a *stylistic choice*. In each case, writers may choose either the clitic *dalet* or pure Hebrew equivalents.

Almost all historical Hebrew texts mix and match these options, to greater or lesser degrees. For instance, here is a sentence from the book Shevet Halevi of Rav Shmuel Wosner (c. 1913-2015, Israel). This single sentence contains all of the aforementioned grammatical markers (bolded, in order: the Hebrew preposition *šel*, the *dalet* clitic as subordinator, a Hebrew construct form indicating a possessive, the Hebrew subordinator *šin*, another construct form indicating a possessive, and finally two instances of the *dalet* clitic as possessive preposition).

(3) būbdā byôm ṭôb **šel** šābûôt **d**'aḥar
in.the.case in.Yom Tov of Shavuos that.after
gmar tpillâ nizkar
conclusion prayer remembered
šā"ṣ **še**'āmar raq ḥăṣî hallēl
shaliach.tzibbur that.said only half Hallel
'im ṣārîk laḥăzōr, w'im ḥôzēr
whether he.should to.return and.if he.returns
'im ṣārîk lbārēk, ûb**dînê**
whether he.should to.bless and.in.laws
hapsāqâ **d**hallēl, wkammâ
interruption of.Hallel and.several
ysôdôt btaqqānâ **d**ehallēl
fundamentals in.institution of.Hallel
'In the case of the Yom Tov of Shavuos after the conclusion of the prayer, the cantor remembered that he had only said half of Prayer of Hallel, whether he should return, and if he returns whether he should bless, and the laws of interruption of Prayer of Hallel, and several fundamentals in the institution of Hallel.'

In order to computationally investigate stylistic trends across historical Hebrew regarding the choice of the Aramaic clitic *dalet* versus the available Hebrew alternatives, we must first disambiguate the function of any given instance of the clitic *dalet*: is it filling the role of a subordinator, or is it filling the role of the possessive preposition? This is a challenging determination for computational methods, because the two functions of the clitic *dalet* are orthographically identical. Until now, no computational method has existed to make this determination, and thus, any large-scale computational investigation of these stylistic trends has been impossible.[1]

In this paper we present a novel solution to this challenge. The key contributions of this paper are:

- We introduce a new BERT model for historical Hebrew in which all prefixes are segmented and encoded as independent tokens, enabling contextual prediction at the level of the proclitic.
- We present an unsupervised method for determining the grammatical role of the *dalet* proclitic - whether subordinator or preposition - by probing the masked language modeling head of the prefix-segmented model.
- We construct and release the first annotated dataset labeling the grammatical function of the *dalet* clitic in historical Hebrew texts.
- We apply this method to a large corpus of historical Hebrew (over 300 million words) in order to analyze stylistic variation between the Aramaic *dalet* and its Hebrew alternatives.
- Using this large-scale analysis, we identify geographic, diachronic, and genre-based patterns in the use of the *dalet* clitic, as well as distinctive stylistic phenomena within the historical Hebrew corpus.

2 The Challenge

We seek to train a BERT classifier to determine whether a given instance of the *dalet* clitic functions as a subordinator or as a preposition.

Two BERT models currently exist for Historical Hebrew (Shmidman et al. (2022), Shmidman et al. (2024a)); however, both of them operate on a per-word basis, and do not provide any mechanism for independent encoding or prediction of the grammatical prefixes. Although it may be possible to

[1] To be sure, in certain situations, morpho-syntactic features can facilitate disambiguation. For example, interrogatives which contain "*kol* + *d-*" such as "*kol 'êmat d-*" (i.e. whenever, "*d-*" as a subordinator) or "*kol hêkā' d-*" (i.e. wherever, "*d-*" as a subordinator), see: Bar-Asher Siegal 2016. Additional guidelines regarding the determination of the *dalet* are discussed by Breuer 2007 and Asis 2010. However, rule-based grammatical determinations such as these only cover a small percentage of the actual cases of the *dalet* clitic, and thus they would be no more than a partial solution.

train a supervised classifier based on these models in order to determine the grammatical role of a given prefix, this would require a substantially-sized labeled training corpus of annotated prefixes, which does not exist for Historical Hebrew.

3 Our Solution: Prefix-Segmented BERT

In order to achieve unsupervised classification of the grammatical function of the *dalet* clitic, we propose a new pretrained BERT model, in which all proclitics are treated as independent tokens, separate from the words to which they are attached. This provides the BERT model with a new expressive capacity that does not exist in the current BERT models for historical Hebrew; specifically, it allows BERT to make contextual predictions on the level of the proclitic. As we will show in the upcoming sections, this approach provides an effective solution for unsupervised classification of the *dalet* clitic.

4 Training Details

4.1 Segmentation Classifier

In order to pretrain a BERT model in which prefixes are treated as independent tokens, we must first preprocess our full BERT pretraining corpus, adding a separator token after each prefix. Thus, we start by training a dedicated classifier to perform prefix segmentation on Rabbinic Hebrew text. We follow the same recipe introduced by Shmidman et al. (2024b) for this segmentation classifier. We release this segmentation classifier to the community: [2]

4.2 Corpus

For pretraining the BERT model, we use a corpus of 810 million words, collected from various sources and genres of Rabbinic Hebrew. We preprocess the corpus with the prefix segmentation classifier described in the previous section. All cases of prefixes are encoded as independent words, followed by prefix indicator mark.

4.3 Pretraining the new BERT model

With the newly preprocessed corpus, we pretrained our new BERT model. We started by training a new tokenizer based on the new prefix-segmented corpus, to ensure that all of the relevant prefix tokens are included in the token vocabulary. We then ran the full BERT pretrain on this preprocessed corpus. The technical details of the training are presented in Appendix B. We release this new BERT model for unrestricted use.[3]

5 Classification of the *dalet* clitic

5.1 Method

Given our new pretrained prefix-segmented BERT model, we can obtain an unsupervised prediction for any given *dalet* clitic, as follows: For each case of a word which opens with a *dalet* clitic, we query the model's masked language modeling (MLM) head at that position and obtain the top k predicted tokens. This converts the classification problem into a masked language modeling prediction task, following the paradigm of pattern-based classification with language models (Schick and Schütze, 2021). From this ranked list we examine only two candidates corresponding to the Hebrew orthographic realizations of the two *dalet* grammatical roles: for *dalet.sub* (Subordinator), we look for the token ש and for *dalet.prep* (Preposition) we look for the token של. If neither candidate appears among the top-k predictions, the instance is labeled *unknown*. If only one candidate appears, that form is selected. If both candidates appear, we choose one ranked higher.

Formally, let x be an input sentence and i the target position. The masked language model returns an ordered list of the top-k predictions

$$\mathcal{P}_k(x, i) = (p_1, p_2, \ldots, p_k),$$

where p_r is the token ranked at position r.

Let c_{shin} denote the token corresponding to ש and c_{shel} denote the token corresponding to של. Define their ranks in the prediction list as

$$r_1 = \begin{cases} r & \text{if } p_r = c_{\text{shin}} \text{ for some } 1 \le r \le k, \\ \infty & \text{otherwise,} \end{cases}$$

$$r_2 = \begin{cases} r & \text{if } p_r = c_{\text{shel}} \text{ for some } 1 \le r \le k, \\ \infty & \text{otherwise.} \end{cases}$$

The predicted form is then determined:

$$\hat{y}(x, i) = \begin{cases} \text{unknown} & \text{if } r_1 = \infty \text{ and } r_2 = \infty, \\ c_{\text{shin}} & \text{if } r_1 < r_2, \\ c_{\text{shel}} & \text{if } r_2 < r_1. \end{cases}$$

[2] https://huggingface.co/dicta-il/BEREL-seg

[3] https://huggingface.co/dicta-il/pre-BEREL

5.2 Test Corpus

To assess the reliability of our method for determining the grammatical role of the *dalet* clitic, we constructed an evaluation dataset consisting of eight rabbinic works spanning different historical periods and geographical regions. Each instance of the *dalet* clitic in the corpus is manually annotated as either *dal.sub* or *dal.prep*. Table 2 summarizes the works included in our dataset and the number of *dalet* instances of each class in each work.[4] This is the first dataset of its kind, and we release it to the community.[5]

5.3 Evaluation

We evaluate our method for a range of values of k, in both masked and unmasked scenarios. In the *masked* scenario, the *dalet* clitic is replaced with a `[MASK]` token and the model must predict the missing token. In the *unmasked* scenario, the clitic itself remains visible in the input and the model ranks candidate tokens for that position based on contextual likelihood. In both scenarios we retrieve the top-k predictions from the MLM; these predictions are then evaluated using the ranking-based decision rule defined in Section 5.1.

Table 1 summarizes performance as a function of k. When the *dalet* clitic is masked, the relevant candidates often fall outside the top-k predictions, producing a larger number of unknown classifications. Increasing k reduces these cases and therefore substantially improves overall accuracy.

By contrast, in the unmasked scenario, when the clitic itself is visible, the model is given a strong morphological cue, which directly guides it toward these candidates, making them appear near the top of the prediction list and greatly reducing the number of unknown cases. However, in the unmasked scenario, the recall for the minority case (dal.prep) fails to climb above 71.8, indicating that the model is unfairly biased toward the majority case when the clitic is revealed. The masked scenario forces the model to make a decision without this bias, resulting in higher scores overall.

In the data experiments below, we adopt the highest performing approach from this evaluation, that is, we use the Masked scenario, with k=500.

6 The Use of the Clitic *dalet* in Historical Hebrew: New Data and a New Analysis

In the previous sections we demonstrated our new computational method to differentiate between the two roles of the proclitic *dalet*: whether as a subordinating conjunction or as a preposition. We now apply this method en masse to a large corpus of historical Hebrew, containing over 300 million words, from the 10th century to the 20th century. Table 4 in Appendix C shows the distribution of the corpus across chronological periods and geographical regions, and Table 5 in Appendix C shows the distribution of the texts across literary genres.

As noted above, the use of the clitic *dalet* in historical Hebrew writing is a stylistic choice. Although it is perfectly normative to interweave Aramaic prefixes within historical Hebrew texts, Hebrew alternatives do exist for both grammatical functions of the *dalet*. Essentially, we have before us two sets of stylistic minimal pairs; every writer of historical Hebrew could optionally decide whether to use the Aramaic *dalet* or the available Hebrew equivalents for every instance of subordination or possessive attachment. Our new computational method allows us to investigate how these stylistic choices come to the fore in historical Hebrew.

6.1 Mapping the use of the proclitic *dalet* across time and place

Our first experiment attempts to identify shifts and trends regarding the stylistic usage of the proclitic *dalet* within historical Hebrew literature.

6.1.1 Experimental Setup

We begin by running each text through the prefix segmentation classifier (Section 4.1) in order to identify cases where the initial *dalet* of a word serves as a proclitic rather than as part of the primary lexeme. For each such case, we run the method described above (Section 5.1) in order to determine whether the *dalet* functions as a subordinator or as a preposition. We thus compute the total number of instances of *dalet*-as-subordinator and *dalet*-as-preposition within each composition.

[4] To be sure, these rabbinic works were also included in the pretraining corpus for the new prefix-segmented BERT model. Nevertheless, it should be emphasized that this does not mean that the results reflect memorization rather than genuine generalization. The unannotated digital Hebrew texts which are input to the BERT model contain no indication whatsoever as to whether a given *dalet* clitic functions as subordinator or as a preposition. Thus, memorization of the text provides no a priori advantage when it comes to classifying the function of the clitic.

[5] https://www.dropbox.com/scl/fi/446x13r4b2hn1h16w1ep9/DaletCorpus?rlkey=1rgpx2y5e1xh2gohx58um1v1z&st=i0oloyte&dl=0

k	Masked							Unmasked						
	dal.sub			dal.prep			Unk	dal.sub			dal.prep			Unk
	Prec	Recall	F1	Prec	Recall	F1		Prec	Recall	F1	Prec	Recall	F1	
10	97.9	71.0	.823	90.3	46.2	.611	31.2%	97.8	70.2	.818	93.8	47.5	.631	31.8%
25	96.1	84.2	.898	89.9	63.7	.745	15.3%	95.6	92.3	.939	93.4	63.7	.757	8.4%
50	96.2	90.9	.935	89.9	72.2	.801	8.0%	94.6	97.8	.962	91.9	70.8	.800	1.2%
100	95.7	95.5	.956	89.4	75.3	.818	3.0%	94.3	98.3	.963	91.4	71.3	.801	0.3%
150	95.4	96.8	.961	89.5	76.2	.823	1.4%	94.3	98.5	.963	91.4	71.8	.804	0.1%
200	95.3	97.1	.962	89.5	76.7	.826	0.9%	94.3	98.6	.964	91.4	71.8	.804	0.0%
300	95.2	97.4	.963	89.5	76.7	.826	0.6%	94.3	98.6	.964	91.4	71.8	.804	0.0%
400	95.2	97.6	.964	89.5	76.7	.826	0.4%	94.3	98.6	.964	91.4	71.8	.804	0.0%
500	95.2	98.0	.966	89.5	76.7	.826	0.1%	94.3	98.6	.964	91.4	71.8	.804	0.0%

Table 1: Per-class performance as a function of the number of retrieved MLM candidates (k), contrasting Masked and Unmasked scenarios. Unk is the percentage of instances where neither candidate appeared in the top-k predictions.

Author	Birth Year	Birth Place	Book	dal.sub	dal.prep
Rabbi Yoseph ibn Migash	1077	Seville, Spain	Responsa of Ri Migash	141	23
Rabbi Israel Isserlein	1390	Wiener Neustadt, Lower Austria	Terumat HaDeshen	128	29
Rabbi Joseph Karo	1488	Toledo, Spain	Shulhan Arukh	141	30
Rabbi Shlomo Luria	1510	Poznan, Poland	Hokhmat Shlomo	125	27
Rabbi Yechezkel Landau	1713	Opatów, Poland	Shut Noda B'Yehuda	141	18
Rabbi Avraham Danzig	1748	Danzig, Poland	Hayei Adam	136	14
Rabbi Yosef Hayyim	1835	Baghdad, Iraq	Ben Ish Hai – Derashot	102	65
Rabbi Ben-Zion Meir Hai Uziel	1880	Jerusalem, Israel	Mishpetei Uziel	145	17

Table 2: The historical Hebrew texts included in the annotated dataset. For each book we report the raw counts of the two classes of the *dalet* clitic: *dal.sub* (used as subordinator) and *dal.prep* (used as preposition).

Furthermore, we compute the total number of times that each composition contains the Hebrew alternatives for these roles. Regarding subordination, we tally the number of times that the Hebrew šin character appears as a proclitic; and regarding the possessive connection, we tally the number of times that the text contains the Hebrew preposition šel, and the number of times that the text contains Hebrew construct forms.[6]

Finally, for each text, we compute two statistics, for each of the two grammatical roles, in order to represent the proportion of cases in which the writer chose the Aramaic *dalet* rather than the Hebrew stylistic alternative. Regarding subordination, we take the count of *dalet*-as-subordinator, and we divide it by the total cases of subordination in the text, whether Hebrew or Aramaic. A value of 1 indicates that the author always chose the Aramaic stylistic alternative for subordination; a value of 0 indicates that the author always chose the Hebrew alternative; and a value of 0.5 indicates that the author chooses evenly between the two alternatives. Regarding the possessive connection we do the same, dividing the number of instances of *dalet*-as-preposition by the total possessive connections, whether Hebrew or Aramaic.

We then visualize these author-level statistics in four interpolated geographic maps, in order to identify diachronic and spatial changes in this stylistic preference. We generate separate maps for the two grammatical roles and, within each role, separate maps for medieval authors (11th–15th centuries) and modern authors (16th–20th centuries). Each map plots author-level values according to the longitude and latitude of the author's birthplace, and the background surface is obtained by inverse-distance-weighted (IDW) interpolation over geographic space.[7]

[6] In order to identify words in the construct state, we use the Hebrew morphological analyzer provided at https://morphology.dicta.org.il.

[7] For technical details regarding the creation of the maps see Appendix A.

6.1.2 Results

We display the results of this experiment in Figures 1 and 2.[8]. In each map, lighter regions indicate higher interpolated values for the Aramaic variant relative to the Hebrew alternative.

We pay particular attention to the two major literary spheres of European Jewry: Ashkenaz and Spain.[9] In Figures 1 and 2, the Ashkenaz region appears primarily in the Franco-German area of Central Europe, approximately east of 4° longitude and north of 47° latitude, whereas the Spanish region lies in the Iberian Peninsula, approximately west of 0° longitude and south of 44° latitude.[10]

Geographically, the clearest observed contrast lies between Ashkenaz and Spain in the medieval period. The division between the two geographical regions here is relatively sharp. In Spain there is a far higher tendency toward the Hebrew stylistic alternative for both subordination and possessive connections, while the Ashkenaz region uses the *dalet* proclitic with far more frequency.

From the diachronic perspective, we cannot meaningfully compare Ashkenaz to Spain, because the modern Spanish corpus contains only a limited number of examples, due to the disruption and eventual dissolution of Jewish communal life in Christian Spain following the expulsions and forced conversions of the late fifteenth century.[11] Instead, we point to a diachronic change within the Ashkenazic corpus itself, that is, medieval writers in Ashkenaz versus modern writers in Ashkenaz. The observed corpus-level patterns suggest that Medieval Ashkenaz is characterized by a consistently strong tendency toward the Aramaic alternatives for both subordination and for the genitive connection, while modern Ashkenaz demonstrates a much more heterogenous distribution: within the same region, we find many writers tending toward the Hebrew alternatives, alongside those who tend toward the Aramaic ones. It is possible that the reason underlying this shift is an increased communal migration in the modern period, which would have led to increased contact between linguistic traditions and the blending of mother tongues and writing conventions.[12]

6.2 Author Variability

Our second experiment investigates the question: how strong are individual author tendencies regarding the stylistic usage of the dalet clitic? How consistent are authors in their usage of the dalet?

6.2.1 Experimental Setup

Of course, in order to assess this question, we must have a set of multiple works authored by the same writer. Therefore, we restricted this part of the analysis to authors in our corpus with at least three distinct books. Applying this threshold yields a set of 62 authors, each represented by between three and twenty-six texts in the corpus. Table 3 presents a representative subset of authors who satisfied this criterion. For each of the books of each of these authors, we calculate the proportions of the use of the *dalet* clitic for each of its two grammatical functions, as compared with the Hebrew alternatives, using the procedure described above (6.1.1). Further, we calculate the standard deviation for each of the two *dalet* proportions across all of the books of each author, in order to examine the consistency of the author's usage of the *dalet* clitic.

We plot these statistics in Figure 3, using the standard deviation of the dalet-as-subordinator proportion as the x-coordinate, and the standard deviation of the dalet-as-preposition proportion as the y-coordinate.

[8] The color scale is identical in both maps. Although the values among modern writers are slightly higher — reflecting a greater reliance on the Aramaic proclitic *dalet* both for subordination and also for expressing a possessive connection — a unified scale was retained in order to facilitate direct visual comparison between the two distributions. In addition, while inverse-distance interpolation is visually compelling, it may create misleadingly smooth patterns that risk overstating spatial coherence. Accordingly, the maps should be interpreted with caution: they serve primarily to visualize broad tendencies for the reader, whereas the substantive analysis and conclusions derive from the statistical evidence.

[9] In its most basic definition, "Ashkenaz" refers to the Jewish communities of medieval Franco-Germany, centered in the Rhineland and northern France. For a historical definition of Ashkenaz and its cultural-linguistic scope, see Marcus 2010. This regional distinction is not merely geographical but also reflects differences in Jewish legal tradition, communal organization, and literary style.

[10] Regarding other regions—such as modern Hebrew writers in North Africa—the available data in the present corpus is insufficient to support statistically robust conclusions.

[11] On the historical development and decline of Jewish communities in Christian Spain, see Yitzhak Baer, Baer 1958

[12] Such migrations of Jewish communities are attested at the end of the fifteenth century following the expulsion of the Jews from Spain. Many Jews from Spain settled in Italy (Beinart 1988); others migrated to the Balkan regions of the Ottoman Empire (Benbassa and Rodrigue 2001). Additionally, beginning in the late Middle Ages and continuing through the fifteenth century, repeated expulsions from German territories and other regions of Western Europe led to large-scale migration of Jewish communities eastward. These movements ultimately contributed to the emergence of major Jewish centers in Eastern Europe, particularly in

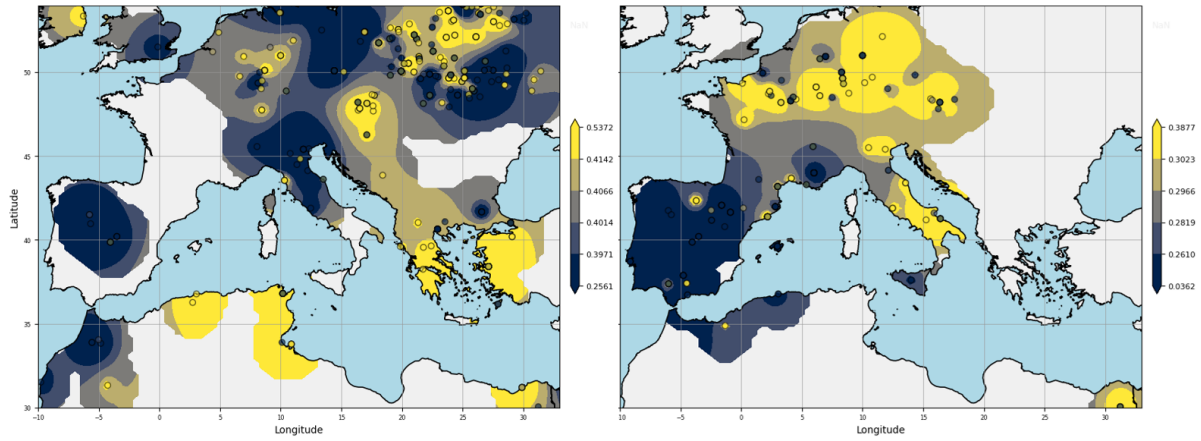


Figure 1: Geographic distribution of the stylistic choice of grammatical subordinator among medieval writers (right) and modern writers (left). Higher values (lighter colors) indicate a tendency towards the Aramaic clitic *dalet*, whereas lower values (darker colors) indicate preference for Hebrew *šin*. Authors are plotted according to their birthplace. The background surface shows inverse-distance-weighted interpolation across geographic space.

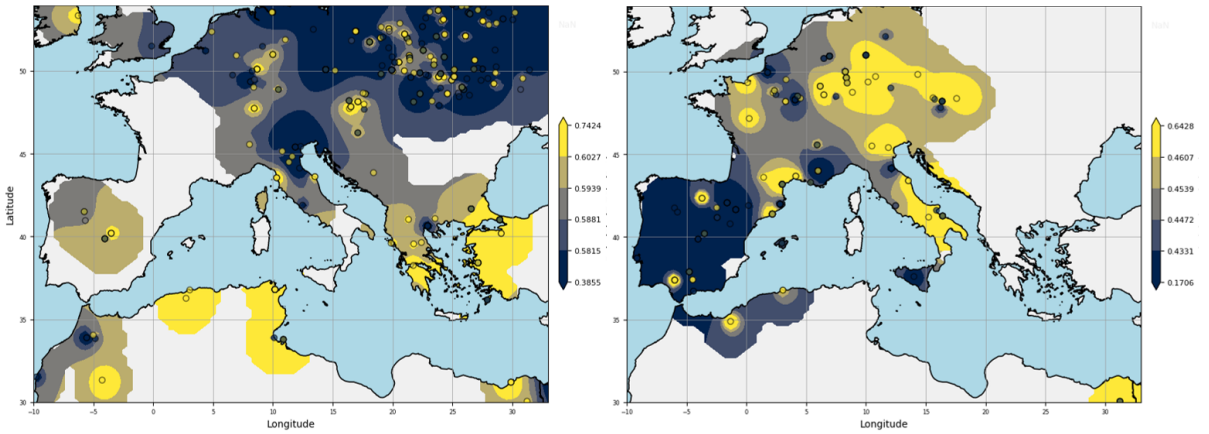


Figure 2: Geographic distribution of the stylistic choice of possessive marker among medieval writers (right) and modern writers (left). Higher values (lighter colors) indicate a tendency towards the Aramaic clitic *dalet*, whereas lower values (darker colors) indicate a preference towards Hebrew alternates. Authors are plotted according to their birthplace, and the background surface shows inverse-distance-weighted interpolation across geographic space.

6.2.2 Results

The plot reveals a clear positive correlation between the two measures (Pearson r = 0.79, R^2 = 0.62), indicating that authors who display greater variability in the use of one construction tend also to exhibit greater variability in the other. The regression line added to the plot highlights this general tendency and suggests that the variability of the two constructions is not independent but part of a broader stylistic continuum. At the same time, it is apparent that there is much more stylistic variation regarding the use of the *dalet.sub*, as compared with *dalet.prep*; this may be because *dalet.prep* is often linked to fixed collocations from Talmudic literature, and thus its free variation may be more limited.

Several authors cluster in the bottom left area of the graph, indicating particularly low variability for both types of dalet. These authors include Abravanel, Ramhal, Moshe ibn Habib, Rabbi Eliyahu Israel, and Ramhal (Rabbi Moshe Hayyim Luzzatto). Indeed, for each of these authors, their multiple compositions belong largely to the same genre. For instance, the four books by Ramhal in our corpus are all books of Jewish Thought. Such genre homogeneity appears to produce relatively stable stylistic behavior.

By contrast, authors located at the upper right of the graph display markedly higher variability

Poland–Lithuania (see Kaplan 2021).

Author	Birth Year	Birth Place	Books	Mean		Median		Std. Dev.	
				dal.sub	dal.prep	dal.sub	dal.prep	dal.sub	dal.prep
Rabbi Shlomo Yitzḥaki (Rashi)	1040	Troyes, France	9	0.265	0.114	0.268	0.107	0.148	0.057
Rabbi Abraham ben David of Posquières (Raavad)	1120	Narbonne, France	6	0.337	0.184	0.304	0.172	0.134	0.039
Rabbi Moses ben Nahman (Naḥmanides / Ramban)	1194	Girona, Spain	12	0.267	0.128	0.272	0.140	0.238	0.099
Rabbi Meir of Rothenburg (Maharam)	1215	Worms, Germany	5	0.412	0.206	0.399	0.219	0.169	0.056
Rabbi Nissim ben Reuven of Girona (Ran)	1290	Barcelona, Spain	6	0.413	0.199	0.518	0.230	0.223	0.079
Rabbi Yoseph Karo	1488	Toledo, Spain	4	0.412	0.234	0.366	0.205	0.288	0.155
Rabbi Judah Loew ben Bezalel (Maharal of Prague)	1520	Poznań, Poland	13	0.210	0.091	0.068	0.077	0.236	0.063
Rabbi Hayyim ibn Attar	1696	Salé, Morocco	6	0.413	0.170	0.412	0.168	0.314	0.130
Rabbi Jonathan Eybeschütz	1696	Pinczów, Poland	6	0.599	0.188	0.698	0.189	0.266	0.061
Rabbi Yosef ben Meir Teomim (Pri Megadim)	1727	Szczuczyn, Poland	6	0.604	0.215	0.597	0.221	0.094	0.042
Rabbi Pinḥas Horowitz	1730	Chortkiv, Ukraine	5	0.570	0.185	0.607	0.214	0.147	0.071
Rabbi Jacob Lorberbaum	1770	Zbarazh, Ukraine	9	0.376	0.144	0.465	0.142	0.183	0.059
Rabbi Naftali Zvi Yehuda Berlin (Netziv)	1817	Mir, Belarus	5	0.378	0.131	0.422	0.164	0.218	0.079
Rabbi Tzadok HaKohen Rabinowitz of Lublin	1823	Krizburg, Lithuania	20	0.272	0.166	0.252	0.155	0.149	0.043
Rabbi Yisrael Meir HaKohen (Hafetz Hayyim)	1839	Radin, Poland	26	0.129	0.061	0.103	0.051	0.100	0.044

Table 3: Evaluation of author consistency. We display a sample of the authors for whom our corpus contains three or more books each. We report the mean, median, and standard deviation of the proportions indicating how often the Aramaic *dalet* clitic is chosen rather than its Hebrew counterparts for the two grammatical functions under investigation: for subordination (*dal.sub*) and for possessive marking (*dal.prep*).

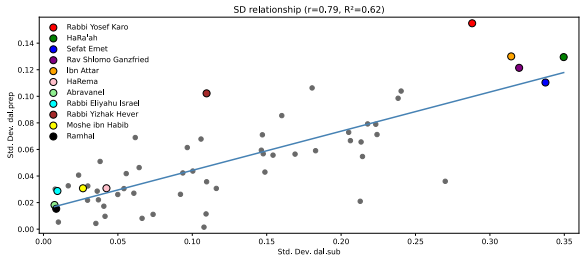


Figure 3: Author-level stylistic variability in the use of the Aramaic *dalet* clitic across multiple works. Each point represents an author with ≥ 3 books in the corpus; the x-axis gives the standard deviation of the *dalet*-as-subordinator proportion and the y-axis gives the standard deviation of the *dalet*-as-preposition proportion. The fitted regression line indicates a strong positive correlation between the two measures (Pearson $r = 0.79$).

in both constructions. Indeed, the works in the corpus written by these authors are spread across several different genres, which naturally increases stylistic variability. For instance, the works of Ibn Attar (cc. 1696-1743, Morocco), represented on the plot by the orange circle, belong to three different literary genres: Biblical Commentary, Legal Responsa, and Jewish Legal Codes and Customs.

6.3 Outliers and Distinctive Stylistic Clusters

The data that we generated regarding all instances of the *dalet* proclitic across the full corpus of hundreds of millions of words also allows us to identify striking outliers and anomalous stylistic clusters. We present two such cases here.

The first stylistic cluster that we examine is found in a group of medieval writers from Provence, including Jonathan ben David ha-Cohen from Lunel (cc. around 1135-1211) and Menachem HaMeiri (cc. 1249–1315). In their writings we find an overwhelming tendency toward the Hebrew options for both subordination and possessive marking, with exceptionally low frequencies of the Aramaic *dalet*. In fact, within their writings, we even find the Aramaic *dalet* missing from Talmudic collocations in which the Aramaic *dalet* is standard and expected. For instance, instead of "gəzērâ dərabbâ" ("a decree of Rabba") - a common expression in Jewish legal literature - HaMeiri writes "gəzērâ dərabbâ" (same meaning, but achieved via a Hebrew construct state). So too, instead of "rôtēaḥ dəṣālî" ("boiling of roasted meat"), he uses "rôtēaḥ šel ṣālî", replacing the Aramaic *dalet* with the Hebrew *šel*. Indeed, these Provençal writers are known to have been influenced by Maimonides (see Ben-Shalom 2008), who was particularly vocal in his insistence on pure Hebrew writing, to the exclusion of Aramaic.

The second anomalous stylistic cluster highlighted by our data involves a group of modern writers in Eastern Europe who exhibit a strong Hebrew tendency when it comes to the possessive marker, yet at the same time demonstrate a preference for the Aramaic *dalet* when it comes to subordination. Within this group are authors such as Rabbi Shimon Shkop, who in his commentary on the Babylonian Talmud employs *dalet* as a subordinator at a rate above the corpus average (52.98%), yet almost entirely avoids *dalet* as a

preposition (8.43%). Similarly, Rabbi El‘azar ha-Kohen of Pułtusk, in his composition Ḥiddushei Maharakh, uses *dalet* as a subordinator more than the corpus average (50.02%), while avoiding its use as a preposition (8.11%). The third author within this group is Samuel ben Uri Shraga Phoebus (52.79% subordinator, 11.08% preposition). This stylistic cluster is quite exceptional, because, as we have seen, when it comes to tendencies towards or away from the Aramaic *dalet*, the two grammatical roles tend to go hand in hand: writers who avoid one also avoid the other, and vice versa. Yet here we have a group of writers who eagerly adopt the *dalet* proclitic for only one of its grammatical roles, while preferring the Hebrew option for the other role.[13]

7 Future Work

The large-scale analyses presented above reveal geographic, diachronic, genre-based, and author-specific patterns in the use of the Aramaic *dalet* clitic relative to its Hebrew alternatives. At the same time, the present study should be understood as an initial corpus-wide exploration of these patterns rather than as a final causal model of the factors governing them. In particular, the geographic and diachronic visualizations are based on author-level aggregate statistics, and although they are useful for identifying broad trends, they do not fully disentangle the potentially overlapping effects of period, region, genre, author identity, and corpus size.

In future work, we plan to extend our analysis by modeling the data using mixed-effects regression. This will make it possible to estimate the contribution of period, genre, region, and text length, while separating author-specific stylistic tendencies from broader historical, geographic, and generic effects. We also plan to add robustness checks for the main geographic and diachronic claims, in order to assess how stable these patterns remain under alternative corpus weightings, regional groupings, and chronological divisions.

8 Conclusion

We have produced a new BERT model, pre-trained from scratch on prefix-segmented historical Hebrew, capable of performing unsupervised proclitic disambiguation in Hebrew historical texts. We demonstrate how this new model allows us to run big-data experiments across the full corpus of historical Hebrew, gaining new insights into the stylistics of historical Hebrew writing; such experiments were previously all but impossible. We are pleased to release the new BERT model to the community for unrestricted use, both for this specific disambiguation task, as well as for any other downstream task which would benefit from token prediction at the level of the segmented proclitic.

[13] No reverse behavior was identified; we did not find authors who make extensive use of the *dalet* as a preposition while strongly avoiding its usage as a subordinator.

9 Acknowledgments

This work has been funded by the Israel Science Foundation (grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829; principal investigators: Avi Shmidman, Bar-Ilan University; Daniel Stökl, EPHE-PSL; Nachum Dershowitz, Tel Aviv University; and Judith Olszowy-Schlanger, EPHE-PSL), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Moshe Asis. 2010. *A Treasury of Jerusalemite Expressions*. The Jewish Theological Seminary of America, New York and Jerusalem. Hebrew; published in 2010 (Hebrew year 5770).

Yitzhak Baer. 1958. *Toledot ha-Yehudim bi-Sefarad ha-Notzrit*. Am Oved, Tel Aviv. 2nd rev. and expanded ed.; Hebrew.

Elitzur A. Bar-Asher Siegal. 2016. *Introduction to the Grammar of Jewish Babylonian Aramaic*, 2 edition. Ugarit-Verlag, Münster. References: §§ 10.1, 4.3.

Haim Beinart. 1988. Jewish–converso relations between spain and italy. In Haim Beinart, editor, *Jews in Italy: Studies Dedicated to the Memory of U. Cassuto on the 100th Anniversary of His Birth*, pages 275–288. Magnes Press, Jerusalem.

Ram Ben-Shalom. 2008. *The Jews of Provence and Languedoc*. Magnes Press, Jerusalem.

Esther Benbassa and Aron Rodrigue. 2001. *The Jews of the Balkans: The Judeo-Spanish Community, 15th–20th Centuries*. The Zalman Shazar Center for Jewish History, Jerusalem. Hebrew translation. Originally published in French.

Yochanan Breuer. 2007. The babylonian aramaic in tractate karetot: According to ms oxford. *Aramaic Studies*, 5(1):1–45.

Yosef Kaplan. 2021. Jews on the move: Early modern jewish migration. The Posen Library of Jewish Culture and Civilization.

Ivan G. Marcus. 2010. Ashkenaz. YIVO Encyclopedia of Jews in Eastern Europe, accessed 2026-03-12.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *Preprint*, arXiv:2208.01875.

Avi Shmidman, Ometz Shmidman, Hillel Gershuni, and Moshe Koppel. 2024a. MsBERT: A new model for the reconstruction of lacunae in Hebrew manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 13–18, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *Preprint*, arXiv:2308.16687.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024b. MRL parsing without tears: The case of Hebrew. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4537–4550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Appendix: Geographic Interpolation Procedure

The geographic visualizations in Figures 1 and 2 were constructed from author-level observations consisting of birthplace coordinates (latitude and longitude) and a scalar feature value representing the relative frequency of the linguistic form under study. These observations were interpolated over a regular grid spanning longitude -10° to 33° and latitude 30° to 54°, sampled at a resolution of 100×100 cells. For each grid cell, great-circle distances to all observed author locations were computed using the haversine formula. Interpolation was then performed by inverse distance weighting (IDW), with each observation weighted by the inverse of its distance from the grid cell. Cells for which no observation occurred within 250 kilometers were omitted from the interpolated surface; when this condition was satisfied, all observations within 5000 kilometers were included in the calculation. Distances were lower-bounded by a small positive constant in order to prevent division by zero. The interpolated values were rendered as filled contours using quantile-based levels (0, 20, 40, 60, 80, and 100 percent) and the `cividis` color map, while the original author locations were superimposed as scatter points on the same scale.

B Appendix: Training Details

We trained the model on a DGX-A100 with 4xA100 40GB cards. The training was done with the fused lamb optimizer combined with AMP (Automatic Mixed Precision). A polynomial warmup learning rate scheduler was used to warm up for a portion of the training steps and then decay the learning rate over the total steps. We followed the training recipe introduced by (Shmidman et al., 2023) with regard to the training objectives & the construction of the training examples.

We used the HuggingFace framework wrapped with NVIDIA libraries[14] which are highly optimized for training compute-heavy machine learning models on NVIDIA hardware. We pretrained the model for a total of 36,200 iterations, each iteration consisting of 8,192 examples. The first 23,200 iterations were done with sequences of up to 256 tokens, followed by 13,000 iterations with sequences of up to 512 tokens. Total training time was 7.28 days.

C Appendix: Corpus Distribution

Period	Region	% of Corpus
Medieval Writers	Spain	7.68
	Ashkenaz	3.74
	Provence	2.80
	France	2.70
	Italy	1.14
Modern Writers	Ashkenaz	53.99
	Modern Hebrew	10.00
	Arabic language sphere	6.85
	Ottoman Empire	4.15
	Balkans	3.22
	Land of Israel	2.70
	Italy	0.93
	Yemen	0.10

Table 4: Distribution of the investigated corpus by historical period and geographical region.

[14] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT

Group	Literary Genre	% of Corpus
Exegetical and Commentarial Literature on Canonical Texts	Biblical Commentaries	6.19
	Commentaries on the Mishnah and Halakhic Midrash	1.25
	Commentaries and Halakhic Works on the Babylonian Talmud	17.09
	Commentaries and Halakhic Works on the Jerusalem Talmud	1.14
Halakhic Literature	Responsa Literature	21.47
	Commentaries and Works on the Shulḥan Arukh	7.80
	Commentaries on the Tur	1.25
	Commentaries on Maimonides	13.43
	Halakhic Codes and Customs	10.33
	Commandment Literature	0.91
Intellectual and Homiletic Literature	Jewish Thought and Ethical Literature	8.14
	Hasidic Literature	7.10
Systematic and Meta-Halakhic Literature	Legal Principles and Chronological Works	1.37
	Thematic and Conceptual Halakhic Works	2.53

Table 5: Distribution of the investigated corpus by literary genre (percentage of total words).

Beyond Genre Categories: How Narrative Pattern Coherence and Spanning Distance Shape Film Success

Zhichao Wang
Tohoku University
Faculty of Arts and Letters
wang.zhichao.p2@dc.tohoku.ac.jp

Zeyu Lyu
Tohoku University
Faculty of Arts and Letters
lyu.zeyu.e8@tohoku.ac.jp

Abstract

Prior research on cultural markets has relied on genre labels to distinguish products, overlooking the specific content features that differentiate films within the same genre. We address this gap using tropes as building blocks of narrative structure. From a dataset of 30k tropes across 18k films (TVTropes.org), we identify 29 narrative patterns via community detection and characterize each film by two measures: coherence (how concentrated its tropes are within a few patterns) and spanning distance (how far apart the patterns it combines are). Regression analyses show that coherence improves both audience evaluations and attention, while spanning distance increases evaluations but reduces attention. These findings extend category-spanning theory from genre labels to the internal narrative composition of films, demonstrating how stories are constructed and shape audience responses.

1 Introduction

Artistic creation is shaped by conventions and boundaries (Bourdieu, 1996; DiMaggio, 1987; Lena and Peterson, 2008). In film, as in other cultural industries, producers and audiences share expectations about how stories are told—what kinds of characters appear, how plots unfold, and what narrative devices are appropriate within a given tradition. Previous studies suggest that a work that crosses these boundaries risks confusing audiences who cannot easily place it(Zuckerman, 1999; Hsu, 2006), but may also be rewarded for its originality(Rao et al., 2005; Askin and Mauskapf, 2017). However, these studies have measured conventions and boundaries using genre labels, thereby often overlooking the specific plot and content features of films. For example, *Rush Hour* and *Kill Bill* are both labeled "Action," yet one is a buddy comedy built on slapstick and mismatched partners, while the other is a revenge tragedy driven by violence and betrayal. To capture such differences in narrative structure, we need to look beyond genre labels and into the narrative elements from which films are constructed.

Tropes are the recurring, recognizable narrative elements in cultural products—plot devices, character archetypes, situational setups—that constitute the narrative structure of stories(García-Ortega et al., 2021). A well-known example is *the Chosen One*: an ordinary protagonist marked for an extraordinary destiny, a device found in films as diverse as *The Matrix*, *Kung Fu Panda*, and *Moana*. Tropes do not appear independently; certain tropes tend to co-occur across films, forming what we call *narrative patterns*: groups of tropes that collectively reflect a recurring combination of narrative conventions. *TVTropes*[1], a community-maintained wiki launched in 2004, has documented approximately 33,000 tropes across 127,000 works, providing a comprehensive dataset on how tropes are used across cultural products. This makes it possible to identify the narrative patterns that tropes form and to examine how films' combinations of these patterns relate to market success.

In this paper, we identify narrative patterns from large-scale trope co-occurrence data and characterize each film along two dimensions. *Coherence* captures how concentrated a film's tropes are within a few patterns. *Spanning distance* captures, among films that combine multiple patterns, how far apart those patterns are from one another. We test three hypotheses regarding how these two dimensions relate to audience outcomes. First, we expect that coherence benefits films because tropes concentrated within familiar patterns make the narrative easier for audiences to interpret, leading to more favorable evaluations(Zajonc, 1968). Moreover, coherent films appeal more directly to audiences who already favor particular

[1] https://tvtropes.org/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 132–137
July 4, 2026 ©2026 Association for Computational Linguistics

patterns(Keuschnigg, 2015). Therefore: **H1:** Films with higher coherence receive more favorable audience evaluations and attract greater audience attention. Second, we expect that spanning distance affects evaluations and attention differently. Films that combine distant patterns create greater novelty and originality, which may lead to higher evaluations(Uzzi et al., 2013). However, when a film combines patterns that are far apart, it appeals to a narrower overlap of audience tastes, which reduces its potential overall audience(Hannan and Freeman, 1989). Therefore: **H2a:** Films with higher spanning distance are associated with higher audience evaluations. **H2b:** Films with higher spanning distance are associated with lower audience attention.

2 Related Work

The category-spanning literature has identified two mechanisms through which boundary crossing penalizes products: reduced niche fitness and audience confusion (Hsu et al., 2009; Negro and Leung, 2013). In film, Keuschnigg and Wimmer (2017) find that confusion plays a relatively larger role. These penalties are moderated by category contrast and the distance between categories spanned (Kovács and Hannan, 2015). Pontikes (2012) further shows that the same ambiguity penalized by consumers can be rewarded by investors, and distinguishes between variety and atypicality as independent dimensions of boundary spanning. These work suggests that what matters is not simply whether a product crosses boundaries, but which boundaries and how far.

A growing body of research has moved beyond category labels to characterize cultural products by their internal composition. Askin and Mauskapf (2017) use acoustic features to show that optimal differentiation predicts chart success; **?** demonstrate that the popularity of stylistic elements depends on their structural embeddedness among other elements rather than on the categories they belong to; Sgourev et al. (2023) show how visual aesthetic features like color serve as positioning devices relative to peers. These studies share a common insight: product-level features capture dimensions of novelty and conformity that coarse labels miss. Yet in film, the dominant approach continues to rely on genre labels as proxies for narrative structure.

We address this gap using tropes as building blocks of narrative structure. Tropes have been used in a range of computational and cultural studies: tracing how tropes propagate across works (Mellina and Svetlichnaya, 2011), mapping the trope co-occurrence networks (García-Ortega et al., 2021), supporting story generation (Álvarez and Font, 2022; Chou et al., 2023), and identifying social biases embedded in trope usage (Gala et al., 2020). We take a different direction: rather than predicting ratings from individual tropes or generating new narratives, we identify narrative patterns from trope co-occurrence and measure how films' combinations of these patterns—their coherence and spanning distance—relate to audience outcomes, connecting the computational study of tropes to the theoretical framework of category spanning.

3 Data

We collect a large-scale dataset containing approximately 30,924 tropes and 1.2 million trope–film links across 18,164 films from *TVTropes.org*. Then, we matched these films to the IMDb datasets to obtain film metadata (release year, genres) and audience outcomes: *IMDb rating* (audience evaluation) and *number of votes* (audience attention). The dataset covers films released from the 1920s to the 2020s. Table 1 reports descriptive statistics.

Table 1: Descriptive statistics

Variable	Mean	Median [Min; Max]
Tropes per film	54.66 (64.59)	32 [1; 483]
Films per trope	32.20 (93.33)	10 [1; 3788]
Genres per film	2.37 (1.02)	2 [1; 7]
IMDb rating	6.21 (1.18)	6.40 [1.20; 9.80]
IMDb votes	42967 (114517)	6423 [5; 3152337]

4 Methods

Our research proceeds in three steps. First, we identify *narrative patterns*—groups of tropes that tend to co-occur across films—by detecting trope communities in a co-trope network (Section 4.1). Second, we use each film's distribution of tropes across these patterns to construct two film-level variables: *coherence*, which captures how concentrated a film's tropes are within a few patterns, and *conditional spanning distance*, which captures how narratively distant the bridged patterns are (Section 4.2). Third, we estimate regression models relating these variables to audience outcomes (Section 4.3).

4.1 Narrative Pattern Identification

Certain tropes tend to co-occur in the same films, and these co-occurrence regularities reveal the underlying patterns that organize narrative conventions (Cawelti, 1976). To identify these patterns empirically, we construct a weighted co-trope network in which each node is a trope and each edge weight reflects how often two tropes appear together in the same film. We apply the Leiden algorithm (Traag et al., 2019) to partition this network into communities, yielding 29 groups. Each community is treated as a *narrative pattern*: a data-driven cluster of tropes that collectively reflect a recurring trope combination.

The resulting partition has a distinctive structure. One large community (Community 0, containing 10,573 tropes) appears across nearly all films in the dataset. Its central tropes—such as *Shout Out*, *Foreshadowing*, and *Chekhov's Gun*—are general-purpose narrative devices not specific to any genre or thematic cluster. This community functions as a baseline layer of narrative conventions shared across films. The remaining 28 communities are smaller and more specialized, each organized around a distinctive thematic or narrative domain: for example, one centers on supernatural and horror elements, another on science fiction and futuristic settings, and another on combat and weaponry. Community film coverage among these 28 ranges from 1 to 10,475 films (mean = 4,116; SD = 3,252). Table 3 in the Appendix reports characteristics of five example communities.

4.2 Measuring Coherence and Spanning Distance

Given the 29 narrative patterns identified above, we characterize each film along two dimensions. The first, *coherence*, captures how concentrated a film's tropes are within a few patterns—whether the film commits to a clear narrative identity or disperses its elements across many different conventions. The second, *spanning distance*, captures how narratively distant the patterns are that a film combines—whether it bridges patterns that frequently co-occur (close distance) or patterns that rarely appear together (far distance).

For film m, let n_{mi} denote the number of tropes assigned to community i, and let $N_m = \sum_i n_{mi}$ be the total trope count in film. The share of film m's tropes belonging to community i is $w_{mi} = n_{mi}/N_m$.

4.2.1 Coherence

Some films draw their tropes primarily from one or two narrative patterns; others spread them broadly. We expect this distinction to matter because a film whose tropes cluster within a small number of recognizable patterns presents a clearer narrative backbone, making it easier for audiences to interpret and evaluate. To capture this, we need to measure how concentrated a film's tropes are across patterns. We use the Herfindahl–Hirschman Index (HHI), a standard measure of concentration widely used in economics and organizational research to quantify how a distribution is spread across categories (Rhoades, 1993). In our context, HHI measures the degree to which a film's tropes are concentrated within a few narrative patterns:

$$\mathrm{HHI}_m = \sum_i w_{mi}^2.$$

HHI approaches 1 when a film draws almost all of its tropes from a single pattern, and decreases as tropes spread more evenly.

4.2.2 Spanning Distance

Beyond how concentrated a film's tropes are, we also want to know how far apart the patterns it combines are from one another. Consider two films, each spanning two narrative patterns. One mixes two closely related types of action tropes; the other mixes horror tropes with romantic comedy tropes. Even though both films span the same number of patterns, the second is likely to bridge a larger narrative gap.

To measure how far apart two patterns are, we look at how often they appear together in the same films. Patterns that frequently appear together in films represent familiar combinations that creators and audiences are accustomed to; patterns that rarely co-occur represent more unusual pairings. Formally, let F_i denote the set of films containing at least one trope from community i. The Jaccard distance between trope communities i and j is:

$$d_{ij} = 1 - \frac{|F_i \cap F_j|}{|F_i \cup F_j|}.$$

Given these pairwise distances between patterns, we aggregate them at the film level. Each film's spanning distance is the weighted average distance between the patterns it draws from:

$$\mathrm{SD}_m = \frac{\sum_{i<j} w_{mi}\, w_{mj}\, d_{ij}}{\sum_{i<j} w_{mi}\, w_{mj}}.$$

Spanning distance is high when a film combines narratively distant patterns, and low when the patterns it mixes tend to co-occur in other films as well. 1,056 films draw all of their tropes from a single community, making spanning distance impossible to compute; these films are excluded from the estimation sample.

4.3 Regression Model

We examine how coherence and spanning distance relate to two outcomes: *IMDb rating*, which captures audience evaluation, and *number of IMDb votes*, which captures audience attention. Because vote counts are highly right-skewed, we use $\log(1 + \textit{Votes})$ as the dependent variable.

We estimate OLS regressions with HC3 robust standard errors:

$$Y_m = \beta_0 + \beta_1 \mathrm{HHI}_m + \beta_2 \mathrm{SD}_m + \beta_3 \, \textit{trope_count}_m + \gamma_{\textit{year}} + \delta_{\textit{genre}} + \varepsilon_m.$$

Both HHI and SD are mean-centered prior to estimation. All models control for the total number of tropes in a film, because films with more tropes tend to span more communities simply because they have more tropes to distribute(García-Ortega et al., 2021). We also include year fixed effects to absorb temporal trends in both film production and TVTropes annotation coverage(García-Ortega et al., 2020), and main-genre fixed effects to ensure that the estimated relationships are not driven by genre differences in ratings or votes(Francemone et al., 2023; Shahid and Islam, 2023).

5 Results

Table 2 reports the main regression estimates. The correlation between HHI and SD in the estimation sample is 0.09, and the variance inflation factors are close to 1, indicating no multicollinearity concern.

HHI is positive and significant in both models ($\beta = 1.089$, $p < 0.01$ for rating; $\beta = 1.207$, $p < 0.01$ for votes). These results support H1, indicating that films whose tropes concentrate within fewer narrative patterns tend to receive higher evaluations and more audiences. Spanning distance shows opposite signs across the two outcomes. For ratings, the coefficient is positive and significant ($\beta = 0.477$, $p < 0.01$); for votes, it is negative and significant ($\beta = -0.557$, $p < 0.01$). These results support H2a and H2b: films that bridge more distant narrative patterns tend to receive higher evaluations but attract smaller audiences.

Table 2: Regression results for trope concentration and spanning distance

	(1) Rating	(2) $\log(1 + \text{Votes})$
HHI_c	1.089^{***}	1.207^{***}
	(0.048)	(0.082)
CTD_c	0.477^{***}	-0.557^{***}
	(0.104)	(0.182)
Trope count	0.005^{***}	0.017^{***}
	(0.0001)	(0.0002)
Year FE	Yes	Yes
Main genre FE	Yes	Yes
Observations	17,108	17,108
R^2	0.258	0.409
Adj. R^2	0.251	0.404

Notes. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

6 Discussion

Our results show that narrative structure shapes audience response, and that it does so differently for evaluation and attention. Coherence benefits films on both dimensions. We extended category-conformity arguments (Zuckerman, 1999; Askin and Mauskapf, 2017) from market-level genre labels to the narrative itself: a clear narrative structure helps not only in attracting audience attention but in generating favorable evaluations. Spanning distance reveals that evaluation and attention respond to different mechanisms. Films combining distant patterns may be harder to categorize from external signals, reducing their attention (Hsu, 2006). Yet for audiences who have already watched, narrative novelty may be appreciated rather than penalized (Uzzi et al., 2013). More broadly, this study shows that tropes can be treated not only as descriptive labels but as building blocks of narrative patterns whose combination helps explain success in cultural markets.

Limitations

This study has several limitations. First, TVTropes is a user-edited wiki, so popular films inevitably receive more thorough annotation. This creates a systematic bias between trope coverage and audience attention that our design cannot fully address. Second, the narrative patterns are sensitive to the resolution parameter of the Leiden algorithm, and the presence of a general community may compress spanning distance estimates.

References

Alberto Álvarez and Jose M. Font. 2022. TropeTwist: Trope-based narrative structure generation. In *Proceedings of the International Conference on Computational Creativity*.

Noah Askin and Michael Mauskapf. 2017. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944.

Pierre Bourdieu. 1996. *The rules of art: Genesis and structure of the literary field*. Stanford University Press.

John G. Cawelti. 1976. *Adventure, Mystery, and Romance: Formula Stories as Art and Popular Culture*. University of Chicago Press.

Hsi-Ping Chou, Changxu Huang, and Yun-Nung Chen. 2023. TaleStream: Story generation with tropes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Paul DiMaggio. 1987. Classification in art. *American Sociological Review*, 52(4):440–455.

C Joseph Francemone, Kevin Kryston, and Matthew Grizzard. 2023. Examining genre success, co-occurrence, release, and production of 9,068 films over twenty years. *Poetics*, 99:101794.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science*.

Rubén Héctor García-Ortega, Pablo García-Sánchez, and Juan Julián Merelo. 2021. Trope community structure in TV series. In *Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design*. Springer.

Rubén Héctor García-Ortega, Pablo García-Sánchez, and Juan Julián Merelo-Guervós. 2020. Startroper, a film trope rating optimizer using machine learning and evolutionary algorithms. *Expert Systems*, 37(6):e12525.

Michael T Hannan and John Freeman. 1989. *Organizational ecology*. Harvard university press.

Greta Hsu. 2006. Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly*, 51(3):420–450.

Greta Hsu, Michael T. Hannan, and Özgecan Koçak. 2009. Multiple category memberships in markets: An integrative theory and two empirical tests. *American Sociological Review*, 74(1):150–169.

Marc Keuschnigg. 2015. Product success in cultural markets: The mediating role of familiarity, peers, and experts. *Poetics*, 51:17–36.

Marc Keuschnigg and Thomas Wimmer. 2017. Is category spanning truly disadvantageous? new evidence from primary and secondary movie markets. *Social Forces*, 96(1):449–479.

Balázs Kovács and Michael T Hannan. 2015. Conceptual spaces and the consequences of category spanning. *Sociological science*, 2:252–286.

Jennifer C Lena and Richard A Peterson. 2008. Classification as culture: Types and trajectories of music genres. *American sociological review*, 73(5):697–718.

Chris Mellina and Svetlana Svetlichnaya. 2011. Trope propagation across cultural products. *arXiv preprint*.

Giacomo Negro and Ming D Leung. 2013. "actual" and perceptual effects of category spanning. *Organization Science*, 24(3):684–696.

Elizabeth G Pontikes. 2012. Two sides of the same coin: How ambiguous classification affects multiple audiences' evaluations. *Administrative Science Quarterly*, 57(1):81–118.

Hayagreeva Rao, Philippe Monin, and Rodolphe Durand. 2005. Border crossing: Bricolage and the erosion of categorical boundaries in french gastronomy. *American Sociological Review*, 70(6):968–991.

Stephen A Rhoades. 1993. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79:188.

Stoyan V Sgourev, Erik Aadland, and Giovanni Formilan. 2023. Relations in aesthetic space: How color enables market positioning. *Administrative Science Quarterly*, 68(1):146–185.

Muzammil Hussain Shahid and Muhammad Arshad Islam. 2023. Investigation of time series-based genre popularity features for box office success prediction. *PeerJ Computer Science*, 9:e1603.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Ezra W. Zuckerman. 1999. The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology*, 104(5):1398–1438.

Table 3: Example trope communities identified by the Leiden algorithm for the co-tropes network.

ID	#Tropes	#Films	Dens.	Center trope	Representative tropes
0	10,573	18,164	0.170	Shout-Out	Shout-Out, Oh Crap!, Foreshadowing, Big Bad, Chekhov's Gun
1	3,658	10,475	0.084	Eldritch Abomination	Eldritch Abomination, Our Ghosts Are Different, Demonic Possession, Our Vampires Are Different, Sealed Evil In A Can
2	2,565	9,428	0.083	Doomed Upgrade	Doomed Upgrade, Sci-Fi Flyby, Humans Advance Swiftly, Proof of Commitment, No Warping Zone
3	1,703	7,878	0.090	Converted Into a Weapon	Converted Into a Weapon, Older Than Steam, The Speedster, Single-Power Superheroes, Superhero Capital of the World
4	1,453	7,288	0.129	Face Framed In Shadow	Face Framed In Shadow, Dramatic Unmask, Bank Robbery, No Honor Among Thieves, Evil Wears Black

Register Mixing Is the Norm on the Web

Erik Henriksson, Alireza Razzaghi, Tuomas Lundberg, Antti Kanner, Veronika Laippala
TurkuNLP, University of Turku
{eihenr,alireza.razzaghi,tuomas.w.lundberg@utu.fi,antti.kanner,mavela}@utu.fi

Abstract

Nearly all studies on web registers—online text varieties associated with characteristic social contexts and linguistic features—use full documents as the unit of analysis. However, web documents often contain sections in different registers. A cooking blog, for instance, may combine personal storytelling, recipe instructions, user comments, and promotional text within a single URL. This internal variation raises doubts about the validity of document-level register labeling. In this paper, we propose an LLM-based approach that identifies register-homogeneous segments within documents and apply it to a 10,000-document English sample from HPLT 3.0. We show that segmentation addresses persistent problems in register analysis, including low inter-annotator agreement and category fuzziness. Strikingly, it also reveals that most web documents contain more than one register, making register mixing the norm rather than the exception on the web.

1 Introduction

Register, the linguistic variety associated with a particular communicative situation and social context, has long been recognized as a central organizing principle of language use (Halliday, 1978; Biber, 1988; Matthiessen, 2019; Biber and Conrad, 2019). On the web, register variation is particularly diverse, ranging from news articles and recipes to forum discussions and product reviews, each with distinct situational configurations and linguistic features. Corpus linguists have focused on identifying and classifying these varieties, with the Corpus of Online Registers of English (CORE; Egbert et al., 2015; Laippala et al., 2023) being the foundational large-scale example. CORE established a taxonomy of web registers and a document-level annotation methodology widely adopted in corpus linguistics and NLP (e.g., Laippala et al. 2019; Biber et al. 2020; Repo et al. 2021; Henriksson et al. 2024, 2025a; Myntti et al. 2025; Burchell et al. 2025).

However, document-level analysis oversimplifies web text by treating individual documents as instances of a single register. Within a single URL, a cooking blog post may consist of multiple sections with distinct communicative purposes. These sections may include promotional text, information about the author, cooking instructions, or a comment section. Some studies have addressed this document-internal register mixing by allowing individual annotators to assign multiple register labels to a single document (e.g., Skantsi and Laippala, 2023), which better reflects the mixed nature of web text. However, multi-labeling at the document level lacks the granularity needed to distinguish true register-mixing within a passage from register shifts between distinct text segments (Egbert and Gracheva, 2022). The mismatch between theory and annotation unit has concrete empirical consequences. Inter-annotator agreement on CORE reaches only $\kappa = 0.47$ (Egbert et al., 2015), and no machine learning model has achieved above 80% F1 on document-level web register classification (Henriksson et al., 2025b).

In this paper, we introduce a pipeline for analyzing the internal register composition of web documents (`github.com/TurkuNLP/register-segments-hplt-en`). We utilize the XML structure of the HPLT 3.0 web corpus (Oepen et al., 2025) to segment documents into paragraphs and use a large language model (LLM) to label each paragraph, merging consecutive same-register segments. Human evaluation shows that the LLM achieves near-human annotation quality, and a classifier fine-tuned on the resulting labels outperforms document-level baselines. Applied to a 10,000-document English sample, the pipeline reveals that the majority of web documents contain many registers, with characteristic positional patterns, co-occurrences, and transition sequences. Linguistic analysis confirms that the segment-level labels represent genuine dimensions of register variation.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 138–149
July 4, 2026 ©2026 Association for Computational Linguistics

2 Background

2.1 Document-level register annotation

The Corpus of Online English (CORE) (Egbert et al., 2015; Laippala et al., 2023) established a widely used approach for corpus-linguistic web register research, in which annotators assign a register label to documents from a taxonomy of eight main registers with optional subregisters. In this context, a *document* refers to the content retrieved from a single URL by a web crawler and stored as a single record in a corpus. In document-level analysis, register labels often correspond to what are sometimes interchangeably called web *genres* (Kuzman and Ljubešić, 2025). The sub-document approach we take here arguably makes the *genre* vs. *register* distinction clearer and comes closer to a traditional understanding of register as a functional text variety (Goulart et al., 2020). In our approach, a web genre such as *recipe* or *blog* may contain multiple register varieties, each with their own communicative purpose within the document.

The CORE approach has been applied to various languages besides English and used to develop automatic register classifiers (e.g., Repo et al., 2021; Skantsi and Laippala, 2023; Henriksson et al., 2025b). However, no register classifier has exceeded 80% micro F1 (Henriksson et al., 2025a). This ceiling may partly reflect the inherent fuzziness of register categories (Biber et al., 2020; Henriksson et al., 2024), but another likely factor is the unit of analysis itself. Many web documents are composites made of multiple discourse units with different communicative purposes rather than unified wholes (Biber et al., 2020; Biber and Egbert, 2023). This within-document variation is invisible to document labels and contributes to annotator disagreement when different raters attend to different sections of the same page (Egbert et al., 2015).

2.2 Sub-document register segmentation

Text segmentation is the task of dividing a text into coherent, non-overlapping units (Hearst, 1994, 1997; Hearst and Plaunt, 1993; Galley et al., 2003; Liu et al., 2021). In the context of web register analysis, the goal is to identify spans of text that are internally consistent in terms of register.

Henriksson et al. (2025a) presented the first attempt to segment web documents by register, using a recursive binary splitting algorithm applied to ModernBERT-based register predictions (Warner et al., 2024). Working with the plain-text CORE documents, which have no structural markup, the algorithm identified register boundaries by iterating over sentence boundaries and splitting wherever the predicted register distributions of the two resulting halves diverged sufficiently. They found that sub-document segments carry a cleaner register signal than full documents: classifier F1 scores improved, embedding-based silhouette scores showed more distinct register clusters, and within-register linguistic variance decreased. The approach was limited, however, by the absence of structural markup in the source documents and by its dependence on an imperfect classifier for segmentation.

The present work extends this line of research by using an LLM to perform both segmentation and labelling in a single pass, and by applying the approach at scale to a modern web dataset.

3 Methods

3.1 Data

We sample 10,000 English documents from HPLT 3.0 (Oepen et al., 2025), a large multilingual web corpus that provides documents in an XML format encoding structural information such as paragraphs, headings, lists, and comment sections. The HPLT 3.0. corpus is sharded by Web Document Score (WDS; Pavón et al., 2023), a measure of text quality, and only documents with an overall WDS of 9–10 are selected.

We convert each document to a simplified plain-text format that retains structural markup as section headers (`main`, `comments`) and numbered paragraph indices, while stripping character-level formatting tags (e.g. `<hi>`). This yields a representation that is both human-readable and compact enough for LLM processing.

3.2 Classification scheme

We annotate three orthogonal facets per segment:

- **Source**: whether the text is *written*, *transcribed* speech, or *cannot_rate* (boilerplate, navigational, or incoherent content).
- **Interactive**: either *true* (participatory context: forums, comment sections, interviews) or *false* (monologic). Null for *cannot_rate* segments.
- **Function**: the primary communicative function, chosen from: *narrating*, *describing*, *instructing*, *opinionated*, *promoting*, or *lyrical*. Null for *cannot_rate* segments.

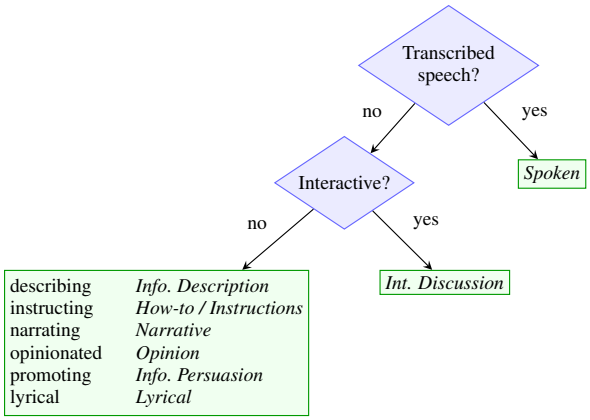


Figure 1: Mapping from the three annotation facets to CORE register labels, applied in hierarchical order.

The faceted scheme maps directly onto the CORE taxonomy following the decision tree in Biber and Egbert (2018, p. 17), as shown in Figure 1. Specifically, any segment with the source *transcribed* maps to CORE's *Spoken*; if the segment is *interactive*, it maps to *Interactive Discussion*; and the remaining written, non-interactive segments map directly via their function. In our analyses, we report results for both the original three-facet scheme and the CORE mappings.

3.3 LLM annotation pipeline

We annotate segments using two LLMs: Gemini 3 Flash (Google DeepMind, 2025) and Qwen3 235B A22B (Qwen Team, 2025). These models were selected for their strong performance, low inference cost, and availability through the OpenRouter API[1], which we used for the task. Each model receives a zero-shot prompt of approximately 1,100 words that defines the classification scheme (see Appendix B). Documents are converted into numbered segments following a simplified markdown format that distinguishes headings, paragraphs, blockquotes, lists, tables, and code snippets. The model classifies paragraphs, blockquotes, and list items by register; headings, tables, and code snippets are used as context but assigned *cannot_rate*. Segments are labeled 20 at a time with a sliding window of ± 5 segments of surrounding context.

After labeling, consecutive segments with identical register labels are merged into single spans, treating *cannot_rate* segments as transparent during merging so that boilerplate or navigational content does not artificially fragment register spans. The resulting register segments thus range from single paragraphs to multi-paragraph stretches of consistent register. The paragraph represents the resolution limit of the approach (i.e., sub-paragraph register shifts are ignored); however, the granularity is substantially finer than document-level analysis.

3.4 Evaluation and analysis

We evaluate our approach in three ways. First, to validate annotation quality, two human annotators independently labeled 100 randomly sampled segments; we compute Cohen's κ and F1 scores between the two humans (H1 vs. H2) and between each LLM and the human average. Second, to assess the learnability of the resulting labels, we finetune a ModernBERT (Warner et al., 2024) model on the LLM-generated segment labels and compare its performance against document-level baselines. Third, to verify that the register labels correspond to genuine linguistic differences, we tag the segments with 96 Biber-style features using BiberPlus (Alkiek et al., 2025) and apply PCA to the standardized register means.

Beyond validation, we use the segmented annotations to analyze the internal register structure of web documents, including register distributions, within-document combinations, positional patterns, co-occurrences, and sequential transitions.

4 Results and Discussion

4.1 Human evaluation of segments

Table 1 reports inter-annotator agreement and F1 scores for human–human and human–LLM comparisons for 100 randomly sampled segments.

Gemini 3 Flash reaches agreement comparable to the human annotators. Its κ of 0.77 on CORE-mapped labels and F1 of 0.81 are in the same range as the human–human κ of 0.76 and F1 of 0.80, though the small evaluation sample ($n = 100$) limits the precision of these estimates. Qwen3 235B A22B performs consistently lower, with a CORE-mapped κ of 0.63 and F1 of 0.69. The gap between the two models, despite Qwen3's relatively large size, suggests that register annotation is a non-trivial task where LLM choice matters. Based on these results, all subsequent analyses use Gemini-generated labels.

To assess whether the results in Table 1 generalize beyond the 100-segment sample, a single annotator independently labeled an additional 500 randomly sampled segments. Compared against Gemini 3 Flash, agreement was consistent with the smaller evaluation across all fields (κ = 0.88 /

[1] https://openrouter.ai

	Cohen's κ		F1	
Field	H1/H2	Gemini	H1/H2	Gemini
Source	0.84	0.74	0.96	0.94
Interactive	0.92	0.88	0.97	0.95
Function	0.70	0.74	0.76	0.80
CORE label	0.76	0.77	0.80	0.81

Table 1: Inter-annotator agreement (Cohen's κ) and F1 between human annotators (H1/H2) and between humans (averaged) and Gemini 3 Flash. Qwen3 235B A22B performs consistently lower (κ = 0.67, 0.76, 0.56, 0.63; F1 = 0.91, 0.90, 0.83, 0.69 for the same fields).

0.94 / 0.81 / 0.83 for source, interactive, function, and CORE label respectively; weighted F1 = 0.96 / 0.97 / 0.86 / 0.86). A human–human baseline at this scale was not conducted, however.

Importantly, both human–human and Gemini agreement substantially exceed the κ of 0.47 reported for document-level CORE annotation (Egbert et al., 2015), suggesting that paragraph-level segments constitute a more stable unit for register annotation than full documents.

Most disagreements between humans and the LLM involve conceptually proximate functions, most notably *describing* versus *narrating*. Consider the following example:

> *[Context -1]* Lila also received some money for her savings account from MeMe and Henry.
>
> **[TARGET]** These were the gifts that Andrew and I gave her: the Jesus Storybook Bible, a PraiseBaby CD and DVD set, and a pretty, wooden cross frame.
>
> *[Context +1]* We love you baby girl! Our hope is that you grow to love Jesus and know him as your friend and Saviour!

One human annotator labeled the target segment as *describing*, reading it as an informational list. The LLM and the other human labeled it as *narrating*, interpreting it as part of an account of an event. The segment does refer to a concrete past event, but it lacks temporal sequencing, making both readings defensible. The disagreement thus reflects genuine ambiguity in the text rather than misclassification. Most of the LLM's "errors" are of a similar kind.

4.2 Learnability of segments

To assess whether the segmented labels yield a learnable register signal, we fine-tuned a ModernBERT-large (Warner et al., 2024) model on the LLM-generated segment labels (segments $\geq$ 100 words; $n \approx$ 36k). The supervised model achieves 0.83 F1 on CORE-mapped labels. For reference, document-level English CORE classifiers have not exceeded 0.76 F1 (Henriksson et al., 2024), though the two settings differ in training data, label source, and text granularity, making direct comparison difficult. The higher figure is nonetheless consistent with the finding in Henriksson et al. (2025a) that segmentation produces a cleaner register signal, and suggests that register segments are an easier classification target than full documents. A caveat is that the fine-tuning data consist of LLM-generated labels, so systematic annotation biases may propagate into the model.

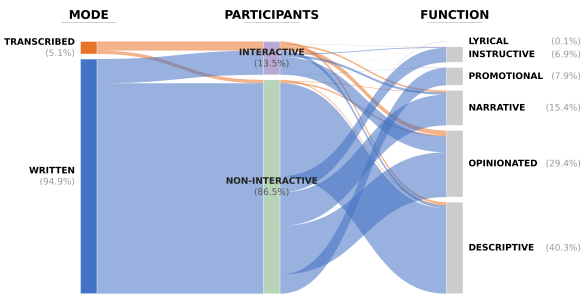


Figure 2: Distribution of register facets across all labelled segments (excluding *cannot_rate*). Flow widths are proportional to segment counts.

4.3 Distribution of register facets

Figure 2 shows how labeled segments (excluding *cannot_rate*) distribute across the three annotation facets source, interactivity, and function. The vast majority of segments are *written* (94.9%), with *transcribed* content accounting for only 5.1%.

The middle column reveals an asymmetry between the two source types. Written segments are predominantly non-interactive (86.5% overall), whereas transcribed segments flow almost entirely into the interactive category. This follows naturally from the nature of transcribed web content, which consists largely of interviews, podcast transcripts, and similar participatory varieties.

Turning to communicative function, the dominant category is *describing* (40.3%), followed by *opinionated* (29.4%) and *narrative* (15.4%). The prevalence of *opinionated* segments among interactive content is unsurprising: comment sections and forum replies are prototypically evaluative. More noteworthy is the relative scarcity of *narrative* at the segment level. In document-level CORE, *Narrative* is the single most frequent category, accounting for approximately 40% of documents (Egbert et al., 2015). One possible explanation is that documents labeled as narratives at document level are

Figure 3: Distribution of the number of distinct registers per document, under the full faceted scheme (blue) and CORE-mapped labels (orange).

internally heterogeneous, combining a narrative frame with stretches of other text varieties.

4.4 Within-document register combinations

Figure 3 shows the distribution of distinct register counts per document under the faceted and CORE labeling schemes. Since *cannot_rate* segments are treated as transparent during merging (Section 3.3), these counts reflect genuine register shifts rather than fragmentation by boilerplate or navigational content.

Fewer than 20% of documents contain a single register, and most often, documents contain two or three registers, with one and four registers being the next most common. The view offered by the segment analysis is therefore radically different from the document-level analysis in CORE, where 69% of documents received a single consensus register label from multiple annotators (Egbert et al., 2015, p. 10). More broadly, this perspective challenges the common assumption in corpus linguistics that register categories can be meaningfully assigned to entire documents.

Figure 4 zooms in on the specific combinations of registers, showing the 15 most frequent CORE-label co-occurrences across documents. Of these, only four are single-register: *Informational Description* (9.1%), *Informational Persuasion* (3.6%), *Opinion* (3.1%) and *Narrative* (2.3%). The most frequent combination is *Informational Description* + *Opinion* (9.7%), which alone outnumbers any single-register category. Three-way combinations are also common. *Informational Description* + *Opinion* + *Narrative* ranks third overall (8.6%); and the seventh most frequent pattern combines as many as four registers.

Figure 5 shows the number of distinct registers per document as a function of document length. Although longer documents tend to contain more registers, the relationship is weak: even short documents ($\sim 10^3$ characters) frequently contain two or three distinct registers, and the modal count remains at two across most of the length range. Register mixing is thus not simply a byproduct of document length, but a persistent feature of web text regardless of size.

4.5 Positional distribution of registers

To examine where registers tend to appear within documents, we computed the probability mass distribution of each CORE-mapped label as a function of relative character position, accumulated into 128 equal-width bins and smoothed with a Gaussian filter ($\sigma = 2$). Each label is normalized independently, so curves reflect positional *shape* rather than overall frequency.

The results reveal clear positional patterns (Figure 6). *Interactive Discussion* rises near-monotonically from near zero to a peak at position 100%, directly reflecting the convention that comment sections appear after main content on most webpages. Conversely, *Narrative* is front-loaded, peaking in the first 5–10% and declining steadily. *How-to/Instructions* peaks later (around 75–80%), likely following introductory or contextualizing material. *Spoken* content peaks early (15–20%) and declines gradually, plausibly reflecting embedded interview excerpts near document openings.

These patterns suggest that register mixing is not random but *structurally organized*: registers occupy characteristic positions that follow conventions of web genre composition.

4.6 Register co-occurrence patterns

To examine which registers tend to co-occur within the same document, we compute two measures. The conditional probability matrix shows $P(\text{col} \mid \text{row})$: given that a document contains the row register, the probability that it also contains the column register. The pointwise mutual information (PMI) matrix shows $\log \frac{P(A \cap B)}{P(A) \cdot P(B)}$: how much more often two registers co-occur than expected by chance. As PMI is symmetric, only the lower triangle is shown.

The conditional heatmap (Figure 7) shows that *Informational Description* co-occurs with almost every register at high rates, largely as a base-rate effect of its overall frequency. The PMI matrix (Figure 8) controls for this and reveals that *Informational Description* has near-zero PMI with most registers, suggesting it in fact functions as a register-neutral backdrop in web documents.

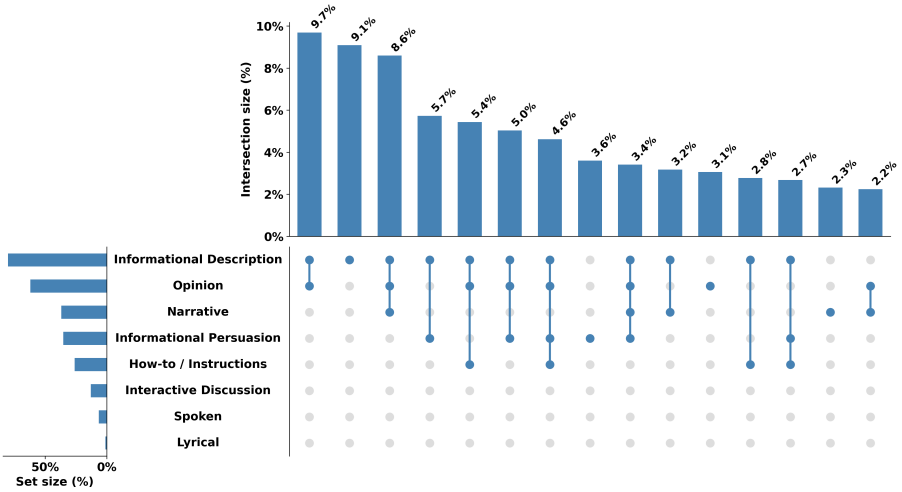


Figure 4: UpSet plot of the 15 most frequent CORE-mapped register combinations across documents. Horizontal bars show per-register document percentages; vertical bars show combination percentages.

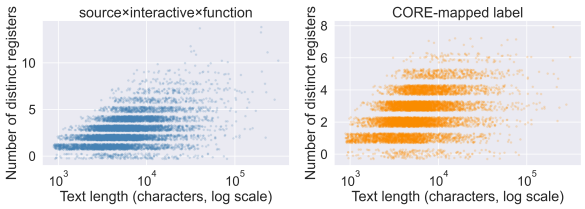


Figure 5: Number of distinct registers per document as a function of document length (log scale), under the faceted scheme (left) and CORE-mapped labels (right).

The clearest pattern is the isolation of *Informational Persuasion*, which has negative PMI with nearly every register except *How-to/Instructions* (0.25), plausibly because product and service pages often include usage instructions. *Interactive Discussion* is the opposite: positive PMI with *Narrative* (0.31), *Lyrical* (0.22), and *Opinion* (0.16), but negative with *Informational Persuasion* (–0.41) and *Informational Description* (–0.14). Comment sections attract subjective and narrative content but repel informational and promotional material.

The strongest positive associations are *Spoken+Lyrical* (1.07) and *Narrative+Lyrical* (0.71), though both should be interpreted cautiously given that *Lyrical* and *Spoken* are among the rarest registers (Figure 4), making PMI sensitive to small counts. The LLM may also conflate poetic prose with informal spoken-style writing, which would inflate their apparent association.

4.7 Register transitions

To examine sequential register structure, we compute a PMI matrix over register transitions, where each value reflects how much more or less likely a given register-to-register sequence is relative to base rates. Consecutive segments with identical labels have been collapsed before counting (Section 3.3), so the matrix captures genuine register shifts.

The results (Figure 9) reveal clear sequential preferences. The strongest avoidances involve *How-to/Instructions*, which almost never transitions into *Spoken* (-2.52) or *Lyrical* (-1.21), and *Informational Persuasion*, which strongly avoids *Lyrical* (-1.64) and *Spoken* (-1.15). In other words, promotional and instructional content is rarely followed by lyrical or conversational text.

On the positive side, *Informational Description* functions as a transitional register, showing above-chance transitions to and from most other registers, even when base rates are taken into account. This contrasts with the co-occurrence analysis, where *Informational Description* showed near-zero PMI with most registers. It appears that when documents shift between registers, they tend to pass through descriptive stretches.

Narrative and *Spoken* show near-symmetric mutual transitions (0.89 and 0.86), suggesting genuine alternation. This *Spoken*↔*Narrative* pattern likely reflects genre-level structure rather than register affinity per se: interview or podcast transcripts are a plausible host genre in which spoken framing and narrative content naturally interleave: for example, a guest might recount an experience (*Narrative*) within a conversational exchange (*Spoken*). This points to a theoretical limitation of flat register segmentation: genres such as interviews impose a hierarchical structure within which registers alternate in conventionalized ways. Modeling this hierarchy

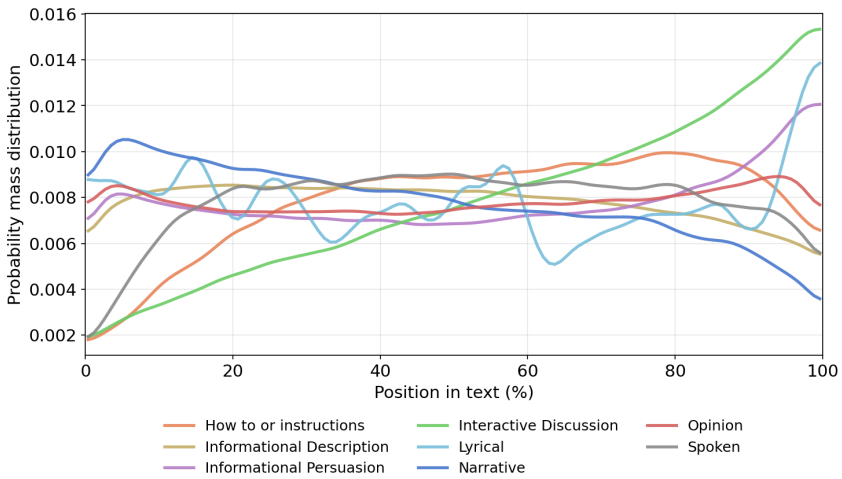


Figure 6: Distribution of CORE-mapped register labels by relative document position. Each label is normalised independently; curves reflect positional shape rather than overall frequency.

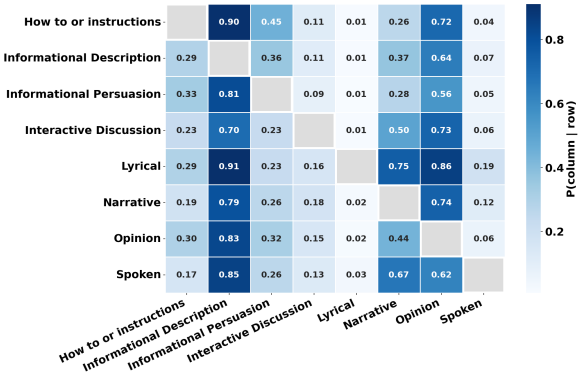


Figure 7: Conditional probability heatmap: $P(\text{col} \mid \text{row})$, the proportion of documents containing the row register that also contain the column register.

Figure 8: Pointwise mutual information (PMI) between register pairs. Positive values indicate above-chance co-occurrence; negative values indicate avoidance.

explicitly is left for future work.

4.8 Linguistic analysis

To verify that the segment-level register labels reflect genuine register variation, we analyze the

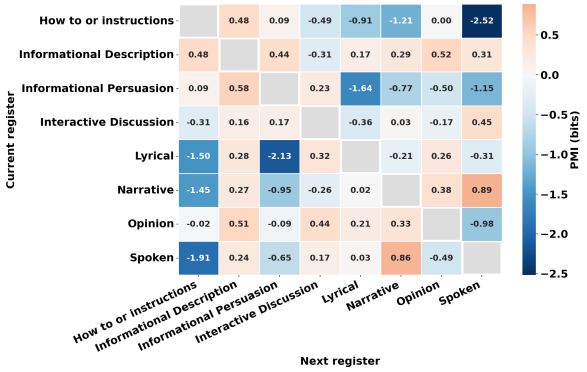


Figure 9: PMI matrix over register transitions (rows = current, columns = next). Positive values indicate above-chance sequences; negative values indicate avoidance.

Biber-style feature profiles of each register using BiberPlus (Alkiek et al., 2025), a neural tagger that estimates 96 lexico-grammatical features per text. Segments shorter than 50 words are excluded, and features with near-zero variance or pairwise correlations above 0.95 are removed to reduce redundancy. We compute mean feature values per register label and apply PCA to the z-scored means. We focus on PC1 and PC3, which together account for 52.3% of variance and yield the most interpretable dimensions; PC2 was less clearly interpretable and is omitted for brevity. Table 2 reports the five highest-loading features at each pole of each component.

PC1 (37.4%) corresponds to the involved versus informational opposition that has been identified as the primary dimension of register variation across numerous Multi-Dimensional studies (Biber, 1988; Biber and Egbert, 2016). The positive pole is defined by private verbs, contractions, that-deletion,

Dimension	Feature	Loading
PC1 (37.4%) – Involved (+) vs. Informational (−)		
+	Private verbs	+0.179
+	Contractions	+0.171
+	That-deletion	+0.170
+	*because*	+0.169
+	Direct WH-questions	+0.161
−	Mean word length	−0.168
−	Present participles	−0.165
−	Other nouns	−0.163
−	Present participial WHIZ deletion	−0.140
−	Attributive adjectives	−0.132
PC3 (18.4%) – Instructional (+) vs. Narrative (−)		
+	Second-person pronouns	+0.246
+	Infinitive verbs	+0.227
+	Predictive modals	+0.216
+	*if/unless*	+0.190
+	Possibility modals	+0.179
−	Sentence relatives	−0.218
−	By-passives	−0.197
−	Third-person pronouns	−0.196
−	Conjuncts	−0.195
−	Perfect aspect	−0.192

Table 2: Top 5 feature loadings at each pole of PC1 and PC3, from PCA on z-scored BiberPlus register means.

because, and direct WH-questions; the negative pole by mean word length, nouns, present participles, and attributive adjectives. In Figure 10, the registers spread along this axis as expected: *Interactive Discussion*, *Lyrical*, and *Spoken* on the involved side, *Informational Description* on the informational side.

PC3 (18.4%) separates directive from narrative discourse. The positive pole is loaded with second-person pronouns, infinitive verbs, predictive and possibility modals, and conditional subordinators; the negative pole with sentence relatives, by-passives, third-person pronouns, conjuncts, and perfect aspect. This component appears to merge aspects of Dimensions 5 (Irrealis vs. Informational Narration) and 3 (Oral Narrative vs. Written Information) from Biber and Egbert (2016) into a single axis. In the figure, *How-to/Instructions* and *Informational Persuasion* score highest, while *Narrative* occupies the extreme negative end. *Spoken* is worth noting: it is highly involved on PC1 but leans toward the narrative pole on PC3, which is consistent with interview and podcast transcripts that recount events conversationally.

Together, the two components account for over half the variance in 96 features and align with dimensions established in previous work, strongly suggesting that the segment-level labels reflect real linguistic differences rather than artifacts of the LLM annotation.

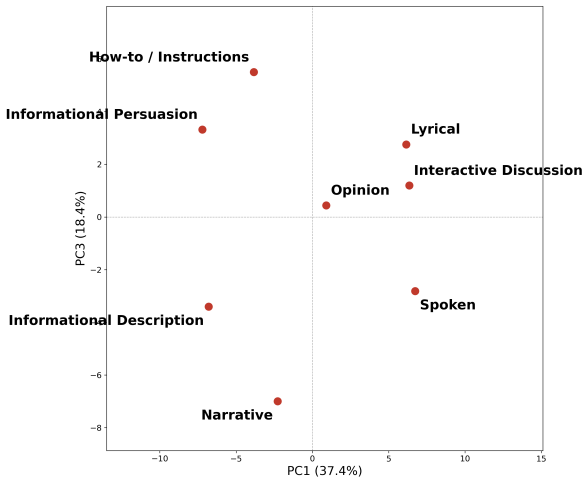


Figure 10: Mean register positions on PC1 (Involved vs. Informational) and PC3 (Directive vs. Narrative), from PCA on z-scored BiberPlus features.

5 Conclusions

We have presented an LLM-based pipeline for register segmentation that achieves annotation quality comparable to human annotators. Applied to 10,000 web documents from HPLT 3.0, the pipeline reveals that register mixing is the norm online: fewer than 20% of documents contain a single register, contrasting sharply with the document-level view in CORE where 69% received a single consensus label. Segmentation also improves inter-annotator agreement (κ = 0.77 vs. 0.47) and classifier learnability (F1 = 0.83 vs. 0.76). Moreover, the segmented annotations reveal that registers occupy characteristic document positions, show systematic co-occurrence and transition patterns, and align with known dimensions of linguistic variation. Together, the findings suggest that the web document is not a natural unit for register analysis.

Limitations

Our human evaluation covered only 100 segments with two annotators; a single-annotator evaluation on 500 segments yielded similar results, but larger-scale evaluation is left for future work. As the ModernBERT classifier is trained on LLM-generated labels, systematic biases may propagate into it. The segmentation also relies on a simplified version of HPLT's (itself noisy) XML structure, and our sample was restricted to high-quality documents (WDS 9–10), leaving open whether register mixing is equally common in lower-quality text. Future work should explore the relationship between genre-level structure and register segmentation.

Acknowledgements

Alireza Razzaghi received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101177564 (HAIF). Cofunded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Furthermore, this work was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union – NextGenerationEU instrument), "Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora" (project 362459), "Green NLP: Controlling the Carbon Footprint in Sustainable Language Technology" (project 353167), and the "Centre of Excellence Human Diversity through Contacts" (project 374223). We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

Kenan Alkiek, Anna Wegmann, Jian Zhu, and David Jurgens. 2025. Neurobiber: Fast and interpretable stylistic feature extraction. *Preprint*, arXiv:2502.18590.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2023. What is a register?: Accounting for linguistic and situational variation within–and outside of–textual varieties. *Register Studies*, 5(1):1–22.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long papers, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Jesse Egbert and Marianna Gracheva. 2022. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash model card. Technical report, Google DeepMind.

Larissa Goulart, Bethany Gray, Shelley Staples, Amanda Black, Aisha Shelton, Douglas Biber, Jesse Egbert, and Stacey Wizner. 2020. Linguistic perspectives on register. *Annual Review of Linguistics*, 6(Volume 6, 2020):435–455.

Michael Alexander Kirkwood Halliday. 1978. *Language as Social Semiotic*. Arnold, London.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, page 59–68, New York, NY, USA. Association for Computing Machinery.

Erik Henriksson, Saara Hellström, and Veronika Laippala. 2025a. Analyzing register variation in web texts through automatic segmentation. In *Proceedings of the 5th International Conference on Natural*

Language Processing for Digital Humanities, pages 7–19, Albuquerque, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, and Veronika Laippala. 2024. From discrete to continuous classes: A situational analysis of multilingual web registers with LLM annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 308–318, Miami, USA. Association for Computational Linguistics.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2025b. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Taja Kuzman and Nikola Ljubešić. 2025. Automatic genre identification: A survey. *Language Resources and Evaluation*, 59(1):537–570.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2021. End-to-end segmentation-based news summarization. *Preprint*, arXiv:2110.07850.

Christian MIM Matthiessen. 2019. Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. In *Conference on Language Modeling (COLM)*, Montreal, Canada.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models. *Preprint*, arXiv:2511.01066.

Pablo Pavón and 1 others. 2023. Web docs scorer. `https://github.com/pablop16n/web-docs-scorer`. GitHub repository.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *arXiv preprint arXiv:2102.07396*.

Valtteri Skantsi and Veronika Laippala. 2023. Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, 48(1):1–31.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

A Appendix: Classification scheme details

The three-facet classification scheme assigns the following CORE-mapped labels based on combinations of Source, Interactive, and Function:

- Written + Non-interactive + Describing → **Informational Description**
- Written + Non-interactive + Instructing → **How-to/Instructional**
- Written + Non-interactive + Narrating → **Narrative**
- Written + Non-interactive + Opinion → **Opinion**
- Written + Non-interactive + Promoting → **Informational Persuasion**
- Written + Non-interactive + Lyrical → **Lyrical**
- Transcribed (any) → **Spoken**
- Written + Interactive (any function) → **Interactive Discussion**

B Appendix: Annotation prompt

The following prompt was used for zero-shot LLM annotation, prepended to each batch of 20 segments presented to the model.

Line format conventions Each line is numbered like [1], [2], etc. Lines use simplified markdown: `# text` = heading; plain text = paragraph; `> text` = blockquote; `- item1; item2; item3` = list items joined by semicolons; `TABLE: ...` = table data; `CODE: ...` = code snippet. Use headings, tables, code, and surrounding lines to understand the document's overall purpose, but classify every line in the [CLASSIFY THESE LINES] section.

Categories For each line, assign three fields.

source (always required)

- *written* — originally composed as written text. This is the default for most web content. Informal or ungrammatical writing is still "written" — poor grammar does not mean speech.
- *transcribed* — text that was originally spoken aloud and then written down. Use ONLY when there is clear evidence of oral origin: explicit transcript labels, speaker turn markers (e.g. "Interviewer:", "Q:", "[Speaker]"), or the text is clearly a transcript of a speech, podcast, interview, or hearing. Do NOT use just because the writing is informal, conversational, or ungrammatical. A letter or written document that is quoted or embedded in a page is "written" even if introduced with "here is the full text."
- *cannot_rate* — use for: (1) machine-generated, boilerplate, navigational, or structural content with no substantive human-authored message (e.g. cookie notices, auto-generated footers, breadcrumb navigation, "Related posts" lists, copyright lines, "Share on Facebook", short bylines/datelines, section labels like "7 Parts:", calls to edit/contribute like "Click EDIT to write this answer", and metadata). Note: an author bio with full sentences describing qualifications is substantive — classify it by its function, not as cannot_rate. (2) Incoherent, garbled, or spam-like text that lacks coherent human intent, including bad machine translations and keyword stuffing. (3) Fragments too short to classify — isolated numbers, single words. A complete sentence with clear communicative intent is always classifiable as written or transcribed, not cannot_rate.

interactive (required if source is *written* or *transcribed*; null otherwise)

- *true* — the text is produced in a participatory context where responses from other participants are expected or possible. This includes: forum threads, comment sections, discussion boards, social media posts with replies, chat messages, and interviews with alternating speakers. A single comment or reply is interactive even without visible back-and-forth, because it is produced within a discussion context.
- *false* — monologic: one author/speaker addressing a general audience with no participatory context. Articles, blog posts (the post itself, not comments), encyclopedia entries, guides, stories.

Important: editorially structured Q&A content is NOT interactive. This includes FAQ pages, wiki-style Q&A, and how-to sites where questions and answers are organized as reference content rather than real conversation between participants.

function (required if written/transcribed; null otherwise) The key question: what communicative act is the author performing? Classify by the DOMINANT purpose of the line. If a line serves multiple functions (e.g. describes a feature while also promoting it), choose the function that best captures the author's primary intent. Comments and forum posts can serve any function — people in discussions also narrate events, describe facts, ask questions, and give instructions.

Valid values (no other values are permitted):

- *narrating* — recounting specific events or stories in temporal sequence. Requires specific events that actually happened (or are presented as having happened). This includes news reports of events, historical accounts, and biographical passages that recount what someone did. Factual reference material about a topic is NOT narration even if it mentions dates or timelines. Hypothetical scenarios and descriptions of typical situations are also NOT narration.
- *describing* — presenting factual information, concepts, states of affairs, or general knowledge. This covers neutral reference material, encyclopedic content, and explanatory or analytical writing. Minor framing or incidental

evaluative language does not disqualify a line from "describing" — the test is whether the primary purpose is to inform the reader about facts rather than to argue a position or express a judgment. However, if evaluative language is prominent or the line's purpose shifts toward advocacy or critique, prefer "opinionated."

- *instructing* — teaching the reader how to do something, or providing direct answers to how-to or problem-solving questions. How-to guides, tutorials, recipes, technical instructions, troubleshooting steps, and Q&A content where the answer provides information or guidance in response to a question. Recipe components such as ingredient lists, quantities, and preparation notes are "instructing" — they are part of the instructional act of telling the reader how to make something.

- *opinionated* — expressing subjective views, evaluations, arguments, complaints, praise, or commentary where the evaluative stance is the primary purpose of the line. Includes text that uses narrative framing primarily to support an evaluative point (when an anecdote serves to complain, critique, praise, or argue, the function is "opinionated"). Also includes short expressive utterances like "Thanks!", "Great post!", "Agreed" — these express a stance or sentiment, not information. Signals include: sustained value judgments, subjective adjectives ("essential", "crucial", "best"), rhetorical questions, prescriptive statements ("you should", "it's a must"), and recommendations. If text argues what should be done or advocates a position, it is "opinionated" even if the topic is factual.

- *promoting* — selling, advertising, or marketing a product, service, brand, or organization. This includes: explicit ads and calls to action; business pages describing their own services or qualifications to attract clients; SEO-style content stuffed with keywords for commercial purposes; fundraising appeals and donation requests; text where the author represents an organization and frames its offerings positively to potential customers. "Promoting" does not require explicit "buy now" language — if the text's purpose is to attract clients or donors, it is promoting.

- *lyrical* — poetry, song lyrics, verse, or artistic literary expression presented as primary content. A poem or song quoted within a review or critical essay should be classified by the outer framing purpose, not as lyrical.

Output format Return ONLY a JSON array with one object per line in [CLASSIFY THESE LINES]. Each classified line gets exactly one JSON object. No markdown fences, no commentary, no extra text.
Example:

```
[{"line":1,"source":"written","
    interactive":false,"function":"
    describing"},
{"line":2,"source":"cannot_rate","
    interactive":null,"function":null}]
```

Scaling Sentence Similarity for Classical Tibetan with Automatic Annotations

Shay Cohen[1,2], Jingyi Yang[3], Gal Rabinovitz[1,2], Sonam Choden[3], Ofir Shtrosberg[1,2], Nicola Bajetta[3], Goody Ben Horin[1,2], Rebecca Sundén[3], Omri Drori[1,2], Sonam Jamtsho[3], Dorji Wangchuk[3], Kfir Bar[1,2], Orna Almogi[3], Shai Fine[1]

[1]Data Science Institute, Reichman University, Herzliya, Israel
[2]Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
[3]University of Hamburg, Germany

Correspondence: shay.cohen02@post.runi.ac.il

Abstract

Identifying intertextual parallels is central to philology, traditionally requiring labor-intensive manual analysis. While digitized historical corpora enable automated approaches using semantic sentence embeddings, training such models requires large annotated datasets, which are scarce for low-resource languages. We address this challenge by introducing a scalable automatic annotation pipeline for training semantic embedding models for Classical Tibetan. Our method combines unsupervised contrastive bootstrapping with iterative pair mining, generating silver-standard similarity labels through two complementary annotation strategies: (1) an ensemble of embedding models and rerankers, and (2) an LLM-as-a-judge committee using best–worst scaling. When combined with a domain-specific, gold-standard annotated dataset for sequential fine-tuning, the resulting text-similarity model achieves a state-of-the-art Spearman correlation of 0.864 on the STS task. This enables effective semantic search in Classical Tibetan and provides a framework for automatic supervision in low-resource languages used in digital humanities.

1 Introduction

The digitization of global textual heritage has unlocked vast new opportunities for computational analysis within the Digital Humanities. A prime example is the Tibetan Buddhist literary corpus, the focus of this study, which represents one of the largest and most complex collections of pre-modern literature in the world. Spanning philosophy, logic and epistemology, ritual and meditation, traditional sciences, language and literature, and the arts. This textual archive, now being preserved by organizations such as the Buddhist Digital Resource Center (BDRC[1]), offers a profound window into human intellectual history. However, its sheer scale and the linguistic complexities of Classical Tibetan make it largely inaccessible to modern, large-scale computational methods.

A central scholarly interest in working with such corpora is intertextuality: the reuse, adaptation, and transformation of earlier materials across texts, genres, intellectual circles, and historical periods. Tibetan literature, in particular, is deeply interwoven through quotation, borrowing, commenting, and doctrinal reformulation. Tracing these relationships is essential for understanding intellectual transmission and historical development. However, rather than mere verbatim repetition, intertextual connections often involve paraphrase, stylistic harmonization, condensation, or expansion, making simple string matching insufficient. To detect intertextuality at this non-trivial semantic level, we require methods capable of identifying meaning-preserving variation across passages. A foundational tool for this task is a sentence similarity model. Such a model would enable scholars to perform semantic search, identify textual reuse, trace the evolution of concepts, and cluster related passages—tasks that are currently laborious or impossible. Yet, the development of high-performance models, such as sentence transformers (Reimers and Gurevych, 2019), is fundamentally dependent on massive, human-labeled datasets.

For Classical Tibetan, such resources are virtually nonexistent, making this requirement a critical bottleneck. The number of experts worldwide with the linguistic and doctrinal depth needed to annotate semantic similarity is exceptionally small. As a result, creating a gold-standard dataset through traditional annotation is logistically impractical and

[1]https://www.bdrc.io

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 150–166
July 4, 2026 ©2026 Association for Computational Linguistics

highly labor-intensive.

To address this data-scarcity challenge, we introduce and evaluate a novel pipeline for training a text-similarity model for Classical Tibetan using synthetic data, achieving performance comparable to models trained on gold-standard annotations. By creating an iterative pipeline that mines candidate pairs from an unannotated corpus and synthetically annotates them using model ensembles, we construct a large-scale, cost-effective, and high-quality training dataset. We use this data, alongside expert-annotated examples, to train `Tib-Bi-Dharma`, a dedicated sentence transformer for Classical Tibetan. Our contributions are **threefold**:

1. We release `Tib-Bi-Dharma`, a new foundational text-similarity embedding model that unlocks a high-dimensional semantic search and new possibilities for computational study of the vast Tibetan literary heritage.
2. We release comprehensive Classical Tibetan sentence-pair annotation datasets, comprising both a large-scale synthetically scored corpus and a novel expert-annotated gold-standard dataset, providing a crucial benchmark for future text-similarity research.
3. We introduce an end-to-end pipeline that dynamically mines candidate text pairs and automatically annotates them, serving as a methodological framework for overcoming data scarcity in low-resource languages.

2 Related Work

The paradigm for large-scale semantic similarity was established by Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which maps sentences independently to dense vector spaces to bypass the quadratic computational bottleneck of standard cross-encoder architectures (Devlin et al., 2019). Subsequent work, such as SimCSE (Gao et al., 2021), further improved embedding quality by introducing advanced contrastive learning objectives, notably by reformulating Natural Language Inference (NLI) datasets into discrete triplets. However, because these discrete classification objectives push pairs either together or apart, they can introduce a discrepancy between the training loss and continuous evaluation metrics. To resolve this, CoSENT (Huang et al., 2024) utilizes a ranking-based loss to handle continuous similarity labels, ensuring the embedding space accurately reflects fine-grained semantic differences.

While highly effective, these methods fundamentally rely on massive, human-labeled training datasets. For low-resource domains like Classical Tibetan, this requirement is often insurmountable. A common alternative is the use of large-scale multilingual models, such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020). Although these models exhibit impressive cross-lingual transfer, their representation spaces are diluted across over 100 languages, causing them to underperform compared to dedicated monolingual models.

To address general data scarcity, Thakur et al. (2021) proposed Augmented SBERT, utilizing a cross-encoder for labeling and Kernel Density Estimation (KDE) for pair sampling. We follow the spirit of this approach in our zero-resource setting by substituting the standard cross-encoder with either a multi-model representation ensemble or a generative LLM-as-a-judge committee for automated data labeling (Section 3.2.3). Furthermore, we replace static KDE with an iterative, active semantic mining loop for dynamic pair sampling (Section 3.2.2).

In recent years, significant progress has been made in Tibetan NLP (Huang et al., 2025; Meelen et al., 2024). Foundational models, most notably the monolingual TiBERT (Liu et al., 2022) and Tibetan-BERT-wwm (Liang et al., 2024), have been effective for tasks like text classification and sentence boundary disambiguation. Furthermore, multilingual models like CINO (Yang et al., 2022) have shown strong performance across the broader Sino-Tibetan language family.

Despite this progress, a critical research gap remains: the lack of a dedicated, high-performance sentence-transformer model optimized for Classical Tibetan. While Modern and Classical Tibetan share the same script, the archaic vocabulary and the distinct morphology of historical texts render modern-trained models suboptimal. Furthermore, while models like TiBERT offer strong general-purpose language capabilities, they are not calibrated for sentence-level semantic similarity out-of-the-box. Consequently, applications like semantic search or clustering, even for Modern Tibetan (Engels and Barnett, 2025), currently must rely on unoptimized multilingual baselines. This paper addresses this gap with `Tib-Bi-Dharma`. By integrating state-of-the-art contrastive bootstrapping, synthetic data generation, and continuous scaling objectives, our approach provides a holistic, empirically validated solution for Classical Tibetan

sentence embeddings.

3 Methodology

Training a high-quality text similarity model requires a large number of semantically aligned text pairs. In the context of historical corpora, such pairs typically take the form of textual parallels: passages that convey equivalent or closely related meanings, even when expressed through paraphrase, condensation, expansion, or stylistic variation. Unlike surface-level duplication, these parallels reflect deeper semantic correspondence and are essential for modeling meaning-preserving variation. For a sentence transformer to learn a meaningful embedding space, it must observe both positive examples of such semantic equivalence and contrasting negative examples that differ in meaning. Constructing this type of parallel data at scale is therefore central to effective similarity learning.

However, in Classical Tibetan, such parallel data is extremely scarce and costly to annotate manually. To address this limitation, we adopt a two-tiered strategy. First, we curate a high-quality "gold-standard" dataset with human experts, primarily serving as a robust evaluation benchmark. Second, to address the scalability limitations of manual annotation, we propose and evaluate a pipeline for automatically annotating text parallels at scale, and use the resulting data to train and evaluate text-similarity models.

Although Classical Tibetan uses its own script, many academic and computational resources adopt the Wylie transliteration scheme (Wylie, 1959), which represents Tibetan characters with the Latin alphabet. We use Wylie for all experiments unless otherwise stated.

3.1 Gold-Standard Dataset of Text Parallels

Seven scholars of Tibetan Studies compiled the set, which comprises hand-picked pairs drawn from known Buddhist corpora, including the *bKa' 'gyur*, *bsTan 'gyur*, *rNying ma rgyud 'bum*, as well as *Rong zom gsung 'bum*. To ensure semantic diversity and strict traceability, each candidate pair picked by the team was pre-categorized into four similarity levels and recorded on a corresponding spreadsheet. To maintain rigorous data provenance, we explicitly logged both the identity of the scholar who originally sourced each candidate pair and the annotator who provided the final evaluation. Once there were enough pairs in each category pool, we selected 1,000 pairs with a balanced class distribution to facilitate unbiased relative scoring.

These 1,000 pairs were distributed among the seven scholars and annotated in four discrete batches: an initial warm-up batch of 100 pairs, followed by three subsequent batches of 300 pairs each. To mitigate the subjective biases inherent in absolute rating scales, we employed Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991; Louviere et al., 2015), implemented via an established BWS framework (Saif M. Mohammad). During the task, annotators were presented with randomly sampled 4-tuples of Classical Tibetan sentence pairs and instructed to isolate exactly one "most similar" (best) and one "least similar" (worst) pair per tuple (see Table 4 in Appendix A for examples). To maintain strict annotation quality, any highly ambiguous or challenging comparisons were escalated to a senior annotator for guidance and final adjudication. To convert these discrete categorical selections into a continuous similarity score S_i for each sentence pair i, we applied the standard BWS scoring formula: $S_i = \frac{B_i - W_i}{N_i}$ where B_i and W_i represent the total number of times pair i was selected as the most and least similar, respectively, and N_i denotes the total number of times the pair appeared across all evaluated tuples. This rigorous comparative approach, combined with senior oversight, allowed us to compute highly reliable continuous similarity scores for the final publicly released gold-standard dataset[2]. Table 5 (Appendix B) provides examples of these final scored pairs.

3.2 Text-Similarity Model Creation

We describe a process for building a Classical Tibetan text-similarity model, in which candidate training pairs are iteratively retrieved from the raw corpus and annotated through active learning.

3.2.1 Corpus Preprocessing and Segment Pool Generation

Our raw corpus consists of the Kangyur and Tangyur collections[3]. In order to generate text segments from the continuous, space-less Classical Tibetan script, we apply a custom segmentation algorithm that produces segments of 10–30 words. We use the first two million segments (out

[2] https://huggingface.co/datasets/Intellexus/tibetan-sts-gold

[3] Sourced from https://github.com/Esukhia/derge-kangyur and https://github.com/Esukhia/derge-tengyur

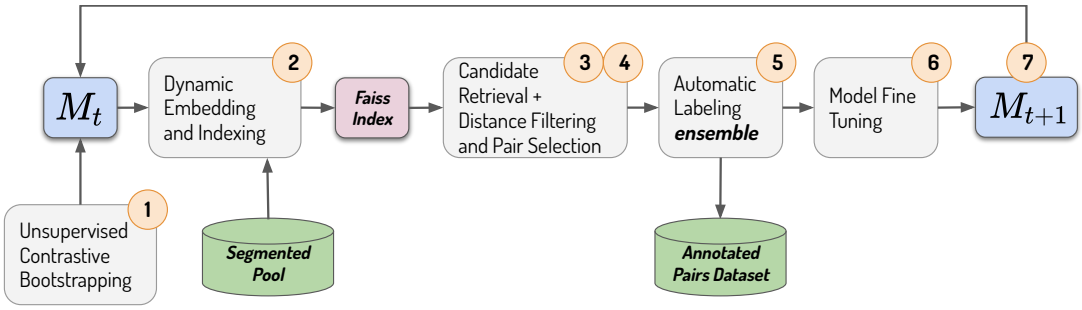


Figure 1: End-to-end active learning pipeline for candidate pair mining. M_t represent the model at the i'th iteration.

of 2,823,001) as our primary training pool. Full preprocessing and segmentation details are provided in Appendices H and I.

3.2.2 Automatic Annotation of Segment Pairs

Training a text-similarity model in a contrastive setting requires not only semantically aligned positive pairs, but also informative hard negatives to provide a robust learning signal (Xiong et al., 2021). Our preliminary experiments show that without explicit embedding-oriented fine-tuning, a foundation model exhibits severe anisotropy, consistent with prior findings on representation degeneration (Ethayarajh, 2019; Gao et al., 2019). As reported by Li et al. (2020), text-segment vectors collapse into a narrow cone in the embedding space, and we observe the same phenomenon: cosine similarities from the uncalibrated model are uniformly high regardless of true semantic overlap.

Our objective is to extract semantically related pairs from this vast, unannotated pool, allowing the model to learn an embedding space in which meaning-preserving variations yield high cosine similarities. To overcome this initial anisotropy and accomplish our objective, we designed a pipeline (illustrated in Figure 1) that actively mines candidate segment pairs based on the model's evolving similarity estimates.

1. **Unsupervised Contrastive Bootstrapping:** We start with a base model. As a preliminary step to mitigate anisotropy, we bootstrap the model by fine-tuning it with unsupervised contrastive learning over the entire pool of segments. Specifically, we train the model for three epochs by treating the same segment as anchor-positive alignment pairs and optimizing with the Multiple Negatives Ranking Loss, where the negatives consist of the other examples in the same batch, following the SimCSE framework. This effectively forces the collapsed vector space to expand and disperse prior to our mining phase.
2. **Dynamic Embedding and Indexing:** The current iteration of the sentence embedding model encodes all remaining, unused textual segments in the training pool into dense vector representations. The newly generated embeddings are then ingested and indexed using `Faiss` (Douze et al., 2024) to facilitate rapid, high-dimensional similarity operations.
3. **Candidate Retrieval:** To optimize computational efficiency, we randomly sample a batch of 250 segments from the training pool simultaneously, which we use as queries. Using their dense vector representations, we perform a batched query against the `Faiss` index to retrieve the top-k ($k = 1,000$) most semantically similar candidate segments based on cosine distance.
4. **Distance Filtering and Pair Selection:** For each query, we randomly select a target segment from its top-k neighbors, enforcing a strict cosine distance threshold of ≥ 0.6 to guarantee the extraction of semi-hard candidates with high information gain. Selected pairs are permanently removed from the pool to prevent redundancy. If filtering yields fewer than 250 valid pairs, we iteratively repeat Step 3 until the target batch size is reached or the pool is exhausted. In the unlikely event that no valid pairs can be extracted, the pipeline halts and requires manual recalibration of the initial threshold settings.
5. **Automatic Labeling:** Extracted pairs are automatically labeled using a strategy selected from Section 3.2.3 and kept fixed for the entire pipeline run. We experiment with two strategies and report their performance.
6. **Model Fine-Tuning:** The sentence embedding model is contrastively trained using the CoSENT loss on this newly acquired batch of

automatically annotated pairs, effectively updating its weights and refining its spatial alignment.

7. **Iteration:** The loop restarts at Step 2. The newly updated model generates fresh embeddings for the remaining pool, ensuring that the next batch of mined pairs is selected based on an incrementally improved vector space.

To facilitate future research, we publicly release the code for our pipeline framework[4]. Using this framework across multiple iterations, we generated automatically annotated pairs and aggregated them into a final, comprehensive dataset to train our downstream text-similarity models.

3.2.3 Automatic Annotation of Text Parallels

In Step 5, the extracted segment pairs are automatically annotated. Here, we explore and evaluate two different automated annotation approaches. The first leverages the capabilities of existing text-similarity models (text embeddings and re-rankers), while the second leverages generative large language models (LLMs) instructed to emulate the human expert annotation process using BWS.

Approach 1: Sentence Embeddings and Re-ranking Models (EMR) In our first approach, we evaluated a range of pre-trained sentence embedding models and encoder–decoder re-ranking models (Guo et al., 2016) to assess their ability to capture semantic similarity in parallel texts. Sentence embedding models encode each sentence independently into a fixed-dimensional vector space, with similarity computed via cosine similarity, whereas encoder–decoder re-rankers jointly process both sentences as a concatenated input and directly predict a similarity score. To leverage complementary strengths, we construct a late-fusion ensemble in which the final similarity score is computed as a weighted average of the top-performing models' outputs.

We begin with a zero-shot evaluation of several state-of-the-art embedding models and encoder–decoder re-rankers on our manually curated gold-standard dataset of Classical Tibetan sentence pairs. Performance is measured using standard correlation metrics, Spearman and Pearson, to assess alignment between model predictions and human similarity judgments. Full results are provided in Table 6 (Appendix C). Based on this evaluation, we select the three strongest models: BGE[5] (Chen et al., 2024b; Xiao et al., 2024), GemmaEmbedding-300m[6] (Schechter Vera et al., 2025), and Qwen3-Reranker-8B[7] (Zhang et al., 2025). The first two are embedding-based models, while the third is a re-ranking architecture. For the ensemble, we experiment with different weight combinations (listed in Table 7 in Appendix D) and select a balanced configuration that achieves nearly the highest Spearman correlation. Table 1 summarizes the performance of the three individual models and their combined ensemble.

Model	Spearman (ρ)	Pearson (r)
BGE	0.809	0.806
Qwen3-Reranker-8B	0.809	0.753
GemmaEmbedding-300m	0.785	0.782
Ensemble	**0.839**	0.827

Table 1: Zero-shot evaluation of text-similarity models.

Approach 2: Generative LLMs Using BWS As an alternative to conventional embedding-based similarity, we leverage chat LLMs by framing them as a committee of expert annotators, an approach that has shown strong effectiveness in automating complex annotation tasks (Bagdon et al., 2024; Zhang and Feng, 2025). We employ the same BWS protocol used in our human expert annotation process (Section 3.1). To construct the ensemble, we adopt an early-fusion strategy: instead of computing separate BWS scores per model and averaging them, we merge the raw categorical selections ("best" and "worst" for each 4-tuple) produced by the LLMs into a single annotation pool. The BWS standardization procedure is then applied to this combined set of judgments, effectively simulating a multi-annotator setting.

To identify effective models for this task, we benchmark a diverse set of state-of-the-art generative LLMs using the BWS protocol against our gold-standard dataset (full results in Table 8, Appendix E). To keep the pipeline financially scalable, we focus on lightweight, cost-effective variants. Our final committee therefore includes two high-performing instruction-tuned models from different vendors: `Gemini 3 Flash (preview)` (Google DeepMind, 2026a) and `Claude 4.5 Haiku` (Anthropic, 2025). Table 2 reports the correlation of

[4] https://github.com/Intellexus-DSI/sensim

[5] BAAI/bge-multilingual-gemma2

[6] google/embeddinggemma-300m

[7] Qwen/Qwen3-Reranker-8B

each model, and their ensemble, with human expert annotations on the gold-standard dataset.

Model	Spearman (ρ)	Pearson (r)
Gemini 3 Flash	0.843	0.837
Claude 4.5 Haiku	0.811	0.809
Ensemble	**0.847**	0.844

Table 2: Evaluation results of LLMs using BWS.

As demonstrated in both Table 1 and Table 2, both ensembling techniques consistently outperform their respective individual models.

3.2.4 Training Text-Similarity Models

Using the automatically annotated data produced by the pipeline described above, we train two encoder-based language models: BERT and ModernBERT. Prior to contrastive learning for text similarity, both models undergo continual pretraining using the standard masked-language modeling (MLM) loss on a large corpus of Wylie-transliterated Classical Tibetan to better capture the language's lexical and syntactic properties. The resulting models are referred to as Tib-B (BERT-based) and Tib-MB (ModernBERT-based). Additional details are provided in Appendix J.

4 Experiments and Results

4.1 Experimental Setup

We first run our annotation pipeline for 20 iterations for each supervision type to generate two collections of labeled segment pairs: one using EMR supervision (3.2.3) and another using LLM-based BWS supervision (3.2.3). In Step 1 of the annotation pipeline, we use Tib-B as the foundation model to select candidate pairs from the training pool. For the BWS approach, we employ a $4N$ 4-tuple strategy (N denoting the total number of sentence pairs), intentionally exceeding the standard $1.5N$–$2N$ baseline as Kiritchenko and Mohammad (2017) show that further oversampling improves annotation quality.

Before training Tib-B and Tib-MB on the automatically annotated pairs, we perform a preliminary training step to adapt the model for text similarity. We experiment with two approaches for this initialization step: (1) **Anchor–Positive (AP):** Following Gao et al. (2021), identical segments with dropout noise are used to form anchor–positive pairs. We use this standard setup as a reference baseline for the bootstrapped model described in Step 1 of the annotation pipeline. (2) **Consecutive Segments (CS):** To capture narrative continuity and discourse flow, we construct positive pairs by coupling adjacent text segments from the unannotated corpus (i.e., treating segment t_{i+1} as the positive target for t_i). This approach follows self-supervised methods that exploit document-level context by treating neighboring sentences as semantically related (Logeswaran and Lee, 2018; Giorgi et al., 2021).

Both initialization approaches are optimized using the multiple-negatives-ranking loss (Henderson et al., 2017), which combats initial representation degeneration by pulling positive pairs together and treating all other in-batch segments as hard negatives.

Finally, the models are fine-tuned for seven epochs using the CoSENT objective (Huang et al., 2024), chosen because it naturally handles our continuous scores while maintaining robust contrastive alignment. Training utilized the AdamW optimizer (learning rate 2e-5, temperature $\lambda = 20$) on NVIDIA RTX 3090/GB10 GPUs (see Appendix K for the full hyperparameter configuration).

4.2 Evaluation Metrics

Following standard practices for Semantic Textual Similarity (STS) tasks (Reimers and Gurevych, 2019), we compute the cosine similarity between sentence embeddings. Model performance is then evaluated using Spearman's rank correlation coefficient (ρ) between these predicted similarities and the human judgments in our novel Classical Tibetan gold dataset (Section 3). Additional details regarding the STS setup are provided in Appendix F. To ensure robust evaluation, all experiments employ 4-fold cross-validation to generate the training, validation, and test splits. Appendix G provides further details and the data distributions in Table 9.

4.3 Comparison with Tibetan Foundation Models

We compare our models to several state-of-the-art Tibetan language models: LaBSE (Feng et al., 2022), TiBERT (Liu et al., 2022), CINO (Yang et al., 2022), and T-RoBERTa[8] . Since these models are trained on native Unicode Tibetan script, we convert our data from Wylie transliteration to Unicode before evaluation, allowing them to fully utilize their Tibetan tokenizers and vocabularies.

[8] https://huggingface.co/sangjeedondrub/tibetan-roberta-base

Model	Baseline	Gold	EMR	BWS	EMR →Gold	BWS →Gold	EMR+BWS →Gold
LaBSE	0.803±.02	0.831±.03	0.830±.02	0.840±.03	0.836±.02	0.845±.02	0.844±.02
CINO	0.615±.01	0.766±.02	0.771±.01	0.763±.02	0.793±.02	0.788±.02	0.805±.02
T-RoBERTa	0.778±.03	0.798±.03	0.804±.03	0.795±.04	0.806±.03	0.804±.04	0.810±.03
TiBERT	0.691±.03	0.747±.02	0.755±.04	0.738±.03	0.770±.03	0.773±.04	0.781±.04
Tib-B	0.795±.02	0.846±.01	0.839±.02	0.843±.02	0.853±.02	0.855±.02	0.856±.02
Tib-B-AP	0.807±.02	0.833±.01	0.827±.02	0.838±.02	0.844±.02	0.842±.02	0.853±.01
Tib-B-CS	0.801±.01	0.844±.01	**0.845**±.01	0.847±.02	0.849±.01	0.850±.02	0.852±.02
Tib-MB	0.702±.02	0.832±.02	0.824±.02	0.827±.03	0.842±.02	0.842±.03	0.848±.02
Tib-MB-AP	0.689±.04	0.754±.04	0.797±.02	0.792±.03	0.811±.02	0.815±.02	0.823±.01
Tib-MB-CS	**0.808**±.01	**0.851**±.01	0.841±.02	**0.850**±.02	**0.859**±.01	**0.864**±.02	**0.858**±.01

Table 3: Spearman (ρ) correlation results for all evaluated models. **AP** denotes initialization using anchor-positive pairs with multiple-negatives-ranking loss, while **CS** denotes initialization using consecutive segments. The model Tib-B-AP was used as the unsupervised bootstrap model in the automatic annotation pipeline.

4.4 Data Configurations

We train the models with contrastive learning on different data sources and combinations: (1) **Baseline:** no training; (2) **Gold**[9]**:** training on the gold dataset only; (3) **EMR**[10]**:** EMR-annotated pairs; (4) **BWS**[11]**:** LLM-based BWS-annotated pairs; (5) **EMR→Gold:** EMR followed by Gold; (6) **BWS→Gold:** BWS followed by Gold; and (7) **EMR +BWS→Gold:** EMR+BWS followed by Gold.

4.5 Results and Discussion

Unsupervised Bootstrapping Efficacy. As shown in Table 3, unsupervised initialization substantially improves baseline performance. The CS strategy consistently outperforms AP, raising the Tib-MB Spearman correlation from 0.702 to 0.808 and indicating that adjacent segments provide a stronger signal of semantic continuity.

The Advantage of Prior Sentence Alignment. As a multilingual sentence embedder, LaBSE shows a strong zero-shot advantage, reaching a baseline Spearman correlation of 0.803. Under the BWS→Gold regime, it achieves 0.845, indicating that cross-lingual sentence spaces adapt well to Classical Tibetan.

Automatic vs. Gold Training. Training only on the BWS dataset yields performance comparable to the human-annotated Gold set. For instance, Tib-MB-CS scores 0.850 on BWS versus 0.851 on Gold, while LaBSE and Tib-B-CS perform slightly better on BWS. This suggests that the scale and diversity of BWS effectively compensate for the absence of expert annotations.

Effective Knowledge Transfer from the Ensemble A notable result from the standalone automatic training experiments is the effectiveness of knowledge transfer. Trained only on EMR or BWS data, and without any Gold labels, our lightweight Tib-MB-CS model achieves Spearman correlations of 0.841 and 0.850, surpassing the zero-shot performance of the larger ensemble used to generate the labels.

Two-Stage Training Yields Best Performance Table 3 shows that two-stage training consistently outperforms single-dataset training. The best configuration, BWS→Gold, first trains on the large automatic pool and then refines on Gold, yielding a peak Spearman correlation of 0.864 for Tib-MB-CS. We release Tib-Bi-Dharma[12] with this top configuration, demonstrating that a lightweight model trained with automatic and gold data can outperform large multilingual ensembles.

4.6 Qualitative Analysis of the Automatic Annotation Pipeline

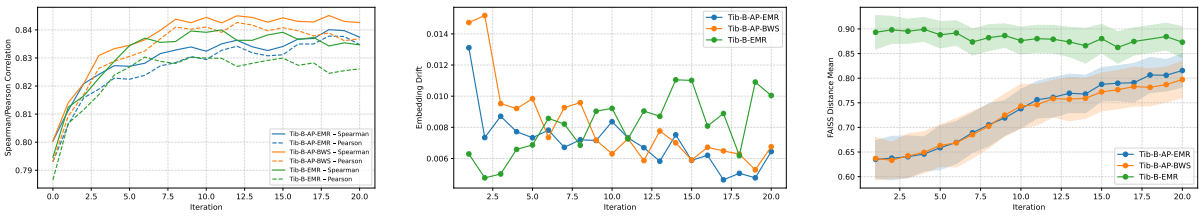
To analyze how the embedding space evolves across iterations in the annotation pipeline, we track three metrics over 20 training iterations. Figure 2 shows the trajectories of downstream test performance (Spearman and Pearson correlations), embedding-space drift of the first 10^5 segments in the training pool, and candidate retrieval dynamics for both EMR and BWS supervisions. We explain and

[9] https://huggingface.co/datasets/Intellexus/tibetan-sts-gold

[10] https://huggingface.co/datasets/Intellexus/tibetan-sts-synthetic-emr

[11] https://huggingface.co/datasets/Intellexus/tibetan-sts-synthetic-bws

[12] https://huggingface.co/Intellexus/Tib-Bi-Dharma-Tibetan-EWTS-v2



(a) Downstream STS Correlations (b) Embedding Drift (c) Mean `Faiss` Retrieval Distance

Figure 2: STS performance correlation (a), embedding drift (b), and mean `Faiss` retrieval distance (c) across 20 iterations of the annotation pipeline for different models.

analyze each metric below. To measure the effect of the AP initialization on the `Tib-B` model, we compare the initialized and uninitialized versions across all three metrics. For brevity, the uninitialized baseline shown is the model trained with `EMR`.

Performance Trajectory and Convergence Figure 2a shows Spearman and Pearson correlations on the test set across iterations. Performance improves rapidly in the early stages, indicating that the first batches of automatically annotated pairs effectively establish the basic semantic structure. After iterations 10–15, the metrics plateau, suggesting the model has largely captured the available signal and that additional pairs yield diminishing returns.

Embedding Drift Analysis We track embedding drift (Figure 2b), defined as the average cosine distance between a segment's embedding at iteration t and $t-1$. For the initialized models (`Tib-B-AP-EMR` and `Tib-B-AP-BWS`), drift is highest in the early iterations, reflecting rapid reorganization after unsupervised bootstrapping, and gradually stabilizes as the pipeline adds semantic supervision. In contrast, the uninitialized model (`Tib-B-EMR`) starts from a collapsed anisotropic space and requires more iterations to separate the representations. By iteration 20, the initialized models reach a stable lower drift, indicating that the learning loop has largely stabilized the embedding space.

Retrieval Dynamics and Candidate Hardness Figure 2c tracks the mean and standard deviation of `Faiss` cosine distances for retrieved candidate pairs, where the mean distance reflects candidate hardness. In the initialized models, the initial dispersion yields high retrieval distances that gradually decrease as the model learns to bring semantically related segments closer together. In contrast, the uninitialized model (`Tib-B-EMR`) begins in a collapsed anisotropic space with artificially low distances and exhibits the opposite trend, as the iterative process progressively separates the representations. Despite these different starting points, the bootstrapped models consistently operate within the 0.60–0.80 range, an effective regime for mining informative training pairs (Xiong et al., 2021).

5 Conclusion and Future Work

In this work, we introduced `Tib-Bi-Dharma`, significantly improving semantic similarity modeling for Classical Tibetan and enabling reliable semantic search across Tibetan Buddhist literature. By validating a fully automated annotation framework, we demonstrate that the long-standing expert annotation bottleneck can be mitigated through automatic supervision: training solely on automatically annotated data matches gold-standard performance, and combining both yields a peak correlation of 0.864. Our approach allows scholars to trace conceptual development and textual reuse across centuries of texts. Beyond Classical Tibetan, the methodology provides a general blueprint for building high-quality retrieval systems in other low-resource cultural heritage domains. Our work opens several directions for future research. First, we plan to move from sentence-level alignment to corpus-scale discovery, using `Tib-Bi-Dharma` to detect intertextual parallels across the entirety of Tibetan Buddhist literature (including both allochthonous and autochthonous). This will enable large-scale analysis of conceptual transmission and textual reuse. We also aim to extend the embedding space cross-lingually by linking Tibetan passages to their Sanskrit sources and applying the pipeline to other low-resource textual traditions. Finally, we will refine the annotation pipeline with improved acquisition strategies that target uncertain or contested sentence pairs, further reducing annotation effort.

Limitations

This work has several limitations. First, the automatic annotation pipeline introduces noise into the training data, as automatically labeled parallels may include semantically weak or incorrect pairs despite our ranking and BWS-based scoring strategies. Second, the method relies on the availability of large unannotated corpora from which candidate pairs can be extracted, an assumption that may not hold for languages with extremely limited textual resources. Third, because candidate parallels are extracted from the same corpus, the learned representations may partly reflect corpus-specific stylistic or topical patterns rather than general semantic equivalence.

Acknowledgments

This study is supported in part by the European Research Council (Intellexus, Project No. 101118558). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them.

References

Anthropic. 2025. Claude 4.5 haiku (version 20251001). Large Language Model, API model string: `claude-haiku-4-5-20251001`. Accessed: 2026-02-22 to 2026-03-03.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

James Engels and Robert Barnett. 2025. Developing a semantic search engine for modern tibetan. *Revue d'Etudes Tibétaines*, 74:262–283.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 879–895, Online. Association for Computational Linguistics.

Google DeepMind. 2025a. Gemini 2.5 flash. Large Language Model, API model string: `gemini-2.5-flash`. Accessed: 2026-02-22 to 2026-03-03.

Google DeepMind. 2025b. Gemini 2.5 flash-lite. Large Language Model, API model string: `gemini-2.5-flash-lite`. Accessed: 2026-02-22 to 2026-03-03.

Google DeepMind. 2026a. Gemini 3 flash (preview). Large Language Model, API model string: `gemini-3-flash-preview`. Accessed: 2026-02-22 to 2026-03-03.

Google DeepMind. 2026b. Gemini 3.1 pro (preview). Large Language Model, API model string: `gemini-3.1-pro-preview`. Accessed: 2026-02-22 to 2026-03-03.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM'16, page 55–64. ACM.

Kai Golan Hashiloni, Shay Cohen, Asaf Shina, Jingyi Yang, Orr Meir Zwebner, Nicola Bajetta, Guy Bilitski, Rebecca Sundén, Guy Maduel, Ryan Conlon, Ari Barzilai, Daniel Mass, Shanshan Jia, Aviv Naaman, Sonam Choden, Sonam Jamtsho, Yadi Qu, Harunaga Isaacson, Dorji Wangchuk, and 3 others. 2025. DharmaBench: Evaluating language models on buddhist texts in Sanskrit and Tibetan. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2088–2110, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *Preprint*, arXiv:2501.01028.

Cheng Huang, Nyima Tashi, Fan Gao, Yutong Liu, Jiahao Li, Hao Tian, Siyang Jiang, Thupten Tsering, Ban Ma-bao, Renzeg Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Jin Zhang, Xiao Feng, Hao Wang, Jie Tang, Guojie Tang, Xiangxiang Wang, and 3 others. 2025. Tibetan language and ai: A comprehensive survey of resources, methods and challenges. *Preprint*, arXiv:2510.19144.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. 2024. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *Preprint*, arXiv:2312.15503.

Yatao Liang, Hui Lv, Yan Li, La Duo, Chuanyi Liu, and Qingguo Zhou. 2024. Tibetan-bert-wwm: A tibetan pretrained model with whole word masking for text classification. *IEEE Transactions on Computational Social Systems*, 11(5):6268–6277.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, working paper.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Marieke Meelen, Sebastian Nehrdich, and Kurt Keutzer. 2024. Breakthroughs in tibetan nlp; digital humanities. *Revue d'Etudes Tibétaines*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*,

pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5 mini. Large Language Model, API model string: `gpt-5-mini-2025-08-07`. Accessed: 2026-02-22 to 2026-03-03.

OpenPecha. a. Botok: a powerful python library for tokenizing tibetan text. GitHub repository.

OpenPecha. b. pyewts: Python tibetan unicode to wylie (ewts) converter. GitHub repository.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Saif M. Mohammad. Best-worst-scaling-scripts: Code to assist with best-worst-scaling annotations.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Octen Team. 2025. Octen series: Optimizing embedding models to 1 on rteb leaderboard.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Turrell V Wylie. 1959. A standard system of tibetan transcription. *Harvard Journal of Asiatic Studies*, 22:261–267.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peng Yu, En Xu, Bin Chen, Haibiao Chen, and Yinfei Xu. 2025. Qzhou-embedding technical report. *Preprint*, arXiv:2508.21632.

Mengchen Zhang and Xiang Feng. 2025. Automated annotation of academic emotion intensity in online learning comment texts: A bws method based on llms. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 317–323, New York, NY, USA. Association for Computing Machinery.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2025. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *Preprint*, arXiv:2506.20923.

A Gold BWS 4-tuple pairs example

This table presents five examples of 4-tuples used during the annotation process. For each tuple of four sentence pairs, the annotator (whether human or synthetic) must select the "most similar" (best) and "least similar" (worst) pair.

Ex.	Pair	Sentence A	Sentence B
1	1	slob ma ni dad pa che zhing sgrub pa la dga' ba/ gsang sngags la mos shing dbang don du gnyer pa/	slob ma dad ldan sgrub la dga'// rtsa ba'i sngags bzlos byed pa dang // gsang sngags rgyud la dga' sems shing // gzhan yang phyi ma'i don 'dod pa'o//
	2	sgyu ma'i skyes bus sgyu ma'i skyes bu dag la chos bstan pa bzhin no/	sems rgyun yan lag can dang mtshungs//
	3	de la ni chos rdul tsam yang mi dmigs so//	de ni nam mkha'i mtshan nyid do//
	4	de ni bcad par gyur pa yin zhes bstan// gang du che ba bzhi yis btab nas su// sbrul gyis ni zlog par gyur te// bstan nas bsad pa yin zhes bshad pa ste//	de ni bcad par gyur pa yin zhes bstan// gang du mche ba bzhi yis btab nas su// sbrul gyi lus ni zos par gyur pa de// btab nas zos pa yin zhes bshad pa ste//
2	1	nam mkha'i mtshan nyid snga rol na// nam mkha' cung zad yod ma yin// gal te mtshan las snga gyur na// mtshan nyid med par thal bar 'gyur// mtshan nyid med pa'i dngos po ni// 'ga' yang gang na'ang yod ma yin//	bdag dang bdag gir 'dzin byas pa// 'dus byas spyod yul can chags ni// rgyu yin de yi gnod byed ni// bdag med mthong ba 'gal ba yin//
	2	de'i dngos po'i ngo bo nyid rnam par brtags na rdul phra mo tsam gyi chos rnams kyang dmigs su med do//	gang tshe 'dus byas 'dus ma byas dang dkar nag chos// shes rab rnam par bshig nas rdul tsam mi dmigs tshe// 'jig rten dag na shes rab pha rol phyin grangs 'gro// nam mkha' gang la'ang cung zad mi gnas de dang 'dra//
	3	de bzhin du rig byed smra ba la sogs pa ltar shes pa po ni bdag yin no snyam du rtog cing / rnam par shes pa'i ngo bo nyid du bzung ste/	bdag tu 'dzin pa'i rnam shes ni// bdag las gal te skye 'gyur na// des na kun bdag nyid rnam shes//
	4	sems ni bza' btung tshogs la chags mi bya//	bza' btung sogs la sems ni chags mi bya//
3	1	sems ni sgyu ma'i rnam pa ste// byang chub kyang ni sgyu ma 'dra//	mya ngan las 'das pa ni mya ngan las 'das pas stong ngo//
	2	de ltar sangs rgyas bsam mi khyab// sangs rgyas chos kyang de bzhin te// bsam mi khyab la dad pa yi// rnam par smin pa'ang de bzhin no//	sangs rgyas bcom ldan yon tan bsam mi khyab// dam pa'i chos kyi yon tan bsam mi khyab// 'phags pa'i dge 'dun yon tan bsam mi khyab// dkon mchog bsam mi khyab la mngon dad pa'i// rnam smin bla na med pa'ang bsam mi khyab//
	3	de bzhin byang chub sems ni ma bskyed cing // sangs rgyas ma dad lha gzhan la brten pas// nga yi gsang sngags bzlas na phung bar 'gyur//	de dag nga yi dkyil 'khor gzhug / drag po g.yo dang mi g.yo ba// drag tu 'khrul cing sems can rnams// nga yi dkyil 'khor nang du gzhug /de mthong sems can 'chi bar 'gyur//
	4	chos thams cad ni sems tsam ste/ sems dang mtshungs par ldan pa yin no//	sems nyid kyis ni sems can 'jig rten dang // snod kyi 'jig rten shin tu sna tshogs 'god// 'gro ba ma lus las las skyes par gsungs// sems spangs nas ni las kyang yod ma yin//
4	1	sring mo de ltar na byang chub dang mig gi rnam par shes pa'i khams de ni gnyis su med de gnyis su dbyer med do//	sring mo de ltar na byang chub dang mig gi rnam shes kyi khams de ni gnyis su med de gnyis su dbyer med do//
	2	dkyil 'khor chen po bshad du gsol//	gzugs brnyan dkyil 'khor bri bar bya//
	3	yan lag bcu gnyis yod ma yin// mtha' dang mtha' med pa yang med// lta ba thams cad spang ba'i phyir// sems tsam du ni ngas bshad do//	gang gis thugs brtse nyer bzung nas// lta ba thams cad spang ba'i phyir// dam pa'i chos ni ston mdzad pa// gau tam de la phyag 'tshal lo//
	4	dka' thub nges par mi bzad pas// bsten pas 'grub par mi gyur gyi//	dka' thub sdom pa mi bzad pa// brten pas 'grub par mi 'gyur te//
5	1	dod chags dang / zhe sdang dang / gti mug gang yin pa ste/ de dag ni/ spang bar bya ba yin pas yongs su spangs pa la rab tu brtson zhing bslab par bya'o//	gti mug chags dang zhe sdang dang / nga rgyal phrag dog mi spang do/
	2	nyon mongs de dag gang gi yin// de yang grub pa yod ma yin// gal te gang med ci zhig yod// nyon mongs cung zad yod ma yin//	nyon mongs de dag gang gi yin// de yang grub pa yod ma yin// 'ga' med par ni gang gi yang // nyon mongs pa dag yod ma yin//
	3	sdang ba'i dgra bo'i srog rtsa gcod par mdzod//	dam nyams dgra bo'i srog rtsa gcod mdzad pa'i//
	4	dbang ma thob pas mi rig par// sgrub pos 'khor lo de bya'o//	sangs rgyas sprin rnams de yi tshe// de la dbang bskur rab tu rtsol//

Table 4: Examples of 4-tuples evaluated during annotation.

B Gold BWS scored pairs example

This table provides five examples of sentence pairs along with their normalized BWS similarity scores.

Ex.	Sentence A	Sentence B	Score
1	de lta bas na lta ba mtho dman ni snang ba la dngos por zhen pa che chung gi bye brag tsam ste/	lta ba'i bye brag 'di dag kyang rdzas su yod med kyi gzhi las brtsams te/ lta ba'i dbye ba mdor bsdus te bstan pa'o//	0.75
2	byang chub tu sems bskyed pas khros pas gzhan dag 'tshogs pas sdom pa 'chor bar 'dus/	byang chub sems dpar sems bskyed nas dam chos 'drar snang ston pas spong /	0.5625
3	ci'i phyir 'byung ba'i bdag nyid lhan cig skyes pa'i ye shes su gyur// lhan cig skyes pa nyid kyis de brjod ces bya ba'i don to//	de dag rang gis ngo ma shes pa ni/ lhan cig skyes pa'i ma rig pa ste/ ma rig pa de 'khor ba'i gzhir rgyur gyur pa las 'byung khungs kyi sgra/ zhes bya'o//	0.375
4	dag pa dang ba 'od gsal ba// mi 'khrugs 'dus ma byas pa ni// byang chub sems dpa'i spyod yul lo//	sgyu ma'i rjes su mthun pa ni shin tu dang ba'i me long gi gzugs brnyan lta bu'o//	0.1875
5	gal te bdag dang sdug pa dang // rtag dang bde ba yod na ni// bdag shes sdug shes rtag shes dang // bde shes phyin ci log ma yin//	gal te bdag dang gtsang ba dang // rtag dang bde ba yod na ni// bdag dang gtsang dang rtag pa dang // bde ba phyin ci log ma yin//	0.875

Table 5: Sentence pairs alongside their computed normalized BWS scores.

C Sentence Embeddings and Re-ranking Models Evaluation

Table 6: Single-Model Evaluation Results against gold standard

Type	Model	Spearman (ρ)	Pearson (r)
embedding	`BAAI/bge-multilingual-gemma2` (Chen et al., 2024b; Xiao et al., 2024)	0.809	0.806
reranker	`Qwen/Qwen3-Reranker-8B` (Zhang et al., 2025)	0.809	0.753
embedding	`Octen/Octen-Embedding-8B` (Team, 2025)	0.792	0.800
embedding	`tencent/KaLM-Embedding-Gemma3-12B-2511` (Zhao et al., 2025; Hu et al., 2025)	0.792	0.789
embedding	`BAAI/bge-m3` (Chen et al., 2024b)	0.790	0.792
reranker	`BAAI/bge-reranker-v2-gemma` (Li et al., 2023; Chen et al., 2024a)	0.786	0.773
embedding	`Qwen/Qwen3-Embedding-8B` (Zhang et al., 2025)	0.786	0.796
embedding	`google/embeddinggemma-300m`[13]	0.785	0.782
embedding	`sentence-transformers/sentence-t5-xxl` (Ni et al., 2022)	0.754	0.738
embedding	`Kingsoft-LLM/QZhou-Embedding` (Yu et al., 2025)	0.716	0.701

D Sentence Embeddings and Re-ranking Models Ensemble Evaluation

Table 7: Multi-Model Ensemble Evaluation Results against the gold standard. The ensemble highlighted in **bold** was selected as our annotation ensemble for embedding models and rerankers.

Multi-Model Ensembles	Weights	Spearman (ρ)	Pearson (r)
`bge-multilingual-gemma2`	0.40	0.840	0.831
`embeddinggemma-300m`	0.20		
`bge-m3`	0.10		
`Qwen3-Reranker-8B`	0.30		
`bge-multilingual-gemma2`	0.33	**0.839**	0.827
`embeddinggemma-300m`	0.33		
`Qwen3-Reranker-8B`	0.34		
`bge-multilingual-gemma2`	0.30	0.839	0.822
`embeddinggemma-300m`	0.30		
`Qwen3-Reranker-8B`	0.40		
`bge-multilingual-gemma2`	0.30	0.839	0.822
`embeddinggemma-300m`	0.10		
`bge-m3`	0.20		
`Qwen3-Reranker-8B`	0.40		
`bge-multilingual-gemma2`	0.25	0.838	0.832
`embeddinggemma-300m`	0.25		
`bge-m3`	0.25		
`Qwen3-Reranker-8B`	0.25		
`bge-multilingual-gemma2`	0.33	0.830	0.800
`embeddinggemma-300m`	0.33		
`Qwen3-Reranker-4B`	0.34		
`bge-multilingual-gemma2`	0.25	0.828	0.802
`embeddinggemma-300m`	0.25		
`bge-m3`	0.25		
`Qwen3-Reranker-0.6B`	0.25		
`embeddinggemma-300m`	0.25	0.796	0.755
`bge-m3`	0.25		
`Qwen3-Reranker-0.6B`	0.25		
`bge-reranker-v2-gemma`	0.25		

E Generative LLMs Using BWS Evaluation

Guided by Hashiloni et al. (2025), Table 8 details our evaluation of the following generative LLMs:

- `Gemini 3 Flash` (Google DeepMind, 2026a)
- `Gemini 3.1 Pro` (Google DeepMind, 2026b)
- `Gemini 2.5 Flash` (Google DeepMind, 2025a)
- `Gemini 2.5 Flash-Lite` (Google DeepMind, 2025b)
- `Claude 4.5 Haiku` (Anthropic, 2025)
- `GPT-5 Mini` (OpenAI, 2025)

Table 8: Generative LLMs Using BWS evaluation against gold standard

Model	Spearman (ρ)	Pearson (r)
`Gemini 3 Flash`	0.843	0.837
`Gemini 3.1 Pro`	0.836	0.832
`Gemini 2.5 Flash-Lite`	0.796	0.794
`Gemini 2.5 Flash`	0.821	0.818
`Claude 4.5 Haiku`	0.811	0.809
`GPT-5 Mini`	0.794	0.789

F Adapting Annotations for Semantic Textual Similarity (STS)

To evaluate our automatically annotated datasets on the downstream STS task, the EMR and BWS scores must be mapped to a standard Semantic Textual Similarity (STS) scale.

Data Formulation A standard STS dataset consists of sequence pairs and a similarity score. We formulate our annotated dataset as a collection of tuples, $D = \{(x_i, y_i, s_i)\}$, where x_i and y_i represent the i-th pair of text segments (Sentence A and Sentence B), and the label s_i represents the similarity score.

Data Normalization While EMR produces continuous absolute similarity scores in the range $s_i^{\text{EMR}} \in [0, 1]$, BWS yields continuous relative scores in the range $s_i^{\text{BWS}} \in [-1, 1]$. We therefore normalize the BWS scores to a unified scale of $s_i \in [0, 1]$ to ensure comparability across annotation schemes prior to any usage. This normalization is applied to both the gold-standard and synthetic-standard datasets.

Bi-Encoder Inference During both evaluation and training, the language model operates as a strict bi-encoder. Let f_θ denote the sentence transformer model parameterized by θ. For a given pair, (x_i, y_i), the model independently encodes both segments, using a pooling method (as specified in Appendix K), and outputs dense vector representations: $\mathbf{u}_i = f_\theta(x_i)$ and $\mathbf{v}_i = f_\theta(y_i)$. The predicted semantic similarity between the two segments is then computed using cosine similarity: $\hat{s}_i = \cos(\mathbf{u}_i, \mathbf{v}_i)$.

Evaluation Metric The primary evaluation metric, as suggested by Reimers et al. (2016), is Spearman's rank correlation (ρ). To evaluate model performance on the held-out test set, we compute Spearman's rank correlation coefficient (ρ) between the gold-standard scores and the model predictions.

G Gold and Synthetic Sets Distribution

To ensure robust evaluation, the gold dataset was partitioned into four cascading folds with strictly non-overlapping test sets. To completely isolate the validation data across iterations, the validation set for each successive fold was explicitly sampled as a subset of the preceding fold's test set. This cascading structure guarantees that the validation sets are completely disjoint from one another and never overlap with the current test set.

Table 9: Annotated sets

Set	Size
Train	600
Validation	150
Test	250
Gold	1,000
BWS (Synthetic)	5,000
EMR (Synthetic)	5,000

H Corpus Preprocessing and Segmentation Pipeline

The text preprocessing pipeline is designed to transform continuous Classical Tibetan text—specifically in this study, the Kangyur and Tengyur corpora 3—into dynamically sized context windows that respect natural linguistic boundaries optimized for the embedding model. This is achieved through a sequential, three-step process: cleaning the raw input, parsing and aggregating it into semantically coherent spans bounded by strict length constraints, and finally converting those spans into the Extended Wylie Transliteration Scheme (EWTS).

H.1 Text Cleaning

The pipeline begins by standardizing the raw input text. Using regular expressions, the text is filtered

to retain only characters within the Tibetan Unicode block. All non-Tibetan characters, extraneous non-Tibetan punctuation, and Latin scripts are removed, and whitespace is standardized to create a clean baseline for grammatical analysis. Additionally, if the downstream architecture requires strictly contiguous character strings, the pipeline can be configured to completely strip all remaining intra-segment whitespace.

H.2 Segmentation and Span Generation

Once cleaned, the Unicode text is parsed into foundational grammatical "atoms." For this study, we utilize Botok (OpenPecha, a), a high-accuracy token-based engine that analyzes lexical tokens to reliably resolve the syntactic ambiguity of the Tibetan *shad* (Wylie / or |) punctuation mark (detailed in Appendix I). To prevent excessive semantic fragmentation during this atomic phase, we enforce a strict length constraint: any otherwise valid syntactic split is deterministically suppressed if the preceding sequence contains fewer than four syllables.

Immediately following this atomic segmentation, these base Unicode units are dynamically concatenated into discrete, non-overlapping text spans. The pipeline sequentially aggregates adjacent atoms until the cumulative length satisfies specific upper and lower bounds, which in our experiments were configured to target a window of 10 to 30 words. This guarantees that each resulting sequence provides sufficient semantic context without exceeding the embedding model's optimal input capacity.

H.3 Transliteration

In the final step, the correctly sized Tibetan Unicode spans are passed through a deterministic conversion module. This generates the corresponding EWTS strings, safely transliterating the constrained text segments into Latin characters to ensure compatibility with the downstream foundation models.

I Guidelines for Segmenting a Tibetan Corpus

I.1 General Guidelines

Double or quadruple *shad* (Wylie / or |) or *gter tsheg* (Wylie :) should be treated as strong boundaries, as in prose they mark the end of long syntactic units (such as passages, paragraphs, or chapters) or the closing of specific short syntactic units (such as citations or expressions of thoughts), and in verse the end of a line. Segmentation should generally be implemented at these occurrences.

Single *shad* or *gter tsheg* should be treated as weak boundaries, as they mark the end of short syntactic units, such as clauses, phrases, or sentences. Segmentation at these occurrences depends on specific conditions:

a. Split after a single *shad* when it is immediately preceded by:

- Sentence terminators (final particles): *go; ngo; do; no; bo; mo; 'o; ro; lo; so; to;*
- Optative/imperative suffixes, which commonly immediately follow an adjective, a verb, an adverbialized adjective, or an adverbialized verb: *cig; gyur cig; zhig; shig; shog; rogs; rogs gnang.*

b. Split after a single *shad* when it is immediately followed by:

- Section markers/enumerations: *dang po; gnyis pa;* etc.;
- Sentence-initial markers: *de nas; de bas na; de'i rjes; de lta bas na; de yang; der yang; de ma yin; de ma thag; de ma gtogs; de min; de phyir; de'i phyir; de bzhin du; des na; gzhan du na; gzhan yang; yang na; 'o na; 'on te; 'on kyang; gal te; gal srid; slar yang; spyir la; spyir btang; spyir yang; mdor na; mdor bsdus na.*

c. Do NOT split after a single *shad* when it is immediately preceded by:

- Gerund markers: *nas; bzhin; bzhin du; bzhin pa; bzhin par; kyin;*
- Ablative, locative, genitive, conjunctive, and instrumental particles: *la; su; du; na; ru; tu; nas; las; gi; gyi; kyi; -'i; 'i; yi; phyir; dang; zhing; cing; ste; te; kyang; yang; -'ang; 'ang; -'am; 'am; pas; bas; ltar; gis; -n/m/r/l gyis* (excluding *par gyis, bar gyis*); *kyis; yis;*
- Adverbializer markers: *par; bar;*
- Topic marker: *ni.*

d. Do NOT split after a single *shad* when a correlative pair (e.g., *ji ltar* <> *de ltar; ji srid* <> *de srid; ji snyed* <> *de snyed; ji tsam* <> *de tsam*) is present in the text but not complete at the point of the *shad*. In other words, when in the chunk preceding the *shad* any of the following are present: *ci zhig; ci ste; ci 'dra; ci ltar; ci bzhin; ci tsam; ji snyed; ji srid; ji ltar; ji bzhin; ji skad; ji tsam,*

refrain from splitting before any of the following correlative forms appear (if present): *de bzhin; de ltar; de tsam; de skad; de 'dra.*

In the case of citations, split immediately before the chunk containing the opening phrase (e.g., *de ltar kyang bkod pa chen po las//*) and immediately after the chunk containing the closing phrase (e.g., *zhes gsungs pa lta bu'o//*). However, the rules of c. and d. still apply and take priority. For example, in the following case: *da ni dkyil 'khor bri ba'i thabs ston te/ de ci'i phyir zhe na/ gru bzhi pa la sgo bzhi pa// sgo ba rnams dang ldan pa ste// rta babs bzhi dang ldan par bya// lte ba dang ni mu khyud ldan// zhes gsungs te/ gru bzhi ni phrin las rnam pa bzhis 'gro ba'i don byed pa'i rtags yin pas gru bzhir ldan no//,* instead of splitting immediately before *de ci'i phyir zhe na/* and immediately after *zhes gsungs te/*, treat the entire sequence as a single unit.

In the case of *mantra*s and *dhāraṇī*s (i.e., a string of syllables that do not follow the rules for Tibetan syllables), treat them as one separate unit, ignoring all *shad*s in between.

I.2 Clarification on Certain Conventions and Their Transliteration

The rules described above concerning the application of a double *shad* are not always followed by Tibetan scribes. Some apply a single *shad* systematically also after a final particle or at the end of a verse line. In such cases, a single *shad* should be regarded as equivalent to a double *shad*.

A space in the Tibetan script, often following any type of *shad*, is rendered in Extended Wylie with an underscore.

Any space in texts transliterated in Wylie following any type of *shad* (i.e., a space representing an actual space in the Tibetan script) should be distinguished from those typically corresponding to the syllable separator *tsheg* and should therefore be treated as an underscore.

A *shad* is often omitted when it follows the letter *ga*, which can create ambiguity in transliteration. Therefore if in the original Tibetan text the final word of a syntactic unit ends in the letter *ga* (not *nga*) without a *shad*, and the end of the unit is indicated only by a space, it should be interpreted as terminating with a single *shad*. Likewise, if the final word ends with the letter *ga* (not *nga*) with a single *shad*, it should be interpreted as terminating with a double *shad*.

In this way, sometimes the complication is taken into consideration in transliterated texts, so that letter *ga* plus space in Tibetan script is transliterated as *-g/; ga/; gi/; gu/; ge/; go/*, each marked as being followed by an underscore, and letter *ga* plus a single *shad* plus space is transliterated as *-g//; ga//; gi//; gu//; ge//; go//*, each marked as being followed by an underscore. However, that is not always the case, in which scenario only the application of a double *shad* can be deduced with certainty (i.e., after a final particle or at the end of a verse line).

If such spaces as described above are not represented by underscores in the transliteration and are instead rendered as ordinary spaces, cases of omitted *shad* after letter *ga* become difficult to identify.

J Tibetan models

J.1 Data Collection and Processing

We compiled a corpus of e-texts of Classical Tibetan canonical, para-canonical, and non-canonical scriptures and non-scriptures totaling 13.8 GB of clean running text. The data sources include publicly available repositories such as the Spither Tibetan corpus, as well as in-house collections curated by our team of scholars. We also incorporated corpora that may contain limited amounts of modern Tibetan, such as the MC2 corpus. To ensure data quality, we applied document-level deduplication using the MinHash algorithm combined with a Locality-Sensitive Hashing (LSH) index for efficient similarity lookups. Documents with an estimated Jaccard similarity of 0.8 were considered duplicates and removed. Subsequently, we performed rule-based cleaning and normalization to remove comments, HTML artifacts, and long numerical sequences. We also applied paragraph-level filtering, discarding paragraphs with fewer than four tokens, excessively repetitive word patterns, or consecutive duplicates—issues that often result from OCR errors in digitized manuscripts.

J.2 Transliteration Conversion

Some of the data were converted between Unicode and Wylie using a tool we developed which mainly leverages pyewts (OpenPecha, b). The tool operates through a three-stage cascade: (1) a Unicode heuristic that identifies the dominant script family using Unicode code-point categories; (2) a statistical profile classifier that applies extended langdetect n-gram models trained on Sanskrit and Tibetan transliteration profiles; and (3) a deterministic fallback that uses curated regular-expression templates for Sanskrit schemes and structural stack patterns

for Tibetan, validated through pyewts round-trip checks. It supports Sanskrit (Devanāgarī, IAST, SLP-1, Harvard–Kyoto, Velthuis) and Tibetan (Unicode, Wylie/EWTS, ACIP), with additional coverage for Modern Chinese (CJK) and English.

J.3 Tib-B (BERT-based)

We trained a Tibetan-adapted variant of mBERT by continually pre-training bert-base-multilingual-cased (Devlin et al., 2019) on the full 13.8 GB corpus described above. The tokenizer was trained from scratch using the WordPiece model with a target vocabulary size of 32,000 and the standard special tokens ([UNK], [CLS], [SEP], [PAD], [MASK]), optimized for Tibetan. This yields a 30.8 percent reduction in token count compared to the original mBERT tokenizer (fertility of 1.43 vs. 2.07). Continual pre-training followed standard masked language modeling (MLM) with a masking probability of 0.15, a batch size of 8 per device, sequence length of 512, learning rate of 2e-5, AdamW optimizer with weight decay of 0.01, warmup ratio of 0.06, FP16 mixed precision, gradient clipping at a max norm of 1.0, and 3 training epochs.

J.4 Tib-MB (ModernBERT-based)

Tib-MB builds upon jhu-clsp/mmBERT-base (Marone et al., 2025), a multilingual BERT variant based on the ModernBERT (Warner et al., 2024) architecture. The base model comprises 311,871,168 parameters distributed across 22 hidden layers with 12 attention heads, a hidden size of 768, and an intermediate (FFN) size of 1,152. It uses RoPE positional embeddings (both local and global = 160,000) with a maximum positional embedding capacity of 8,192 tokens, and global attention applied every 3 layers. Before training, the tokenizer was extended to support Tibetan EWTS transliteration system. A SentencePiece BPE model was trained on the same 13.8 GB Tibetan corpus with a target vocabulary of 8,000 subword tokens and a character coverage of 0.9995. After filtering tokens already present in the original vocabulary, 5,389 new tokens were added, expanding the final vocabulary from approximately 255,923 to 261,312 tokens. The model's embedding matrix was resized accordingly, with newly added embeddings initialized randomly. The model was continually pre-trained via masked language modeling (MLM) with a masking probability of 0.15. Input documents were tokenized and split using a sliding window strategy: chunk size of 1,024 tokens with a 512-token overlap (stride = 512), yielding multiple overlapping training examples per long document. Chunks shorter than 100 tokens were discarded. Training was conducted for 3 epochs with a per-device batch size of 1 and 8 gradient accumulation steps (effective batch size of 8), learning rate of 2e-5 with linear decay scheduler and 100 warmup steps, AdamW optimizer (= 0.9, = 0.999, = 1e-8) with weight decay of 0.01, and gradient clipping at a max norm of 1.0. Training was performed in FP16 mixed precision with gradient checkpointing enabled to accommodate the longer sequences on a single NVIDIA RTX 4000 Ada Generation GPU (12 GB VRAM).

K Experimental Training Hyperparameters

To ensure the reproducibility of our experiments, Table 10 details the comprehensive set of hyperparameters utilized during the training in our experiments, including the `Tib-Bi-Dharma` model.

Table 10: Models Hyperparameter configurations

Hyperparameter	Value
Epochs	7
Optimizer	AdamW (fused)
Learning Rate	2e-5
Learning Rate Scheduler	Reduce LR on Plateau
Pooling	Weighted Mean
Loss Function	CoSent
Loss Scale	20.0
Warmup Steps	64
Weight Decay	0.1
Max Gradient Norm	1.0
Batch Size (per device)	256
Gradient Accumulation Steps	1
Precision	bf16
Seed	42

PHMartialLawNER: A Tagalog Named Entity Recognition Corpus for the Philippine Martial Law Era

Abdiel Clarence Tabuzo, Vladimir Gray Velazco
Cassandra Cabral, **Moneah Shaila Lacsam** and **Charmaine Ponay**
Department of Computer Science
University of Santo Tomas
encetabuzo@gmail.com, vladimirvelazco0126@gmail.com
collegiancassie22@gmail.com, moneahlacsam@gmail.com, csponay@ust.edu.ph

🤗 PHMartialLawNER 🤗 Trained Models

Abstract

Historical corpora for Tagalog remain limited, particularly texts produced during the Martial Law period under the dictatorship of Ferdinand Marcos Sr. (1972–1986). Much of this material remains undigitized, restricting computational analysis of a significant period in Philippine political history. To support research on historical Tagalog texts, we introduce PHMartialLawNER, a gold-standard named entity recognition corpus constructed from newspapers and underground publications of the Martial Law era. The corpus includes approximately 13k extracted sentence segments (362,000 tokens), consolidated into 8k annotated text spans through a semi-automatic pipeline with manual validation. The reliability of the annotation is measured using Cohen's κ, reaching 0.86 on all tokens and 0.72 on annotated tokens, with a pairwise F1-score of 0.74. The schema defines historically relevant entity categories including Person (Individual, Collective), Organization (Political, Government, Other), Event (Local, International), Production (Media, Government, Doctrine), as well as Time, Numerical Statistics, Location, and Object entities, specifically identifying weapon artifacts. We establish baseline performance using `GLiNER` variants, `calamanCy` models, and transformer-based architectures under zero-shot and few-shot settings. The PHMartialLawNER corpus will be publicly released to support Tagalog NLP, historical text processing, and digital humanities research.

1 Introduction

Despite the growing number of NLP resources for Philippine languages in recent years, most existing datasets have been constructed using modern news and contemporary texts. Efforts such as TLUnified-NER (Miranda, 2023b), HiligayNER (Teves et al., 2025), and CebuaNER (Pilar et al., 2023) have significantly contributed to this progress by providing named entity recognition (NER) datasets for modern language use.

2 Manotoc sa Batasan?
AU-AU MANOTOC
IMEE MANOTOC
Napahiya si Erap !
MAYOR ESTRADA
DEAN GONZALES
MMC VEHICLES
PANGHAKOT NG MGA BOTANTE !
'VIGILANCE RALLY' sa MANILA CATHEDRAL
'LUKSA NG BAYAN'
ANG TINIG NG
Masa
THE people's voice
TAON I • BLG. 42 • MAYO 15-21, 1984 • P1.50 sa METRO MANILA
MGA ALKALDE, INIREKLAMO
UP STUDENTS VS MARCOS

Figure 1: **Historical Newspaper Page from the Martial Law Era.** Example of a raw archival newspaper scan used in the corpus. These historical pages are digitized into machine-readable text before undergoing entity annotation.

However, these datasets, and consequently the models trained on them, are less suitable for historical research because they are built from contemporary corpora that fail to capture linguistic phenomena that appear only in archival materials. These include shifting orthographic conventions, outdated political titles, historical actors whose names rarely appear in modern texts, and domain-specific organizations unique to earlier political periods. As a result, NER models trained on modern corpora usually perform poorly when applied to historical documents.

This limitation is particularly significant for Tagalog materials from the Martial Law era. During this period, press censorship and political repression reshaped the Philippine media landscape, leading to the emergence of alternative and un-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 167–177
July 4, 2026 ©2026 Association for Computational Linguistics

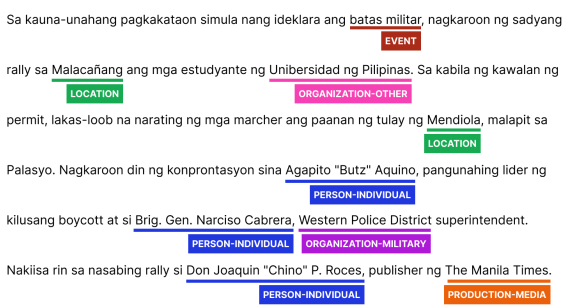


Figure 2: **Entity Annotation on Extracted Newspaper Text.** Example output from the trained NER model showing how domain-specific entities are identified and labeled within a historical Tagalog newspaper passage.

derground publications that documented dissent, resistance movements, and political events. These texts represent an important historical record, but remain difficult to analyze computationally due to the lack of specialized NLP resources.

At the same time, the rapid digitization of archival materials—often described as the "big data of the past" (Kaplan and di Lenardo, 2017)—has created new opportunities for large-scale analysis in digital humanities. However, Tagalog remains underrepresented in these archival NLP efforts, and there is a lack of available NER datasets for historical Tagalog texts.

To address this gap, we present a named entity recognition dataset constructed from digitized Tagalog newspapers and underground publications from the Martial Law period. The dataset enables downstream tasks such as entity-based search, event extraction, and large-scale historical analysis.

2 Background

2.1 Historical Newspapers and Martial Law

Historical newspapers are widely recognized as rich primary sources for reconstructing socio-political events, public sentiment, and cultural shifts (Baumgartner, 1981). In the Philippine context, the Martial Law era (1972–1986) represents a period in which press censorship and state control dramatically shaped the flow of information. During this time, the mainstream media were heavily restricted, prompting the emergence of an alternative press often referred to as the mosquito press, such as Ang Tinig ng Masa (see Figure 1). These publications disseminated dissenting perspectives despite surveillance and repression (Rosario-Braid and Tuazon, 1986).

Underground publications, including We Forum, Malaya, and CPP-aligned outlets such as Ang Bayan, played a crucial role in exposing corruption and human-rights violations (Melencio, 2023; Olea, 2012). Many of these documents have since been digitized through institutional repositories such as the University of Hawai'i at Mānoa's eVols collection, which preserves newsletters, memos, and underground print materials from the Marcos regime (University of Hawai'i at Mānoa Philippine Collection, 2023).

Digitizing such archives is not only essential for preservation, but also enables computational access and large-scale corpus analysis (Oberbichler, 2024). Compared to earlier historical periods—where Tagalog materials are sparse due to colonial suppression (Neumann, 2005; Punzalan, 2006)—the Martial Law era exhibits a greater abundance of Tagalog texts, driven by the resurgence of nationalism and political activism. This makes the period particularly suitable for constructing a historical Tagalog NER dataset.

2.2 Named Entity Recognition

NER facilitates the automatic identification of persons, organizations, locations, events, and other semantic categories (Tjong Kim Sang and De Meulder, 2003). It remains a foundational component of information extraction, powering tasks such as archival search, historical knowledge graph construction, and large-scale socio-political analysis (Ehrmann et al., 2023). An example of this process applied to a Tagalog historical news snippet is shown in Figure 2.

Early approaches relied on rules, lexicons, and pattern-matching, but these systems struggled with the noisy, heterogeneous nature of historical texts—especially when OCR distortions and archaic spellings are present (Todorov and Colavizza, 2022). Traditional machine-learning models like Conditional Random Fields improved generalization but still fell short of modern deep-learning approaches, with performance often in the 60–70% F1 range on historical corpora (Ehrmann et al., 2023).

Deep contextual models, such as BiLSTM-CRF architectures (Ma and Hovy, 2016), and later transformer-based approaches such as BERT (Devlin et al., 2019), RoBERTa, and XLM-R (Conneau et al., 2020), achieved state-of-the-art performance through contextualized embeddings. In historical NER, these methods significantly outperform rule-

based strategies when trained with appropriate domain data (Ehrmann et al., 2023).

A key challenge for historical NER is temporal drift. Entity distributions, naming conventions, and lexical usage shift across decades (Rijhwani and Preotiuc-Pietro, 2020). Models trained solely on contemporary data suffer on temporally distant corpora, but diversified sampling across time improves performance by more than 10% F1-score in low-resource settings. This reinforces the need for domain-specific training data—such as a Martial Law NER corpus.

Recent NER developments relevant to low-resource and historical corpora include `GLiNER` and `calamanCy`, both of which are employed in this study. `GLiNER` (Zaratiana et al., 2024) is a generalist, label-descriptive NER framework capable of zero-shot and few-shot extraction, making it particularly useful for bootstrapping annotations in domains where gold-standard data are scarce. Its variants, such as `GLiNER-lg`, have been referenced as strong baseline models for initializing Tagalog annotations[1]. `calamanCy` (Miranda, 2023a), on the other hand, is a Tagalog-focused NLP toolkit built on spaCy, providing language-specific tokenization, lexical resources, and transformer pipelines optimized for modern Tagalog texts. Prior work demonstrates that `calamanCy` performs effectively when paired with high-quality in-domain data. Together, these systems represent complementary approaches to low-resource NER—`GLiNER` excels in cross-domain generalization through its label-descriptive architecture, while `calamanCy` leverages Tagalog-specific priors— both of which prove essential for addressing the linguistic and historical challenges posed by Martial Law era newspaper corpora.

3 Methodology

Figure 3 shows the general process for creating the gold-standard dataset.

3.1 Data Collection and Preprocessing

Historical newspapers were sourced from the University of Hawai'i eVols archive[2] and the Bantayog ng mga Bayani digital repository[3], both of which contain Martial Law era underground and opposition publications. All scanned pages were processed using OCR systems and subsequently subjected to manual verification to reduce noise, segmentation errors, and character distortions typical of historical print materials. Texts containing excessive OCR corruption were removed to ensure that the resulting corpus reflected the linguistic and historical characteristics relevant to this study.

3.2 Iterative Annotation

To construct a high-quality NER dataset in a low-resource historical domain, a semi-automatic annotation workflow was done; designed to balance efficiency and accuracy. The workflow began with the application of `GLiNER` in zero-shot mode, which generated a preliminary silver-standard layer of entity labels. These outputs served as an initial scaffold for the human annotation process. Rather than annotating from scratch, three human annotators reviewed and corrected the auto-generated spans, a strategy known to significantly reduce annotation time and cognitive load in low-resource NER settings.

The annotations were done in Argilla[4] and was carried out through an iterative cycle, in which each batch of corrected data informed improvements to both the annotation guidelines and subsequent annotation decisions. After each round, the annotated texts were checked for consistency, and cases of disagreement or ambiguity were documented for discussion. These conflict-resolution sessions allowed refinement of definitions for historically specific entity types—such as political organizations, military units, and event references—which often required contextual judgment unique to the discourse of the Martial Law era. The updated guidelines were then redistributed to the annotators, and the next batch of annotations was completed with clearer, more standardized rules. These conflict-resolution sessions allowed for the refinement of definitions for the specific entity types used in this study, as detailed in Table 1.

To quantitatively monitor annotation reliability across all iterations, we calculated the IAA after each batch using Cohen's κ for all tokens and pairwise span-level F1 for labeled entities. These metrics allowed for the detection of inconsistencies, adjusted guidelines when necessary, and ensured that annotation quality improved over time. Through

[1] https://github.com/ljvmiranda921/calamanCy/tree/master/models/v0.1.0-gliner

[2] https://hdl.handle.net/10524/70580

[3] https://bantayogngmgabayani.org/bantayog-digital-library/

[4] https://argilla.io/

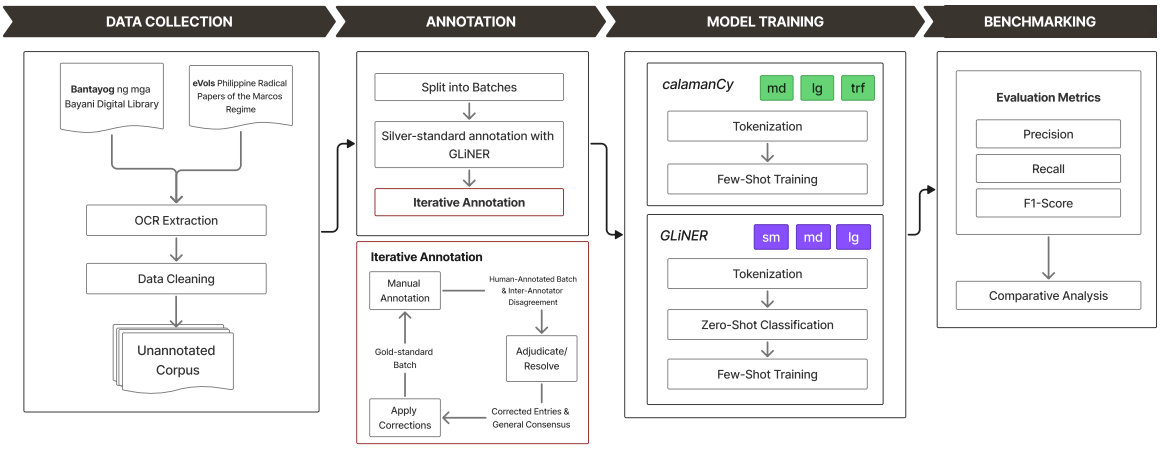


Figure 3: **Complete System Architecture.** Showcases the whole process for creating and benchmarking the gold-standard dataset

repeated cycles of machine-assisted pre-annotation, human correction, guideline refinement, and quality validation, the results gradually converged on a stable annotation scheme. The result of this layered, iterative process is a gold-standard corpus of approximately 362,000 tokens from around 13,000 sentences, representing the NER dataset tailored specifically to Tagalog texts from the Martial Law period.

3.3 Model Training

Supervised and few-shot training procedures were applied for historical Tagalog named entity recognition. The gold-standard corpus was partitioned into training, development, and test sets using stratified sampling (70:10:20) to preserve the distribution of the publication sources.

Two training pipelines were implemented. For `calamanCy` models, annotations were converted to a spaCy binary format to enforce token–entity span alignment and resolve overlapping spans. Three architectures were trained: `calamanCy-md`, `calamanCy-lg`, and transformer-based `calamanCy-trf`, with GPU-optimized hyperparameter settings.

For `GLiNER` models, annotations were reformatted into HuggingFace-compatible token classification datasets by mapping IOB tags to base entity labels. Small-, medium-, and large-scale variants were fine-tuned using a unified optimization pipeline with gradient accumulation to approximate larger effective batch sizes. Training and validation losses were tracked to monitor convergence.

Few-shot experiments were conducted by training models on smaller curated subsets of the training data to evaluate sample efficiency under low-resource historical Tagalog conditions. Following the benchmark design of Abadie et al. (2022), final evaluations were performed on the held-out test set using precision, recall, and F1-score.

Multilingual transformer baselines, including XLM-R and RoBERTa-Tagalog, were fine-tuned for comparison but were not evaluated under few-shot settings.

4 Experiments

4.1 Dataset Quality (IAA)

To evaluate the consistency and reliability of the gold-standard dataset, inter-annotator agreement (IAA) was calculated across multiple annotation batches as reflected in Table 2. Annotator consistency is critical in low-resource corpus development, particularly when dealing with historically complex and domain-specific text. Despite the iterative annotation approach, earlier batches (batches 1 to 3) achieved higher IAA while later batches (batches 4 to 5) showed a gradual decline in agreement. The decrease could potentially be a consequence of annotating progressively larger volumes of data (1200 in the 1st batch to 2000 in the 5th batch). Annotators also increasingly encountered edge cases, rare historical terms, and more complex entity structures that are harder to classify consistently. In the final iteration, the corpus achieved a Cohen's κ of 0.86 on all tokens, 0.72 on annotated tokens, and a span-level pairwise F1 of 0.74, all of which remain within acceptable thresholds for

Entity	Description	Examples
PERSON-INDIVIDUAL	Individual persons, whether deceased or living, real or fictitious, including political figures and activists.	Ferdinand Marcos, Ninoy, 'Raul Segovia, sekretaryo-heneral'
PERSON-COLLECTIVE	Named references to groups of people that are not organizations.	Mga Aquino, Katoliko, Pilipino
ORGANIZATION-POLITICAL	Political parties, movements, or activist groups at national or international level.	Kilusang Bagong Lipunan, Kabataang Makabayan, LP
ORGANIZATION-GOVERNMENT	Government institutions, branches, departments, or geopolitical actors.	Comelec, Estados Unidos, Department of National Defense
ORGANIZATION-MILITARY	Formal armed forces, military units, or alliances.	AFP, NPA, 42nd Infantry Battalion, PC-INP
ORGANIZATION-OTHER	Organizations not covered by other categories.	Unibersidad ng Pilipinas, CBCP, University of Santo Tomas
LOCATION	Geographic entities including administrative regions, buildings, and natural formations.	Ilog Pasig, Plaza Miranda, Camp Crame, timog Korea
TIME	Temporal expressions including dates, years, and time ranges.	Setyembre 21, 1972, Martes, Nobyembre 2
PRODUCTION-MEDIA	Media works, publications, broadcasts, and related artifacts.	Ang Bayan, Radio Veritas, WE Forum, Tinig ng Masa
PRODUCTION-GOVERNMENT	Official documents, decrees, and state-issued materials.	Presidential Decree No. 1081, Batas Pambansa Blg. 880
PRODUCTION-DOCTRINE	Political, philosophical, or religious belief systems.	Marxism-Leninism-Maoism, Demokratiko, Kapitalismo
NUMERICAL STATISTICS	Quantities, monetary values, percentages, and measurements.	500 pesos, 80 porsiyento, sampung kilong bigas
OBJECT-WEAPON	Physical combat-related artifacts including specific weapon models.	M-16, Bolo, Tear-gas, bala, baril
EVENT	Historical events, incidents, or recognized social occurrences.	Araw ng Manggagawa, eleksyon, Lakbayan, kudeta, Olympics

Table 1: Detailed entity annotation taxonomy.

Batch	Size	F1-Score	κ (all tokens)	κ (annotated only)
1	1200	74.13	87.05	74.09
2	1500	74.43	87.04	73.88
3	1500	74.03	86.96	73.89
4	1800	73.86	84.36	69.98
5	2000	71.61	82.17	66.72

Table 2: IAA training scores across annotation batches.

high-quality annotation (Artstein, 2017)–especially for the more fine-grained NER task.

Later batches exhibited unusually low agreement, signaling persistent inconsistencies that could not be resolved through guideline revision alone. Trial trainings from the initial gold-standard dataset resulted in poor model performance on several entities–e.g., *Event-Local*, *Event-International*, and *Object*–due to inherent issues of underrepresentation and fine-grained entity boundaries. In response, we revised the entity list in consultation with a domain expert, combining *Event* into a single label and restricting *Object* to only include weapons. This adjustment stabilized IAA in the final annotation rounds and ensured coherent and reliable labeling across all entity types.

4.2 Final Model Performance

To establish a performance ceiling for the constructed dataset, we trained `calamanCy-trf` alongside other multilingual and monolingual BERT models on the full corpus. As demonstrated in Table 3, these models exhibited the best performance across Precision, Recall, and F1-Score metrics, underscoring the overall strength of these architectures when trained in higher-resource settings.

4.3 Few-shot Model Performance

To evaluate model adaptability in low-resource scenarios, we trained multiple NER models using standardized train–validation–test splits and evaluated them using strict span-level F1-scores, covering both full-data fine-tuning and few-shot training experiments. Since full-data performance often fails to reflect real-world constraints where annotations

Model	Precision	Recall	F1-Score
GLiNER_small	77.84	72.16	74.89
GLiNER_medium	80.96	76.33	78.58
GLiNER_large	78.63	78.38	78.50
calamanCy_md	80.99	80.47	80.73
calamanCy_lg	78.95	80.20	79.57
calamanCy_trf	83.65	84.31	**83.98**
RoBERTa-tagalog (monolingual)	84.67	83.25	83.96
XLM-RoBERTa (multilingual)	84.11	83.82	83.97

Table 3: **Final Model Performance.** Trained models benchmarked on the final corpus.

are scarce, evaluating under few-shot conditions was particularly relevant for historical Tagalog.

As illustrated in Figure 4, despite uneven Inter-Annotator Agreement (IAA) in the later portions of the dataset, several models demonstrated robust adaptability. The Tagalog-specific `calamanCy-trf` and the larger `GLiNER` variants produced the highest F1-scores in both full-data and few-shot scenarios, highlighting their capacity to generalize to historical language patterns.

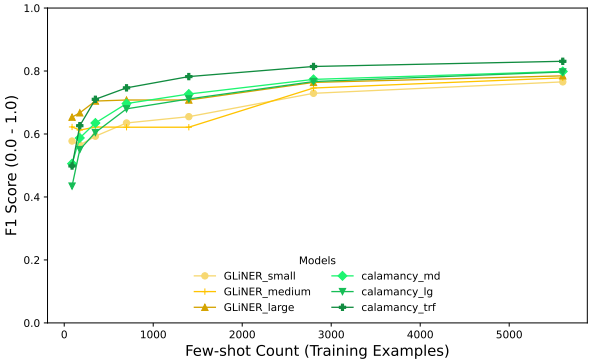


Figure 4: **Few-shot Model Performance.** F1-scores of GLiNER and calamanCy models across increasing training set sizes.

Notably, `GLiNER` exhibited impressive resilience under few-shot training due to its label-descriptive architecture, which allows it to learn entity behavior even from limited data. Together, these results suggest that while high-quality annotated data remains essential for peak performance, both `calamanCy` and `GLiNER` can leverage small, well-annotated subsets effectively, even when portions of the corpus exhibit weaker IAA. Complete quantitative results are presented in Table 4.

4.4 Fine-grained Entity Performance

To further examine model behavior across entity categories, we also conducted a per-entity analysis using the best-performing model, `calamanCy-trf`, as reported in Table 5. The results reveal distinct learning patterns across entity types at increasing levels of data availability. Entities such as *Person-Individual*, *Time*, and *Organization-Military* demonstrated high sample efficiency—*Organization-Military* in particular saw substantial gains, jumping from 41.80% to 91.91% F1 as the model learned to recognize specific acronyms and unit designations prevalent in the corpus, indicating that `calamanCy`'s transformer variant effectively captures structural patterns even with limited data.

Conversely, context-dependent and underrepresented entities exhibited high volatility in low-resource settings. *Production-Media* and *Organization-Other* began with negligible scores (7.66% and 19.14%, respectively) but surged to over 70% F1 by the final split, suggesting that data scarcity was the primary bottleneck for these classes. Notably, the *Event* and *Object-Weapon* categories plateaued or slightly regressed at the largest dataset size (peaking at the 2,801-sample mark), implying that increased data volume may have introduced greater label ambiguity or noise for these specific categories.

5 Analysis

5.1 Qualitative Challenges and Entity Ambiguity

Qualitative analysis highlighted that annotators struggled primarily with boundary consistency for context-dependent entities.

Span	Start	End	Annotator 1	Annotator 2
P5.00 por metro kwadrado	39	63	–	Numerical Statistics
P5.00	39	44	Numerical Statistics	–

Figure 5: **Span-level disagreement for a numerical statistic.** Accepted span highlighted in green; other annotator's span shown in yellow.

- *Numerical Statistics*: Conflicts arose because the annotators failed to include some descriptive phrases (e.g., *anim na anak* 'six children') as statistical data, exposing ambiguity in the initial guidelines. Annotators also disagreed on whether to include preceding units (e.g.,

Model	F1 @ 0-shot	F1 @ 88	F1 @ 175	F1 @ 350	F1 @ 700	F1 @ 1400	F1 @ 2801	F1 @ 5602
GLiNER_small	42.07	57.76 ± 0.63	56.48 ± 0.48	59.29 ± 0.16	63.50 ± 0.33	65.50 ± 2.78	72.92 ± 0.58	76.55 ± 0.78
GLiNER_medium	44.08	62.24 ± 0.06	61.23 ± 0.11	62.21 ± 0.53	62.15 ± 0.52	68.08 ± 1.37	74.60 ± 1.56	77.82 ± 1.47
GLiNER_large	**46.61**	$\mathbf{65.35 \pm 0.32}$	$\mathbf{66.72 \pm 0.47}$	$\mathbf{70.46 \pm 0.99}$	70.79 ± 0.85	72.43 ± 0.10	76.38 ± 0.63	78.47 ± 0.05
calamanCy_md	—	50.16 ± 1.10	58.27 ± 0.58	63.77 ± 0.24	69.94 ± 0.29	72.71 ± 0.39	77.34 ± 0.40	80.28 ± 0.53
calamanCy_lg	—	44.32 ± 1.18	53.39 ± 1.10	61.14 ± 1.20	67.50 ± 0.38	71.72 ± 0.29	76.35 ± 0.11	79.86 ± 0.18
calamanCy_trf	—	50.72 ± 0.89	61.99 ± 0.56	70.40 ± 0.91	$\mathbf{74.36 \pm 0.27}$	$\mathbf{77.74 \pm 0.44}$	$\mathbf{81.64 \pm 0.19}$	$\mathbf{83.04 \pm 0.12}$

Table 4: Few-shot training results.

Entity	F1 @ 88	F1 @ 175	F1 @ 350	F1 @ 700	F1 @ 1400	F1 @ 2801	F1 @ 5602
Person-Individual	81.95	86.75	91.02	90.73	92.48	92.12	**93.57**
Person-Collective	64.98	69.58	73.72	74.76	80.88	78.99	**82.24**
Organization-Political	27.71	58.85	62.53	73.20	71.56	79.58	**81.36**
Organization-Government	41.33	48.73	56.68	60.98	68.45	70.01	**73.51**
Organization-Military	41.80	50.43	72.48	78.15	89.23	89.57	**91.91**
Organization-Other	19.14	32.51	52.05	56.78	59.96	69.57	**73.70**
Location	49.45	72.05	77.37	80.69	79.60	84.34	**85.21**
Time	61.35	79.20	83.13	87.76	90.83	92.29	**92.48**
Production-Media	7.66	12.88	35.81	52.10	62.77	72.55	**76.83**
Production-Doctrine	55.88	64.77	78.29	79.46	83.18	81.18	**84.08**
Numerical Statistics	36.15	51.34	59.18	59.87	66.88	72.39	**74.52**
Object-Weapon	13.43	30.88	34.59	50.33	71.29	**81.11**	74.68
Event	23.74	29.72	49.13	59.71	58.44	**66.53**	64.66
Overall F1	49.84	62.63	71.02	74.65	78.23	81.44	83.09

Table 5: calamanCy_trf Few-Shot Evaluation per entity

"P5.00 por metro kwadrado" versus "P5.00", as seen in Figure 5).

- *Metonymy*: A major source of confusion was the semantic overlap between Organization-Government and Location entities, where terms like "Malacañang" or "US" could refer to either the government institution or the physical place. Similarly, event terms like "Batas Militar" (Martial Law) were sometimes misclassified as political organizations.

These challenges motivated the refinement of entity definitions in the final corpus configuration.

5.2 Impact of Training Data and Model Choice

Assessing the robustness of our core contributions, we conducted a Two-Way ANOVA. The analysis confirmed that the size of the training subset, the type of model, and their interaction exerted an extremely strong and statistically significant influence on the prediction's F1-scores (all $p < .001$) as seen in Table 6. However, post-hoc pairwise comparisons for the Model factor showed that no individual model pairs differed significantly for multiple comparisons (with $\alpha = 0.05$). This indicated that the overall model effect found on the omnibus test was spread across architectures rather than driven by a single dominant model. The large F-values observed for subset size and for the interaction term further show that both the amount of training data and the way different models respond to increasing data volume are critical determinants of NER performance.

Post-hoc comparisons for training subset size in revealed that performance gains, while generally increasing with more data, became less significant and converged nearing 1,400 and 2,801 samples, where several adjacent subset comparisons were no longer statistically significant, e.g. 2801-5602 subset sizes with $p = 0.5278$ and 700-1400 subset sizes with $p = 0.3152$). This range is therefore recommended as a minimum effective training size for reliable generalization on the historical NER dataset, although training on the full dataset would still remain preferable for achieving peak performance.

The significant interaction effect shows that the models improved at different rates. `GLiNER` models performed better with limited training data (all $p < .001$), while Tagalog-specific transformer models, particularly `calamanCy-trf`, achieved higher peak F1-scores as training size increased (only model with all $p < .001$). This highlights the importance

Source	Df	Sum Sq	Mean Sq	F-value	p-value	Sig.
Model	5	979	195.7	296.01	$< 2 \times 10^{-16}$	***
Subset Size	6	8285	1380.8	2088.39	$< 2 \times 10^{-16}$	***
Model × Subset	30	1455	48.5	73.38	$< 2 \times 10^{-16}$	***
Residuals	84	56	0.7	–	–	–

Table 6: Two-way ANOVA results evaluating the effects of model architecture, training subset size, and their interaction on F1-score.

of evaluating NER models across varying training data sizes to better understand how performance changes as more data becomes available.

6 Conclusion

The Martial Law era represents one of the most politically significant, yet computationally neglected periods in Philippine history. Through iterative annotation, guideline refinement, and quality validation, we present PHMARTIALLAWNER as a step toward making this period more accessible. Our experiments have shown that the task is learnable even under low-resource conditions, and that model architecture and training data volume are both critical factors in driving NER performance for historically variable text.

While `calamanCy-trf` emerged as the strongest performer, our per-entity analysis highlights that fine-grained and underrepresented entity types remain a persistent challenge—one that corpus expansion and improved annotation strategies may help address in future work. We hope PHMARTIALLAWNER serves not only as a benchmark resource for Tagalog NLP, but also as a foundation for broader digital humanities research into Philippine archival texts.

7 Acknowledgments

The authors are grateful to Lester James V. Miranda for technical guidance and mentorship in the implementation of this study, and for his prior work that informed this research. We acknowledge Gian Paolo R. Mayo from the Department of History for providing historical guidance and helping connect this work with the humanities. We recognize the members of the Human Rights Violations Victims' Memorial Commission (HRVVMC) who assisted in the annotation and validation of the dataset used in this study. We further acknowledge the Bantayog ng mga Bayani Foundation and the University of Hawai'i at Mānoa Philippine Collection for preserving and providing access to archival materials that served as primary sources for the corpus used in this study. Finally, we appreciate the panelists and other faculty members from the Department of Computer Science for their feedback and contributions.

References

N. Abadie, E. Carlinet, J. Chazalon, and B. Duménieu. 2022. A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories. In *Document Analysis Systems. DAS 2022.*, number 13237 in Document Analysis Systems. DAS 2022., La Rochelle, France. Springer, Cham.

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Joseph Baumgartner. 1981. Newspapers as historical sources. *Philippine Quarterly of Culture and Society*, 9(3):256–258.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Frédéric Kaplan and Isabella di Lenardo. 2017. Big data of the past. *Frontiers in Digital Humanities*, Volume 4 - 2017.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *Preprint*, arXiv:1603.01354.

Gloria E. Melencio. 2023. The fall and rise of the marcoses: From mosquito press to troll farms. *UP Los Baños Journal*.

Lester James V. Miranda. 2023a. calamanCy: A Tagalog natural language processing toolkit. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 1–7, Singapore. Association for Computational Linguistics.

Lester James V. Miranda. 2023b. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

A. Lin Neumann. 2005. The philippines: Amid troubles, a rich press tradition. *Committee to Protect Journalists*.

Sarah Oberbichler. 2024. Large-scale research with historical newspapers: A turning point through generative ai. DH Lab, Leibniz Institute of European History (IEG).

Ronalyn V. Olea. 2012. Underground press during martial law: Piercing the veil of darkness imposed by the dictatorship. *Bulatlat*.

Ma. Beatrice Emanuela Pilar, Dane Dedoroy, Ellyza Mari Papas, Mary Loise Buenaventura, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023. CebuaNER: A new baseline Cebuano named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 792–800, Hong Kong, China. Association for Computational Linguistics.

Ricardo L. Punzalan. 2006. Archives of the new possession: Spanish colonial records and the american creation of a 'national' archives for the philippines. *Archival Science*, 6(3):381–392.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

Florangel Rosario-Braid and Ramon R. Tuazon. 1986. Communication media in the philippines: 1521–1986. Philippine Social Science Council.

James Ald Teves, Ray Daniel Cal, Josh Magdiel Villaluz, Jean Malolos, Mico Magtira, Ramon Rodriguez, Mideth Abisado, and Joseph Marvin Imperial. 2025. Hiligayner: A baseline named entity recognition model for hiligaynon. *Preprint*, arXiv:2510.10776.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Konstantin Todorov and Giovanni Colavizza. 2022. An assessment of the impact of ocr noise on language models. *Preprint*, arXiv:2202.00470.

University of Hawai'i at Mānoa Philippine Collection. 2023. Papers of the underground movement during the marcos regime (philippine radical papers of the marcos regime). Philippine Studies Digital Collection.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Corpus Breakdown and Linguistic Profile

The PHMartialLawNER corpus presents a linguistic profile, heavily shaped by the socio-political realities and media censorship of the Philippine Martial Law era. Unlike contemporary datasets that often rely on standard modern news, this corpus is built from underground "mosquito press" publications, which inherently feature a highly militarized and politically charged vocabulary.

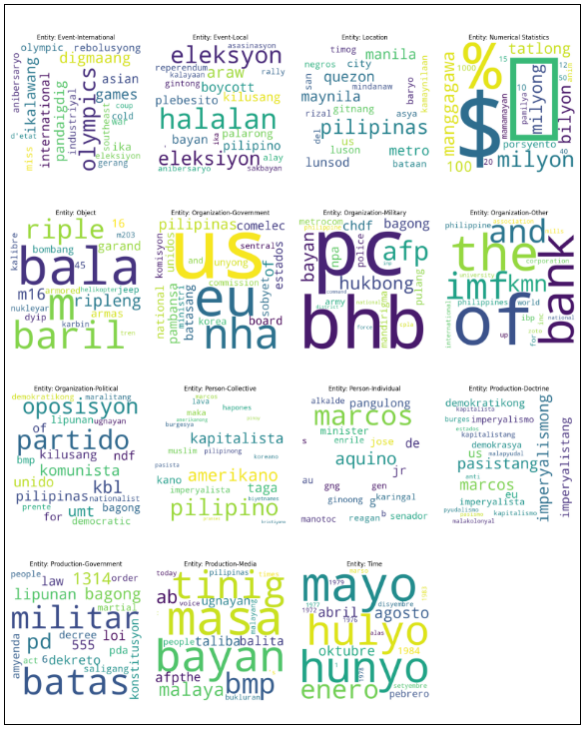


Figure 6: **Top Unigrams per entity** Top unigrams extracted from the corpus, highlighting the core socio-political vocabulary of the Martial Law era, dominated by terms related to key figures, the military, and the resistance (e.g., "Marcos," "militar," "manggagawa" (workers)

An exploratory data analysis of the corpus using Word Clouds (showing the top unigram and trigram of each entity, as seen in Figures 6, 7) vividly illustrates this unique linguistic landscape. The top unigrams and trigrams highlight the dominance of militant and resistance-focused terminologies.

In particular, refer to Figure 6 vocabulary for the Object entity is overwhelmingly saturated with high-frequency terms such as "bala" (bullet) and "baril" (gun). The prominence of these specific combat-related artifacts perfectly illustrates the militarized nature of the texts and the era's focus on armed resistance and state-sponsored military.

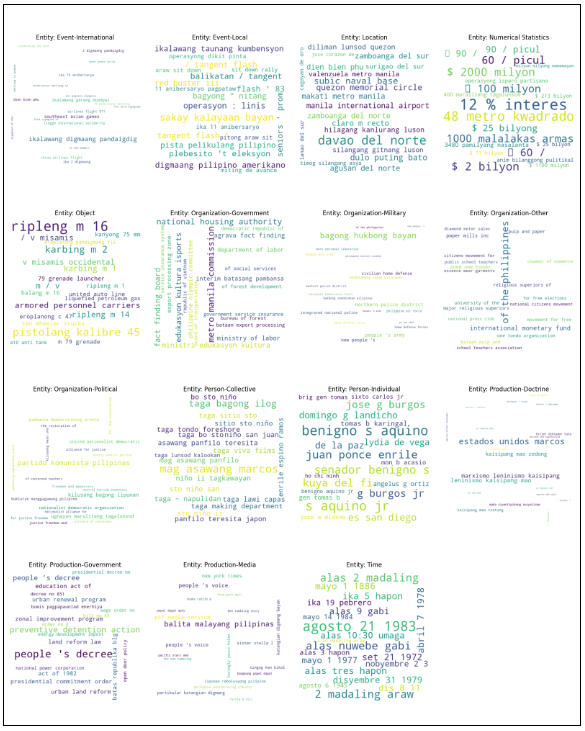


Figure 7: **Top Trigrams per entity** Top trigrams extracted from the corpus, revealing the prominence of specific political organizations, government initiatives, and resistance groups central to the underground press discourse (e.g., "bagong hukbong bayan," (New People's Army) "kilusang bagong lipunan" (New Society Movement)

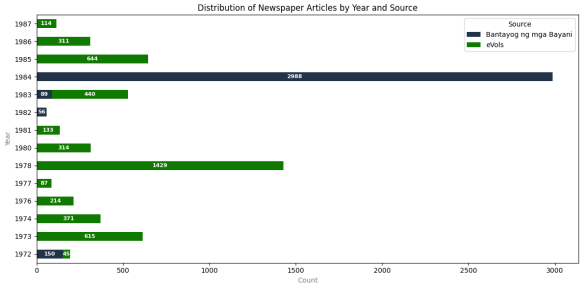


Figure 8: **Year Distribution of Published Newspapers** The digitized newspapers compiled for this corpus were published between 1972 and 1987, a timeframe that encompasses the height of the Martial Law period while intentionally extending into the immediate post-regime aftermath to capture the transitional socio-political discourse.

A.2 Annotation Difficulties and Schema Refinement

Developing a gold-standard dataset from historical texts exposed several challenges that guided the final schema decisions.

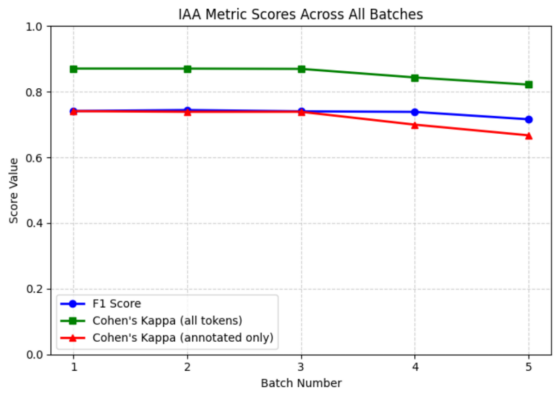


Figure 9: **IAA metrics trend** Tracks the mean Cohen's Kappa (all-tokens and annotated only) and F1-scores across the five annotation batches, illustrating a slight decline in agreement as sample volume increased

A.2.1 Declining IAA Trends

As shown in 9, agreement metrics began strong but declined between Batches 3 and 5. This downward trend correlates with the increasing volume of annotated samples, which exposed annotators to more complex edge cases and historical ambiguities.

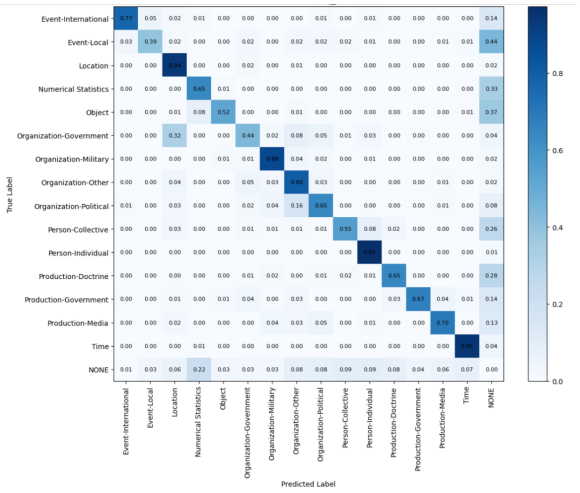


Figure 10: **Confusion Matrix of IAA per entity** A confusion matrix highlighting annotator agreement and semantic overlaps, notably showing the frequent misclassification between government organizations and locations due to metonymy

A.2.2 Semantic Overlap via Metonymy

The Confusion Matrix of Entities (Refer to Figure 10) highlights pervasive boundary ambiguity. Specifically, 32% of Organization-Government mentions were misclassified as Location. This stems from the metonymic use of place names (e.g., "Malacañang" or "Estados Unidos" (United States) to represent political institutions.

A.3 Dataset Availability

The PHMARTIALLAWNER corpus is publicly released to facilitate further research in Tagalog Natural Language Processing and the digital humanities. It is hosted on Hugging Face: https://huggingface.co/datasets/etdvprg/PHMartialLaw-NER_final

Fluency and Faithfulness in Human and Machine Literary Translation

Sarah Griebel* **and Ted Underwood**
School of Information Sciences
University of Illinois Urbana–Champaign
sarahg8@illinois.edu tunder@illinois.edu

Abstract

Literary translation requires balancing target-language fluency with faithfulness to the source. Recent large language models (LLMs) often produce fluent translations, but it remains unclear whether fluency corresponds to semantic preservation in literary text. We examine this relationship using 130,486 translated paragraphs from 106 novels in 16 source languages, including human, Google Translate, and TranslateGemma translations. Fluency is measured as original-likeness with a translationese classifier trained on paragraph part-of-speech n-grams, and faithfulness with the automatic translation evaluation metric COMET-KIWI. We control for paragraph length and find a consistent negative correlation between fluency and faithfulness. The pattern appears for both human and Google Translate, but is weaker and often non-significant for TranslateGemma. These results show that segment length matters for automatic evaluation and suggest a trade-off between fluency and faithfulness in literary translation.[1]

> *"Either the translator leaves the writer in peace as much as possible, and moves the reader toward him; or he leaves the reader in peace and moves the writer toward him."*
>
> Friedrich Schleiermacher, *On the Different Methods of Translating* (1813) transl. Susan Bernofsky

1 Introduction

Since Schleiermacher's famous formulation that a translator must either "move the reader toward the writer" or "move the writer toward the reader", translation theory has often treated fluency and fidelity as opposing poles. Yet this tension has rarely been examined empirically at scale.

Literary translation involves reconciling competing objectives. Translators must produce text that reads thoughtfully, smoothly, and accurately in the target language while preserving the semantic and stylistic content of the original. Figurative language–often present in literary text–adds a unique challenge. For example, a translation that closely follows the source structure of an idiom or metaphor may preserve meaning but sound unnatural or "foreign", while a highly fluent translation may involve paraphrasing or restructuring that alters semantic content, glosses it, or attempts to "domesticate" certain elements of the source text in the target tongue. (Translation theorists use different terms to describe this process; for example, foreignization and domestication (Schleiermacher, 2012; Venuti, 1995) or deforming tendencies (Berman, 2012).)

Recent large language models (LLMs) have demonstrated strong performance in literary machine translation (Zhang et al., 2025; Karpinska and Iyyer, 2023; Wang, 2025). Prior work has primarily evaluated LLM literary translation quality using reference-based metrics or human evaluation, as it is more difficult to assess results from quality estimation (reference-free) evaluation metrics in this domain (Karpinska and Iyyer, 2023). Therefore, less is known about how fluency and quality estimation metrics relate to one another in literary machine translation.

In this work, we examine the relationship between fluency and faithfulness metrics at paragraph level across human and machine translations. We analyze a corpus of 130,486 translation paragraphs from 106 novels spanning 16 source languages, with human translations, Google Translate outputs, and translations generated using TranslateGemma (Finkelstein et al., 2026), a recent open translation model based on the Gemma3 architecture (Team et al., 2025).

We operationalize fluency as original-likeness,

*Corresponding author.

[1]Code and analysis for the project is available at https://github.com/griebels/fluency-faithfulness-mt.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 178–189
July 4, 2026 ©2026 Association for Computational Linguistics

measured using a translationese classifier trained to distinguish original English paragraphs from translated English paragraphs. To isolate syntactic fluency independent of semantic content, we train the classifier on part-of-speech (POS) anonymized text. We operationalize faithfulness using COMET-KIWI scores.

Using these measures, we compute correlations between fluency and faithfulness separately for human translations, LLM translations, and a traditional machine translation system. We hypothesize that LLM translations may exhibit a tradeoff between fluency and semantic faithfulness, reflected in a negative correlation between original-likeness and adequacy scores, while human translations may exhibit weaker or different relationships.

Our contributions are as follows:

- We introduce a large-scale paragraph-level analysis of fluency and faithfulness in literary translation, spanning over 130,000 paragraphs across human, Google Translate, and LLM translations.
- We propose a scalable framework for measuring fluency independent of semantic fidelity using POS-based literary translationese classification with book-level held-out evaluation.
- We apply reference-free quality estimation (COMET-KIWI) to get a proxy for faithfulness across human, Google Translate, and LLM literary translations.
- We provide empirical evidence that a negative relationship exists between fluency and faithfulness signals and is more distinctly observed among human literary translations of similar lengths, signaling that COMET-KIWI does not align with POS-based signals of fluency.

2 Related Work

2.1 Fluency and faithfulness in machine translation evaluation

Machine translation quality has traditionally been evaluated along two primary dimensions: fluency, referring to the naturalness and grammaticality of the target text, and faithfulness (or adequacy), referring to preservation of source meaning. Human evaluation protocols commonly distinguish these dimensions explicitly, recognizing that highly fluent translations may diverge semantically from the source, while highly faithful translations may exhibit syntactic or stylistic interference from the source language.

2.2 Translationese and original-versus-translated classification

Translated text exhibits systematic differences from text written originally in a target language, a phenomenon known as translationese. Prior work has shown that translated text can be distinguished from original text using lexical, syntactic, and distributional features, including function word frequencies (Koppel and Ordan, 2011), string kernel methods (Popescu, 2011), part-of-speech perplexity (Bizzoni et al., 2020), mean word length, syllable ratio, character n-grams, and others (Volansky et al., 2015). These differences typically arise from structural interference from the source language, simplification, and normalization during translation.

In this work, we build on translationese detection methodology to operationalize fluency as original-syntactic-likeness, measured by a classifier trained to distinguish original English paragraphs from translated English paragraphs using features built on ordered parts-of-speech input. We do not aim to produce a perfectly accurate classifier; rather, our goal is to produce a highly interpretable one.

2.3 Faithfulness measures in literary machine translation

Automatic translation evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are limited, as they rely on reference comparisons and often fail to produce high overlap with human judgment. More recent neural metrics such as COMET (Rei et al., 2020) address these limitations by predicting translation quality using multilingual encoders. COMET fine-tunes pretrained multilingual language models to estimate translation adequacy based on reference and target text, and has been shown to have high overlap with human evaluation (Lee et al., 2023). Its reference-free variant, COMET-KIWI (Rei et al., 2022), enables direct estimation of semantic faithfulness without requiring gold reference translations.

While neural evaluation metrics can be considered "black box" as to what their outputs are actually measuring (Karpinska et al., 2022), COMET outputs were demonstrated to rely heavily on multilingual sentence representations derived from pre-

trained encoders (Rei et al., 2023). Therefore, COMET provides a scalable proxy for semantic adequacy, which we interpret here as a signal of faithfulness when analyzed alongside fluency measures.

Reference-based evaluation metrics work well to evaluate the performance of machine translation systems, but they impose a difficulty when comparing human to machine translations by requiring a gold-standard reference. Quality estimation metrics in machine translation are a valuable tool, as they do not require reference translations to compare source and target texts. Moreover, recent advances have shown overlap between QE results and human judgment in machine translation tasks (Zerva et al., 2024; Specia et al., 2020). By correlating COMET-KIWI with syntactic fluency, we are able to study how it performs for literary translation evaluation.

3 Data

3.1 Aligned literary translation corpus

We use the Par3 Dataset from Thai et al. (2022), a paragraph-aligned corpus of literary translations spanning 113 novels and 122,819 source paragraphs. The corpus includes source texts in 16 languages: Czech, Dutch, French, German, Hungarian, Italian, Japanese, Norwegian, Persian, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, and Chinese. For each source paragraph, the dataset includes two or more human translations (maximum 5), as well as machine-generated translations from Google Translate.

We generate LLM translations using TranslateGemma, a fine-tuned translation model suite built on the architecture of Gemma3, evaluated on WMT benchmarks across 55 languages that demonstrates substantial improvements over the base Gemma model (Finkelstein et al., 2026). We use the 4-billion-parameter model version to retain low overhead compute. We run inference on the model through Ollama, an open-source tool to run models locally. Translation prompts are given at the paragraph-level, using the recommended prompt technique for the model.

Paragraph alignment is provided as part of the dataset release and further verified using multilingual embedding similarity filtering (Section 4.3).

Human translations in the corpus span a range of publication periods. Because translation publication dates are not made available with the dataset, we do not condition analyses on this. However, we do discuss implications of translation publication dates on fluency measures (Section 7).

Duplicate and overlapping works. The Par3 corpus contains multiple entries corresponding to the same underlying literary work, including cases where a single novel appears multiple times. To avoid overweighting individual works and introducing near-duplicate material into classifier training, we identify and remove entries that represent the same original text.

When multiple entries corresponded to a single work (e.g., *Les Misérables 1, Les Misérables 2*), we retain only the first representative volume. In contrast, distinct works by the same author (e.g., separate novels within a series) are retained. This filtering reduced the original dataset from 113 books to 106 books.

3.2 Original English comparison corpus

To train a translationese classifier, we use a corpus of 115 English-language novels (i.e., originally written in English) published between 1800 and 1930. In selecting the earliest dates for our original dataset, we aimed to gather texts that would align with an empirical translation lag from the source texts in Par3, which were published between 1399[2]–1982 but mostly clustered in the middle. Our 1930 ceiling reflects standard U.S. public domain constraints.

Paragraphs from this corpus serve as positive examples of original English, while translated paragraphs serve as negative examples.

4 Methods

We measure fluency and faithfulness independently using translationese classification and quality estimation, respectively. This allows us to examine their relationship across human and machine literary translations.

4.1 Fluency as original-likeness

We operationalize fluency as *original-likeness*, defined as the degree to which a translated paragraph resembles original English text.

To isolate syntactic structure independent of semantic content, we anonymize each paragraph by

[2]Though the earliest source text in the Par3 dataset was published in 1399, the majority (nearly 90%) were published after 1800.

replacing tokens with their fine-grained part-of-speech (POS) tags, using a spaCy model. For example:

> Original: "Ivan asked Dmitry to meet him at the inn for lunch today?" Alyosha asked quickly.
>
> POS-anonymized: "NNP VBD NNP TO VB PRP IN DT NN IN NN NN ." NNP VBD RB .

We use the POS-based classifier as our fluency metric because it isolates syntactic structure independent of lexical content, ensuring that fluency reflects structural naturalness rather than topical or lexical differences.

Minimum paragraph length filtering. To ensure that the classifier captures meaningful linguistic structure, we exclude paragraphs shorter than a minimum 20 word threshold. Short segments such as chapter headings (e.g., "IV" or "Chapter VII") or ornamental separators do not contain sufficient syntactic structure to reliably distinguish original from translated text. Including such segments would introduce noise and potentially bias the classifier toward non-linguistic artifacts.

Classifier training, preprocessing, and cross validation. We train a logistic regression classifier using TF-IDF features. Features include unigram, bigram, and trigram n-grams, with a maximum vocabulary size of 20,000 features. The classifier uses L2 regularization with inverse regularization strength $C = 10.0$, `class_weight=balanced`, and a maximum of 2000 optimization iterations.

To minimize the influence of paragraph length on classifier predictions, we downsample the translated class using ten paragraph-length bins derived from the original-English text distribution. For each bin, we randomly sample an equal number of translated paragraphs to match the number of original paragraphs in that bin. This procedure reduces the dataset used for comparison from 264,487 to 130,486 paragraphs.

Because each source paragraph may have multiple translated versions (including human translations, Google Translate, and TranslateGemma outputs), the translated class contains varying numbers of observations derived from the same underlying content. To prevent paragraphs with more translation variants from disproportionately influencing training, we assign each translated paragraph a sample weight inversely proportional to the number of translations available for its source paragraph. Specifically, if a source paragraph has n translated versions, each translation is assigned weight $1/n$. Original English paragraphs are assigned uniform weight.

To prevent leakage of book-specific stylistic features, we perform 10-fold group cross-validation using books as grouping units, ensuring that all paragraphs from a given book appear exclusively in either training or evaluation folds.

Because source languages are unevenly represented, we additionally stratify books by source language when constructing folds. Languages with only one or two books are combined into a single "rare languages" stratum. This ensures that each fold contains representation from as many languages as possible while preserving strict book-level separation, and prevents classifier performance from being driven by language-specific artifacts or imbalanced language representation.

Each paragraph receives an out-of-fold prediction corresponding to the probability of belonging to the translated class, $P(\text{translation} \mid x)$. We define fluency as:

$$\text{Fluency}(x) = 1 - P(\text{translation} \mid x) \quad (1)$$

Higher scores indicate greater similarity to original English and therefore higher fluency.

Translationese classifier performance Table 1 summarizes the performance of the translationese classifier. The classifier achieves strong performance, confirming that translated paragraphs exhibit systematic structural differences from original English at paragraph level.

Features	Accuracy	Macro F1	AUC
POS abstraction	0.760	0.753	0.847

Table 1: Performance of translationese classifier under book-level cross-validation.

4.2 Faithfulness as semantic adequacy

We measure faithfulness using COMET-KIWI, a neural metric trained to predict translation quality without reference translations. COMET-KIWI takes as input the source paragraph and translated paragraph and produces a scalar adequacy score between 0-1.

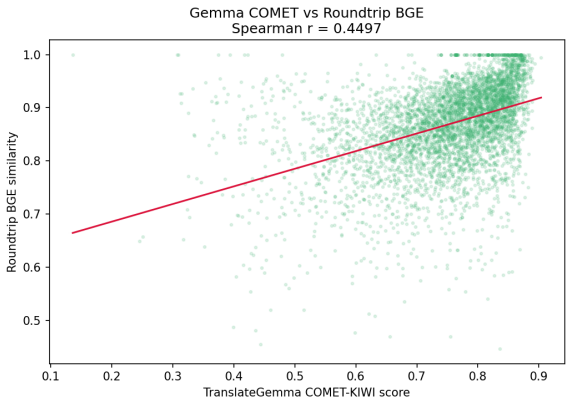


Figure 1: Correlation between COMET-KIWI scores and roundtrip semantic similarity for TranslateGemma translations. Higher COMET-KIWI scores correspond to greater semantic consistency under roundtrip translation.

We use the `Unbabel/wmt22-cometkiwi-da` model for quality estimation.

Sanity check: COMET-KIWI and roundtrip semantic consistency. To verify that COMET-KIWI provides a reliable signal of semantic faithfulness in our dataset, we compare COMET-KIWI scores with roundtrip semantic similarity. For each TranslateGemma output, we generate a roundtrip translation by translating the English translation output back into the source language. We compute cosine similarity between the original source paragraph and the roundtrip text using BGE-M3 multilingual sentence embeddings (Chen et al., 2024), which has shown strong performance for a range of multilingual tasks and supports long inputs.

We observe a strong positive correlation between COMET-KIWI scores and roundtrip semantic similarity (Figure 1), indicating that COMET-KIWI captures semantic preservation consistent with independent embedding-based similarity measures. This provides additional support for using COMET-KIWI as a reference-free proxy for semantic faithfulness in our literary translation task.

4.3 Guardrail filtering for alignment and translation validity

To ensure reliable comparison, we apply several filtering criteria to remove paragraphs with potential misalignment or translation errors.

To make sure LLM translations are valid, we apply a length consistency filter, excluding paragraphs where the TranslateGemma output exceeds the Google Translate output by more than 500 characters, which may indicate degenerate generation. This consistency filter removes 42 LLM translations.

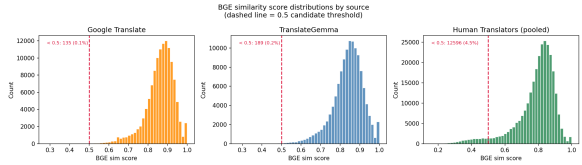


Figure 2: Distribution of source–translation cosine similarity using BGE-M3 embeddings. Low-similarity tail primarily reflects misaligned or invalid paragraph pairs.

We then employ paragraph alignment validation, filtering sentences below a cosine similarity threshold, as described in detail below.

Alignment validation and filtering. Because paragraph alignment in literary corpora may occasionally contain errors or mismatches, we use embedding similarity to identify potentially misaligned translations. We compute cosine similarity between source and translated paragraphs using BGE-M3 embeddings for all translation types.

The distribution of similarity scores reveals a long tail of low-similarity examples, particularly among human translations (Figure 2). To assess whether these low-similarity cases reflect true semantic divergence or alignment errors, we manually inspect a random sample of 30 paragraphs from the low-similarity tail. We find that over 93% of these cases exhibit clear alignment errors, primarily unrelated content. Table 2 shows representative examples of these misalignments.

Based on this analysis, we apply a similarity threshold to exclude paragraphs with low embedding similarity, ensuring that subsequent analyses focus on reliably aligned translations. Importantly, we use embedding similarity rather than COMET-KIWI scores for filtering to avoid circularity, as COMET-KIWI serves as our primary faithfulness metric.

Overall, filtering on this cosine similarity threshold reduced the translation corpus by 2.0%. This primarily affected human translations, rather than translations from Google Translate or TranslateGemma.

source_para	gt_para	translation	cosine similarity
* * * * *	* * * * *	Go ask the artillerymen of Saint-Roch …	0.422
68. Par exemple la Sphynge, une lionne ailée à tête de femme, qui interroge Œdipe.	68. For example the Sphynge, a winged lioness with the head of a woman, who questions Oedipus.	CHAPTER XIV	0.427
— Кому же они привезли?	Whom did they bring?	"You have heard it all.	0.409

Table 2: Examples of low-similarity paragraph pairs manually identified as misaligned. In these cases, the "translation" field contains content inconsistent with the aligned source paragraph (e.g., chapter headings or unrelated narrative content).

5 Results

5.1 Raw fluency and faithfulness scores by paragraph and by book

In Figure 3 we observe that human translations appear at both higher and lower faithfulness thresholds of the graph, while the Google Translate and TranslateGemma results cluster near the top. The contrast becomes more stark when reviewing the bottom right quadrant, indicating lower faithfulness but high fluency. In fact, 27.53% of human translations in our dataset score above 0.5 in our fluency measure, while only 21.37% and 15.26% of Google Translate and TranslateGemma (respectively) score above this threshold.

5.2 Fluency–faithfulness association is length-conditioned

We examine the relationship between fluency and faithfulness using Spearman correlations between fluency (1 - $P(\text{translation} \mid x)$) and COMET-KIWI adequacy scores.

When analyzing all paragraphs, we observe a small overall association ($\rho = -0.0374$). However, stratifying by paragraph length reveals a clear difference between long and short paragraphs: For paragraphs shorter than 100 words, the correlation between fluency and COMET-KIWI tends to be larger (Spearman $\rho = -0.086$, $p < 0.001$). For paragraphs of 100 words or longer, the relationship reduces to nearly zero (Spearman $\rho = -0.0151$, $p < 0.05$).

5.3 Paragraph length correlates with both metrics

The length-stratified difference suggests that paragraph length may confound the fluency–faithfulness relationship. We therefore examine correlations between paragraph length and each metric across the full dataset.

Indeed, we find that paragraph length is negatively correlated with COMET-KIWI scores (Spearman $\rho = -0.2641$, $p < 0.001$) and with fluency (Spearman $\rho = -0.1565$, $p < 0.001$). Notably, paragraph length is essentially uncorrelated with embedding-based alignment similarity used in our guardrail filtering (Spearman $\rho = 0.0074$), suggesting that the negative relationship between length and COMET-KIWI is not driven by increasing misalignment in longer segments.

Together, these results indicate that paragraph length exerts substantial influence on both fluency and adequacy metrics.

5.4 Controlling for length increases the association between fluency and faithfulness

To estimate the direct relationship between fluency and faithfulness independent of paragraph length, we compute a partial Spearman correlation controlling for length. The association between fluency and COMET-KIWI strengthens overall (partial Spearman $\rho = -0.0827$, $p < 0.001$). Parceling this out by translation source, the signal is weakly positive for TranslateGemma ($\rho = 0.0129$) $p < 0.05$). It is negative for human translations ($\rho = -0.0783$ $p < 0.001$) and Google Translate ($\rho = -0.0512$ $p < 0.001$).

Stratifying the analysis by paragraph length reveals a similar pattern. Within each length bin, the partial correlation between fluency and COMET-KIWI remains negative, ranging from $\rho = -0.068$ for shorter segments (20–30 words) to $\rho = -0.125$ for mid-length segments (61–100 words), before slightly declining in association for the longest segments (100+ words). Figure 4 illustrates these bin-level correlations.

Interestingly, the pattern within length bins is also similar to the global pattern: TranslateGemma

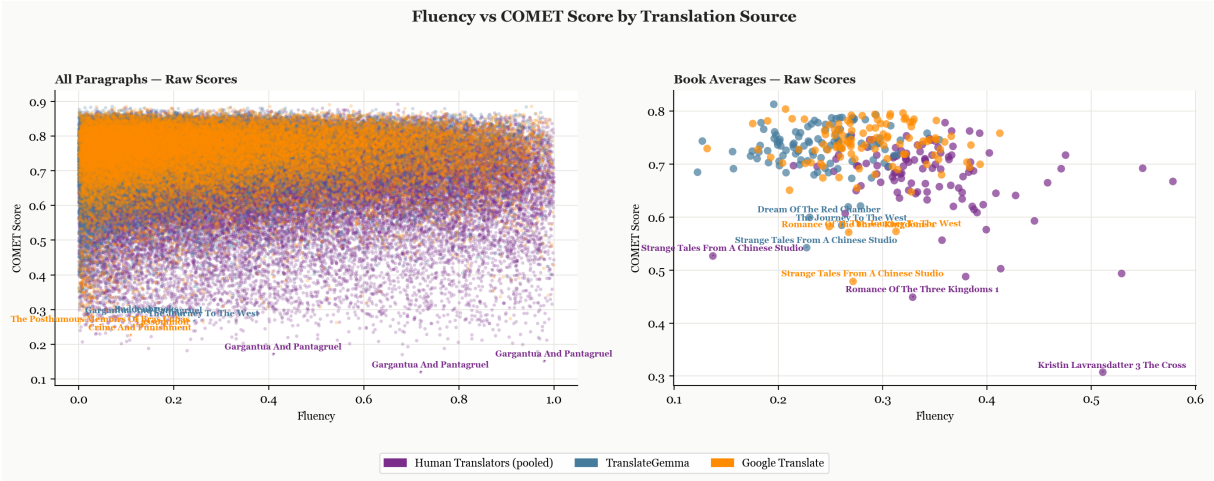


Figure 3: Fluency vs. COMET-KIWI score by translation source. Points represent individual paragraphs (left) and books (right), colored by source: human translators (purple), TranslateGemma (blue), and Google Translate (orange). For human translations, scores are averaged across multiple translations. Labeled points indicate the largest outliers by residual distance from the per-source fit line.

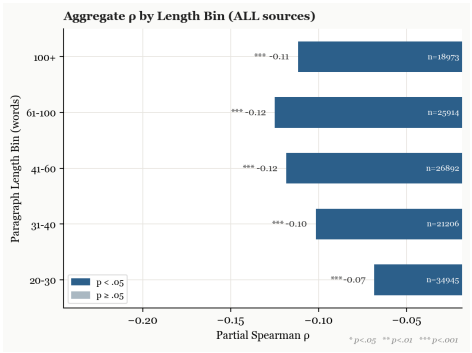


Figure 4: Partial Spearman correlation between fluency and COMET-KIWI after controlling for paragraph length, computed separately for paragraph length bins. The correlation is modest but consistently negative across bins, with the strongest association appearing for mid-length segments (61–100 words).

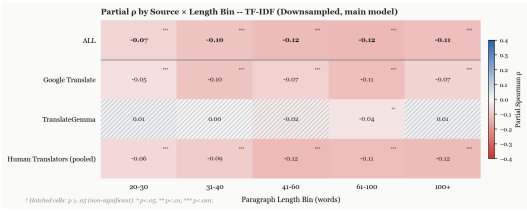


Figure 5: Partial Spearman correlations between fluency and COMET-KIWI by paragraph source and length bin. Stars indicate statistically significant correlations ($p < 0.05$). Human and Google Translate translations show a broadly consistent negative association across bins, while the LLM-generated translations exhibit weaker and often non-significant relationships.

tends to have a weaker, sometimes positive correlation, while human translations show stronger negative correlations. More fine-grained variation across translation sources is shown in Figure 5.

These patterns suggest that the global correlation is affected by the differences between human and machine translation outputs for these literary paragraphs: while human translations show a consistent negative relationship, both TranslateGemma and Google Translate tend to show a weaker, sometimes positive relationship between fluency and faithfulness.

6 Conclusion

This study examined the relationship between syntactic fluency and semantic faithfulness in literary translation using a large corpus of paragraph-aligned texts translated by humans, Google Translate, and TranslateGemma.

Across the full dataset, COMET-KIWI scores exhibited a negative correlation with paragraph length, indicating that longer segments tend to receive lower predicted adequacy scores. To control for this confounding factor, we regressed both fluency and COMET-KIWI on paragraph length and analyzed the residual signals. The resulting length-controlled measures reveal a modest but consistent negative association between syntactic fluency and semantic faithfulness.

Additional analyses further support this pattern. When paragraphs are grouped into length bins, the

negative relationship between fluency and COMET-KIWI becomes more pronounced across most bins. This effect is visible for both human and machine translations, though the signal remains weaker and statistically non-significant for TranslateGemma.

Taken together, these findings suggest a small but systematic tension between producing syntactically native-like English and preserving semantic correspondence with the source text. While the magnitude of the relationship is modest, the consistency of the pattern across analytical approaches indicates that the two signals are not entirely independent.

This pattern also has implications for the interpretation of automatic evaluation metrics. Because COMET-KIWI is trained to approximate adequacy judgments, it may implicitly favor translations that preserve structural or lexical patterns closely aligned with the source text. As a result, the metric may partially reward signals associated with translationese when those signals correlate with surface-level source fidelity.

7 Discussion

One possible interpretation of these findings is that COMET-KIWI captures aspects of translation quality that are not aligned with the notion of fluency operationalized in this study. Our fluency measure reflects the degree to which a paragraph's syntactic structure resembles that of original English prose. By contrast, COMET-KIWI is trained to estimate semantic adequacy relative to the source text. The weak negative association observed here therefore suggests that producing highly native-like syntactic patterns does not necessarily coincide with maximizing semantic correspondence as predicted by COMET-KIWI.

The behavior of the fluency classifier also provides a lens to examine stylistic differences between human and machine translations. Because the model was trained on original English texts published between 1800 and 1930, it captures the syntactic distributions characteristic of that period. Although exact publication dates are unavailable for the human translations in Par3, nearly 90% of the source texts were published after 1800, suggesting that at least some translations may have been produced within a similar historical context. In addition, translators working on historical literary texts may intentionally approximate earlier stylistic conventions in the target language. Contemporary machine translation systems and large language models, by contrast, are typically optimized for modern standard English and may only mirror the language used in popular translations — that is, those that appear most frequently in their training data (Ströbel and Maier, 2025). It is unclear whether modern models would naturally reproduce historical stylistic patterns absent the high-quality human examples in their training data. Consequently, for books with fewer human examples, model outputs may appear less "original-like" to the classifier, even when they are grammatically fluent.

More broadly, these findings echo long-standing discussions in translation theory regarding the tension between producing fluent target-language prose and preserving the foreign structure of the source text. While the correlations observed here are modest, they suggest that computational metrics may capture traces of this trade-off in large-scale translation data.

Limitations

Some limitations should be considered when interpreting these results.

First, the metadata available for the dataset is limited. Information about translators and publication dates of individual translations is not consistently available. As a result, we are unable to analyze how translator identity, translation era, or editorial practices might influence the observed relationship between fluency and faithfulness.

Second, the fluency measure used in this study captures only syntactic signals derived from part-of-speech patterns. While this abstraction helps isolate structural aspects of translationese, it does not account for lexical choices, stylistic register, or rhetorical effects that may also influence perceptions of fluency. Future work could incorporate additional stylistic features or human evaluation to further contextualize these findings.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Antoine Berman. 2012. Translation and the Trials of the Foreign. In Lawrence Venuti, editor, *The Translation Studies Reader*, third edition, pages 240 – 253. Routledge.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. TranslateGemma Technical Report. *arXiv preprint*. ArXiv:2601.09012 [cs].

Marzena Karpinska and Mohit Iyyer. 2023. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing Evaluation Metrics for Translation. *arXiv preprint*. ArXiv:2210.13746 [cs].

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, 11(4):1006. Number: 4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marius Popescu. 2011. Studying Translationese at the Character Level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 634–639, Hissar, Bulgaria. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2023. The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics. *arXiv preprint*. ArXiv:2305.11806 [cs].

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Friedrich Schleiermacher. 2012. On the Different Methods of Translating. In Lawrence Venuti, editor, *The Translation Studies Reader*, third edition, pages 43 – 63. Routledge.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Phillip Benjamin Ströbel and Felix Klaus Maier. 2025. Echoes of Antiquity: Towards Understanding History through Human andLLM-Based Classical Text Translations. In Taylor Arnold, Margherita Fantoli, and Ruben Ros, editors, *Anthology of Computers and the Humanities*, pages 33–49. Association for Computers and the Humanities.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. *arXiv preprint*. ArXiv:2210.14250 [cs].

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*. Routledge.

V. Volansky, N. Ordan, and S. Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Qiufen Wang. 2025. Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator. *PLOS ONE*, 20(10):e0335261.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the Quality Estimation Shared Task at WMT 2024: Are LLMs Closing the Gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

A Classifier Variants

To evaluate the robustness of our findings, we train several alternative versions of the translationese classifier. These experiments test whether the observed relationship between fluency and paragraph length depends on specific modeling choices, particularly the feature representation and the use of downsampling.

Our primary classifier uses TF–IDF weighting on POS n-gram features. One concern with this representation is that TF–IDF emphasizes rare feature combinations, which could cause the classifier to rely disproportionately on infrequent POS patterns rather than capturing broader syntactic tendencies associated with translated text. To examine this possibility, we train additional classifiers using raw frequency counts of POS n-grams instead of TF–IDF weighting.

We also test whether our downsampling strategy, which balances paragraph length distributions across classes, influences the relationship between fluency and paragraph length. To assess this, we train models both with and without downsampling.

Specifically, we train Logistic Regression classifiers under the following configurations:

TF–IDF, full dataset: all samples (264,487 paragraphs), without downsampling Count features, downsampled dataset: POS frequency counts with length-balanced downsampling Count features, full dataset: POS frequency counts using all samples

Across these variants, we observe that the correlation between paragraph length and predicted fluency increases relative to the primary model, indicating that length sensitivity is not unique to the TF–IDF representation.

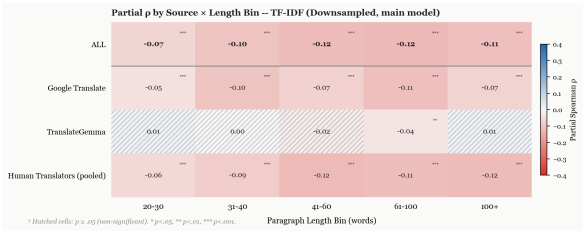
Table 3 summarizes the results.

Despite these differences, the core findings remain stable across classifier variants. In particular, the negative partial Spearman correlation between fluency and COMET-KIWI persists, and the TranslateGemma condition remains largely non-significant across paragraph-length bins.

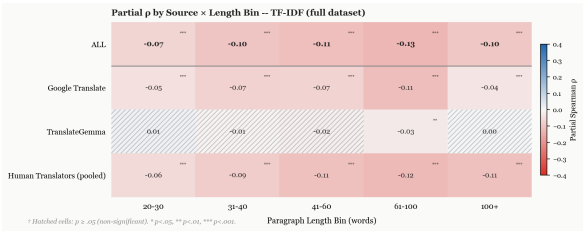
These results are visualized in Figure 6.

Table 3: Robustness checks for alternative translationese classifiers. Length–Fluency ρ denotes the Spearman correlation between paragraph length and predicted fluency. Partial ρ denotes the partial Spearman correlation between fluency and COMET-KIWI controlling for paragraph length.

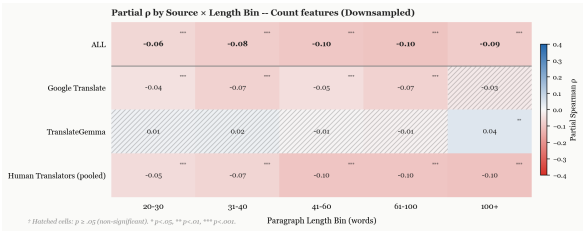
Features	Sampling	Samples	Length–Fluency ρ	Partial ρ
TF–IDF	Downsampled	130486	-0.1565	-0.0827
TF–IDF	Full	264487	-0.2470	-0.0344
Count	Downsampled	130486	-0.3631	-0.0719
Count	Full	264487	-0.4280	-0.0444



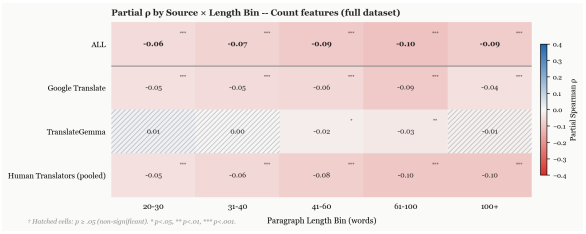
(a) TFIDF (downsampled)



(b) TFIDF (full dataset)



(c) Count features (downsampled)



(d) Count features (full dataset)

Figure 6: Heatmaps showing partial Spearman correlations between fluency and COMET-KIWI across paragraph-length bins for different classifier variants. The overall pattern across paragraph-length bins, as well as the non-significant results for TranslateGemma, remains consistent across feature representations and sampling strategies.

Directional Alignment and Narrative Agency in Human–LLM Co-Writing

Halfdan Nordahl Fundal
TEXT Center
Aarhus University
halfi@cc.au.dk

Yuri Bizzoni
TEXT Center
Aarhus University
yuri.bizzoni@cas.au.dk

Abstract

We investigate narrative agency in human–LLM creative co-writing, asking who drives story development in turn-based collaboration. Using a new corpus of 87 human–LLM co-written stories, we apply sentiment and semantic modeling to quantify affective alignment and semantic novelty in turn-taking, and directional measures to assess which agent shapes narrative progression. Our results show asymmetric influence: human turns introduce greater semantic novelty and are more likely to shape subsequent developments, whereas LLM contributions predominantly elaborate on human-introduced elements. At the sentiment level, alignment is also asymmetric, but more bidirectional: LLMs exhibit stronger turn-level emotional adaptation than humans, but both agents track each other's emotional valence and LLMs show an independent tendency to more positive emotional baselines. These findings indicate a complementary division of labor in human–LLM co-writing, where humans drive narrative innovation and direction, while LLMs act as adaptive amplifiers that sustain coherence and elaborate emerging narratives.

1 Introduction

Large Language Models (LLMs) are becoming frequent "collaborators" in creative writing practices, supporting both story ideation and stylistic transformation. Prior work has evaluated these systems in terms of output quality and controllability, while the turn-by-turn dynamics by which human and model contributions shape a joint narrative remain less well understood.

Creative co-writing is a complex process of negotiation over several dimensions at once, such as emotional tone, content, as well as the story's overall narrative direction. In human–human collaboration, such negotiation involves the necessity of alignment and influence between participants, as well as tensions between different creative aims. As we begin to interact with LLMs as social partners, human–LLM collaboration raises parallel questions: do models tend to adapt to human cues, or do they introduce new trajectories? How does narrative agency work in mixed human–AI teams?

Understanding these dynamics is crucial for both theoretical and practical reasons. From a cognitive and literary perspective, psycholinguistic evaluations of co-writing offer a window into affective coordination and shared meaning between humans and artificial agents. From a design perspective, insights into alignment and influence can inform the development of writing systems that better support creativity or user control. From a digital humanities perspective, studying the affective and informational dynamics of co-authored narratives extends computational approaches to literary sentiment analysis into interactive, multi-agent settings.

In this paper, we introduce a new corpus of co-created human-LLM turn-taking narratives, and examine different aspects of the dynamics that emerge between humans and LLMs through narrative co-creation. Specifically, we assess *emotional alignment* through sentiment modeling between inputs, and evaluate narrative *agency* as the relationship between *novelty*, as information-theoretic deviation of inputs from a baseline, and *resonance*, as the tendency of novel elements to remain active in the subsequent development of the story.

We address the following research questions:

- **RQ1 (Baseline differences):** Do human and LLM turns differ in mean valence or valence distribution?
- **RQ2 (Affective coordination):** Do human and LLM interlocutors align (turn-to-turn), and is alignment symmetric?
- **RQ3 (Narrative influence):** Whose novel contributions are most likely to be taken up in the next turn - humans' or LLMs'?

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 190–201
July 4, 2026 ©2026 Association for Computational Linguistics

Our contribution is thus threefold: (1) a new corpus of human–LLM co-written narratives, (2) directional metrics for affective alignment, and (3) novel applications of information-theoretic methods to quantify linguistic influence, supported by empirical evidence of asymmetric narrative agency in mixed human–AI creative collaboration.

2 Related Work

Although LLMs continue to show impressive performance across NLP benchmarks, research on their creative output is mixed and often highlights limitations in open-ended tasks, such as generating diverse and dynamically evolving narratives (Tian et al., 2024). While an increasing amount of work examines human–LLM collaboration (Li et al., 2025; Wan et al., 2024; Tang et al., 2025), less is known about the turn-by-turn dynamics through which human and model contributions shape one another over the course of a shared narrative.

Human-LLM creativity Findings on human–AI joint creativity remain mixed on whether LLMs increase or inhibit creativity. Access to AI tools can increase absolute contribution of novel artifacts due to the productivity effect (Zhou et al., 2025), and enhance overall writing quality (Noy and Zhang, 2023) compared to the individual. On the other hand, human-LLM interactions in creative fields can result in creative fixation in complex tasks (Cheng and Zhang, 2025) and homogenization of output by narrowing the diversity of ideas (Anderson et al., 2024). Ultimately, human-LLM iteration seems to outperform the individual in creative output, while consistently underperforming human-human collaboration (Tang et al., 2025). Earlier work on human-LLM interaction paradigms has explored facilitation of collaborative co-creative frameworks through dyadic storytelling (Roemmele and Gordon, 2015; Clark and Smith, 2021), but did not touch on the emergent dynamics between agents. Although summary output evaluates general differences of joint creativity, it leaves unclear how the dynamics unfold during collaboration, and how collaborative roles emerge.

Emotional contagion and alignment Collaborative processes, such as joint storytelling, unfold through turn-by-turn negotiations of emotional tone, and the evolving narrative influence between collaborators. In human-LLM collaboration, examining these interpersonal dimensions of collaboration matters in determining whether LLMs function as passive elaborators, equal co-authors, or adaptive partners. In conversational paradigms, interlocutors mimic and synchronize behavior, often referred to as emotional contagion. (Hatfield et al., 1993). Varni et al. (2017) proposed a computational framework to quantify emotional contagion between humans in conversation. They modeled the alignment as matches between interlocutor time series polarity states (positive, negative, neutral), using valence and facial expressions as metrics. Poria et al. (2019) highlighted how challenges to model emotion recognition in text can improve conversational AI-systems, and emphasized how interlocutors exert emotional influence both on their counterpart and on themselves. Studies on human-LLM dialogues have shown convergence in speech rate (Li et al., 2025) and linguistic convergence (Wilkenfeld et al., 2022).

Sentiment analysis in digital humanities Computational sentiment analysis has become a tool for studying emotional dynamics in literary texts. Reagan et al. (2016) used a sliding-window computational framework to identify six dominant emotional arcs, while other approaches modeled progression (Hu et al., 2021) or mood (Öhman and Rossi, 2022). In domain-specific settings, Feldkamp et al. (2024) compared dictionary-based and transformer-based sentiment tools for Danish literature, while Bizzoni and Feldkamp (2023) evaluated similar methods on Hemingway's works, highlighting the challenges of applying general-purpose sentiment models to literary text. Lyngbaek et al. (2025) introduced concept vector projection as a method for deriving continuous sentiment scores tailored to literary and multilingual contexts, which we adopt in the present study. Our work extends this line from single-author literary analysis to interactive, multi-authored narrative frameworks.

Influence and agency On agency and influence, Barron et al. (2018) presented an approach to evaluate the influence of novel ideas on subsequent discourse, introducing the measures of novelty, transience, and resonance. They used a corpus of speeches during the French Revolution, and, through KL-divergence of topic distributions, captured novelty as a speech's deviation from previous discourse, transience as the degree to which that deviation failed to persist, and resonance as the difference between the two, expressing the deviation that is retained in future discourse. Bergey and

Corpus metric	Value
Total stories (n)	87
Total participants (n)	87
Interactions per story	10
Mean LLM word count	29
Mean user word count	26
Total interactions (n)	870

Table 1: Descriptive statistics for the co-writing corpus.

DeDeo (2024) focused on information propagation in dialogue, and used token-level LLM surprisal scores to model conversational information flow between interlocutors. We build on both approaches, applying surprisal-based novelty, transience, and resonance to human–LLM co-writing to evaluate whose contributions exhibit greater narrative influence.

3 A Corpus of Narrative Co-Writing

3.1 Task and experiment design

We designed a controlled collaborative storytelling task that isolates turn-based interaction dynamics. The task models co-writing as a dyadic exchange in which a human participant and one of four LLMs jointly construct a narrative through alternating contributions[1]. Each agent, human or LLM, contributed to the story from where the other left off, for a total of 10 interaction steps, each comprising one user turn and one LLM turn (U_t, A_t), yielding 20 turns per story. Figure 1 illustrates the task flow.

The resulting dataset is structured into four levels of analysis:

1. **Token level (model level):** individual tokens w.
2. **Turn level:** a single agent contribution at turn t, denoted I_t, which can be further specified as either a user turn (U_t) or an LLM turn (A_t).
3. **Interaction level:** paired contributions at the same turn t, represented as (U_t, A_t).
4. **Story level:** a sequence of 10 interactions for session i, denoted S_i.

The following instructions were provided to both participants and the LLM:

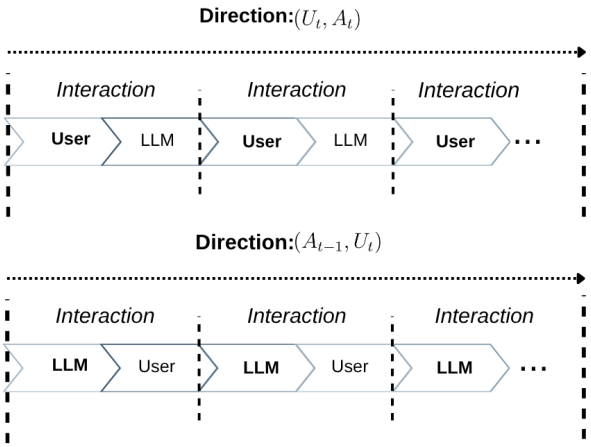


Figure 1: Dyadic task flow, visualizing how the participant and LLM take turns adding to the narrative.

> *"You are an author taking part in a collaborative storytelling activity with another author. Together, you will create a story by taking turns adding to it. Your goal is to continue from where your partner has left off. If there is no story, please begin the story. You have 10 interactions to write the story. Your input may be slightly truncated by a random number of characters."*

The study comprised a total of 91 participants, with a balanced gender distribution (male = 44, female = 46, other = 1) and an age distribution of $M = 24$ years ($SD = 10.4$). This yielded a corpus of 91 stories (910 interactions). After preprocessing exclusions, the corpus comprised 87 stories[2] . Participants were recruited both in a university context in Denmark, through research and student communities, and through public invitations in various online communities. The sample was therefore composed primarily of young adults, fluent in English as a second language, with backgrounds from the humanities, social sciences, and STEM. Each story consists of alternating human and model contributions of approximately 25–30 words. Examples of the stories and the metrics of each turn are in Table 2.

4 Methods

Because the paper combines sentiment modeling, directional alignment, and information-theoretic influence metrics, we summarize the logic of the

[1]Participants were randomly assigned to one of the following LLMs as writing partners: gpt-4.1-2025-04-14, claude-sonnet-4-5-20250929, Llama-3.3-70B-Instruct, or Qwen2.5-72B-Instruct.

[2]The corpus will be made publicly available upon publication.

User (excerpt)	AI (excerpt)	User val.	AI val.	User tok.	AI tok.	User nov.	AI nov.
This is the story of a small-town girl...	She feels trapped in the mundane routine...	−0.241	−0.327	28	31	−0.547	−1.545
Fueled by her reading habit, by books like...	...she's determined to find a way to break free...	0.361	0.667	28	30	−1.389	−2.038
...a more fulfilling future. Everything from...	...though she quickly dismisses the latter...	0.804	−0.786	25	30	−1.336	−2.374
mild-mannered mother, Puritanical father...	...traveling the world as a missionary's wife...	−0.165	0.066	21	30	−0.376	−3.067
smuggled out of Pakistan in the hollow tubing...	...leaving the small town girl to wonder...	−0.673	−0.032	24	31	−1.209	−2.633

Table 2: Example interaction-level features (valence, token counts, novelty) for a single story.

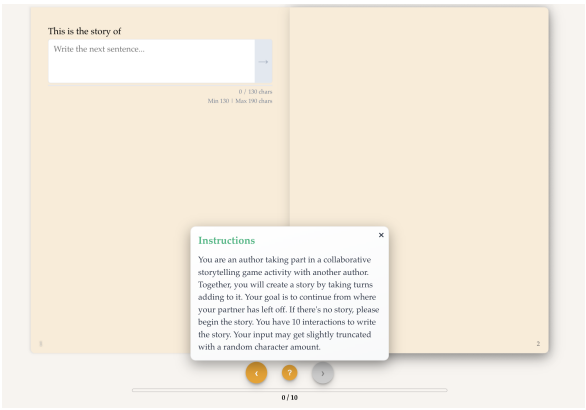


Figure 2: The user interface of the platform used in the experiment with their instructions.

approach before introducing the formal definitions. Valence measures affective tone. Directional alignment tests whether one agent's tone predicts the other's subsequent tone. Novelty measures how much a turn departs from prior context. Resonance measures whether that departure remains predictive of what follows.

4.1 Preprocessing

To ensure accurate downstream analysis, a cleaning pipeline was applied to all text data. First, incomplete stories, classified as stories with fewer than 10 total interactions, were excluded. User-generated text was then spell-corrected using GPT-4o-mini (temperature = 0) via API, with instructions to preserve each sentence as it was, correcting only spelling errors. LLM-generated text required no cleaning as it contained no spelling errors. To ensure corrections did not substantially alter the original text, the Levenshtein edit distance was calculated between original and corrected versions, and stories with user text exceeding an edit distance of 70 were excluded as these cases corresponded to users writing gibberish rather than coherent text. In total, four stories were removed, yielding a final analytical sample of 87 stories. Manual post-hoc inspection of spell-corrected user text confirmed minimal syntactic and semantic changes were made in the cleaning process.

4.2 Valence

First, we investigate the general difference in emotional tone between agents. To get a continuous metric of valence, we compute a scalar valence score for each agent turn using concept vector projection. Using the multilingual sentence-embedding model `mpnet-base-v2`, each turn is embedded as a vector representation $\mathbf{e}_t^{(a)} \in \mathbb{R}^d$. The model was chosen for its strong performance across varying registers and vocabulary. We then compute valence as the scalar projection onto a sentiment concept vector $\mathbf{c}$, constructed as the normalized difference between mean embeddings of positive and negative seed words, following Lyngbaek et al. (2025):

$$v_t^{(a)} = \frac{\mathbf{e}_t^{(a)} \cdot \mathbf{c}}{\|\mathbf{c}\|}. \quad (1)$$

Intuitively, this projection measures how far each turn's embedding lies along the positive–negative sentiment axis in the embedding space. Higher values indicate more positively-toned language, while lower values indicate more negative tone. To evaluate baseline differences in emotional tone between agents, we fitted a linear mixed-effects model:

$$v_{it} = \beta_0 + \beta_1 \text{Agent}_{it} + (1 \mid \text{Story}_i), \quad (2)$$

where v_{it} denotes the valence score for turn t in story i, and Agent_{it} is a binary indicator distinguishing User from LLM turns. Random intercepts for $Story$ account for the non-independence of turns within stories. This approach to valence estimation allows a stronger domain-fit than standard sentiment analysis methods (Lyngbaek et al., 2025), as it can tailor the concept vector to the affective vocabulary most relevant to the data at hand.

4.3 Directional Sentiment Alignment

To measure directionality of alignment as an expression of affective coordination between agents, we model the relationship between predictor-turn valence and response-turn valence under two directional pairing rules:

1. **User→LLM**: (U_t, A_t)
2. **LLM→User** : (A_{t-1}, U_t)

Let $v_{it}^{(X)}$ denote the predictor valence and $v_{it}^{(Y)}$ the response valence under a directional pairing rule in story i. Directional alignment was estimated using the following linear mixed-effects model:

$$\begin{aligned} v_{it}^{(Y)} = \beta_0 + \beta_1 v_{it}^{(X)} + \beta_2 D_{it} \\ + \beta_3 \left(v_{it}^{(X)} \times D_{it} \right) + (1 \mid \text{Story}_i) \end{aligned} \tag{3}$$

where D_{it} is a binary indicator of alignment direction. Here, β_1 expresses baseline alignment strength, and the interaction term β_3 captures directional asymmetry in adaptation of emotional tone.

4.4 Narrative influence

4.4.1 Surprisal

The surprisal scores for turn-level text are computed using the open-source model Llama-3.1-8B-Instruct as the mean negative log-probability of the tokens in a turn given a preceding context. The single surprisal score of a token is thus calculated as the surprisal score of the token w_j given all preceding tokens in the context window plus the preceding tokens within the turn.

$$s(w_j) = -\log_2 p(w_j \mid w_{<j}) \tag{4}$$

where j indexes tokens within a turn. The mean surprisal score of a turn is then expressed as:

$$\bar{s}(T) = \frac{1}{n} \sum_{j=1}^{n} s(w_j) \tag{5}$$

We then define novelty and transience as the difference between conditional and unconditional surprisal. To calculate the negative pointwise mutual information (PMI) of a turn, we subtract the surprisal score of the turn without context from the surprisal score of the turn with context, resulting in a negative score of contextual facilitation, where more negative values indicate increased predictive benefit from the context. For novelty, the score with context is calculated using all preceding tokens in the story as context, denoted $w_{<t}$, and the surprisal score without context is calculated using the model's beginning-of-sequence token, BOS, providing an unconditional baseline.

$$\text{Novelty}_t = \bar{s}(T_t \mid w_{<t}) - \bar{s}(T_t \mid \text{BOS}) \tag{6}$$

For computing transience we measure the information gain of the current turn on the full subsequent partner turn, denoted F_t. For user turns F_t is the immediate LLM response, and for LLM turns F_t is the next user contribution.

$$\text{Transience}_t = \bar{s}(F_t \mid T_t) - \bar{s}(F_t \mid \text{BOS}) \tag{7}$$

Finally, resonance was computed by subtracting transience from novelty.

$$\text{Resonance}_t = \text{Novelty}_t - \text{Transience}_t \tag{8}$$

4.4.2 Influence Modeling

For modeling the relationship between novelty and resonance, we fit a linear mixed-effects model, that evaluates narrative influence by having novelty predict resonance with agent as an interaction effect:

$$\begin{aligned} \text{Resonance}_{it} = \beta_0 + \beta_1 \text{Novelty}_{it} + \beta_2 \text{Agent}_{it} \\ + \beta_3 \text{Novelty}_{it} \text{Agent}_{it} + b_{0i} \\ + \varepsilon_{it} \end{aligned} \tag{9}$$

Supporting this, we also fit a similar mixed effects model, having novelty predict transience with agent as an interaction effect:

$$\begin{aligned} \text{Transience}_{it} = \beta_0 + \beta_1 \text{Novelty}_{it} + \beta_2 \text{Agent}_{it} \\ + \beta_3 \text{Novelty}_{it}\, \text{Agent}_{it} + b_{0i} \\ + \varepsilon_{it} \end{aligned} \tag{10}$$

Because resonance is defined as novelty minus transience, the resonance model is algebraically linked to the transience model. We report both for interpretive transparency, as the transience model isolates the uptake mechanism, while the resonance model captures net narrative influence.

5 Results

5.1 Valence distributions differ (RQ1)

The distributions of valence showed lower mean user valence ($M = -0.089$, $SD = 0.403$) compared to LLM valence ($M = 0.004$, $SD = 0.397$).

RQ	Unit of analysis	Measure(s)	Model / Tool
RQ1	Turn (U_t, A_t)	Valence distributions (sentiment scores per turn)	`paraphrase-multilingual-mpnet-base-v2` + sentiment concept vector projection
RQ2	Directional pairings: (U_t, A_t) and (A_{t-1}, U_t) Self-alignment: (U_{t-1}, U_t), (A_{t-1}, A_t)	Directional alignment: (i) valence coupling (correlation), (ii) semantic alignment (cosine similarity under pairing rules)	`mpnet-base-v2` (valence) `QZhou-Embedding` (semantic)
RQ3	Token w_j with conditional context $p(w_j \mid w_{<j})$ (aggregated to turn level)	Surprisal-based novelty, transience and resonance (informational deviation + uptake)	`Llama-3.1-8B-Instruct` (token probabilities → surprisal)

Table 3: Summary of the employed unit of analysis, measures and models for each research question.

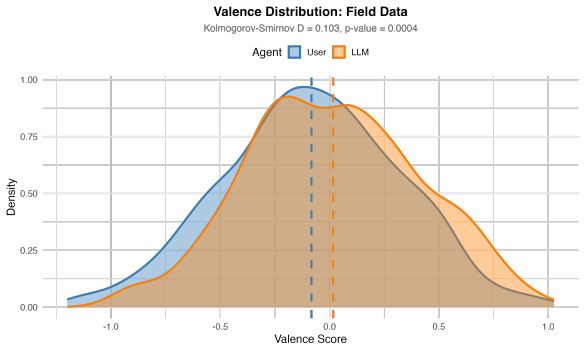


Figure 3: Distribution of turn-level valence scores by agent. Higher values indicate more positive emotional tone. LLM turns show a slightly more positive baseline than user turns., while both distributions overlap

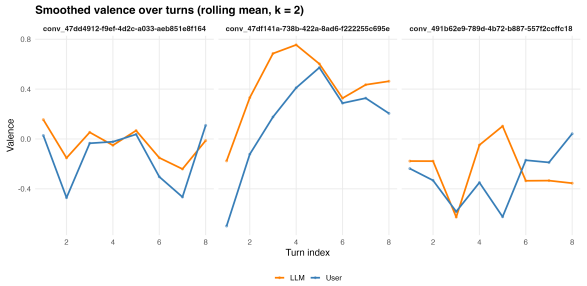


Figure 4: Example of the valence trajectories through three different stories, visualizing baseline differences and temporal alignment.

A linear mixed-effects model predicting valence from agent type showed a significantly higher mean valence for LLM turns ($\beta = 0.093$, $p < .001$). This coefficient reflects the estimated mean difference between LLM and user turns. Figure 3 shows the valence distributions by agent. An example of valence trajectories between agents throughout three stories can be seen in Figure 4.

5.2 Alignment is directional (RQ2)

The emotional alignment between human and LLM co-writers was statistically significant overall ($\beta = 0.16$, $p < .001$), confirming that valence coordination occurs from turn to turn. However, this alignment was asymmetric. The model showed a strong positive relationship in the (U_t, A_t) direction, with $\beta_1 = 0.232$ ($SE = 0.038$, $p < .001$) compared to the significantly weaker (A_{t-1}, U_t) direction, with a significant interaction contrast ($\Delta\beta = -0.141$, $SE = 0.055$, $p = .010$). Figure 5 visualizes the fitted relationships by direction.

This was supported by a Pearson correlation analysis. Mean correlation was higher for (U_t, A_t) ($M = 0.232$, $SD = 0.389$) than for (A_{t-1}, U_t) ($M = 0.091$, $SD = 0.395$). One-sample tests against zero indicated that both correlations were reliably above zero, with a stronger effect for (U_t, A_t). For (U_t, A_t): $t = 5.55$, $p < .001$; for (A_{t-1}, U_t): $t = 2.14$, $p = .035$.

5.3 Narrative influence is asymmetric (RQ3)

User turns showed higher novelty than LLM turns (User: $M = -1.418$, $SD = 0.713$; LLM: $M = -2.077$, $SD = 0.647$). User turns also showed slightly higher transience (User: $M = -1.034$, $SD = 0.441$; LLM: $M = -0.676$, $SD = 0.454$). As a result, mean resonance was higher for user turns (User: $M = -0.384$, $SD = 0.774$) than for LLM turns (LLM: $M = -1.401$, $SD = 0.684$). Since novelty is defined as PMI, it is negative when the context reduces surprisal. Figure 6 shows novelty distributions by agent.

Both agents showed a strong positive novelty-resonance relationship, indicating that more surprising turns tend to stick. In the resonance model, novelty had a strong positive effect for users ($\beta_1 = 0.941$, $p < .001$). The novelty×agent interac-

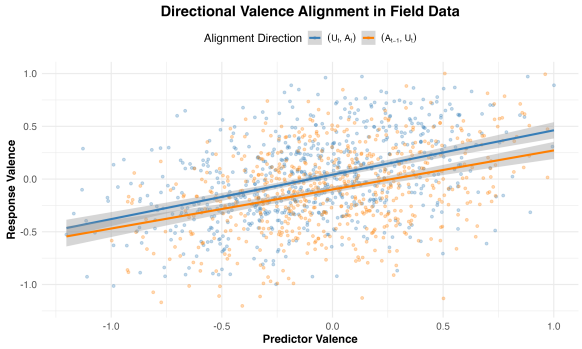


Figure 5: Directional sentiment alignment between agents. Each point is a turn pair (interaction). Lines show fitted slopes of response valence on predictor valence under the two directional pairings. The steeper slope for LLM responding to User (U_t, A_t), compared to User responding to LLM (A_{t-1}, U_t), indicates that LLM responses adapt more strongly to preceding user valence than vice versa.

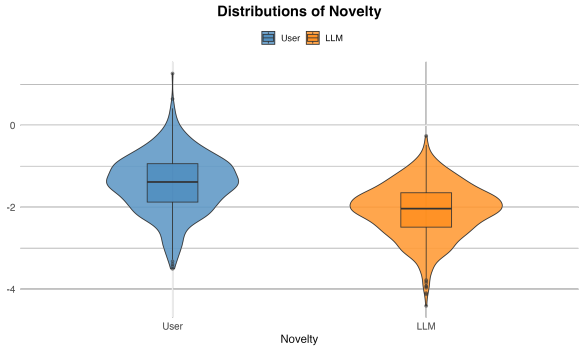


Figure 6: Distributions of surprisal (novelty) for each agent, defined as the surprisal of a turn relative to prior context. User turns exhibit general higher surprisal than LLM turns, indicating more novel contributions.

tion was significant ($\beta_3 = -0.105$, $SE = 0.034$, $p = .0019$), yielding a weaker LLM slope of 0.837, indicating a slightly weaker innovation bias for LLMs than users. This relationship can be seen in Figure 7.

6 Discussion

We investigated agent alignment and narrative agency in human–LLM co-writing, and examined affective and semantic adaptation through embedding-based analyses and information propagation through different measures of innovation and influence. Our findings reveal a consistent asymmetry in collaborative dynamics: while both humans and LLMs exhibit alignment and coordination, humans play a disproportionate role in maintaining emotional autonomy, introducing narrative novelty, and shaping subsequent developments.

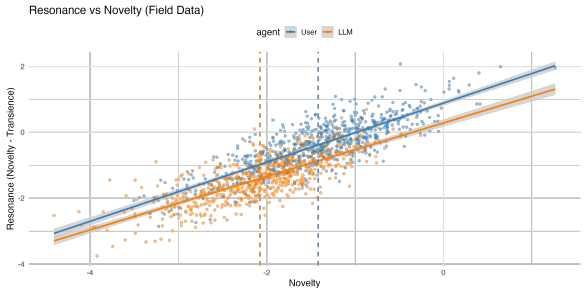


Figure 7: Resonance as a function of novelty. Resonance (y-axis) measures how much a turn's surprising content persists in the subsequent turn. Novelty (x-axis) measures its surprisal relative to prior context. Both agents show a positive novelty–resonance relationship, but the slope is steeper for users (blue), indicating that surprising human contributions exert more influence on the unfolding narrative.

The emotional asymmetry in valence distributions quantifies a fundamental characteristic of LLMs: their responses tend to drift towards a positive baseline. This could be attributed to the removal of strong negative and profane data during training and fine-tuning of the models. While this has been established for isolated LLM output (Muñoz-Ortiz et al., 2024; Zanotto and Aroyehun, 2024), our paradigm tests the autonomy of LLM positivity bias, by having it iteratively react to varying input from human interlocutors, which might force the sentiment of the story into a negative drift. This is equally the case for human partners having to react to positive LLM contributions. This could partly explain the relatively small difference observed in valence distributions, as each partner is faced with a decision to converge emotionally with the counterpart.

The analysis of directional emotional adaptation elaborates on the difference in valence distributions. We found significant bidirectional emotional alignment between co-writers, confirming that human–LLM collaboration involves affective coordination through adaptation of emotional tone. However, the alignment was directional: LLM responses tracked the valence of preceding human turns more strongly than humans tracked LLM affect. This asymmetry indicates that LLMs are more responsive to human emotional cues and suggests that humans maintain a stronger level of emotional autonomy in the co-writing process, resisting immediate emotional adaptation to the LLM counterpart. This is additionally highlighted by the significantly stronger semantic self-alignment of users, in which they

Result	User / $A_{t-1}\rightarrow U_t$	LLM / $U_t\rightarrow A_t$	Test	p
Mean valence	-0.089	0.004	Agent difference (LMM)	$p < .001$
Alignment slope (β)	0.091	0.232	Slope difference (interaction)	$p = .010$
Alignment frequency (%)	49.2	58.9	Proportion difference	$p = .001$
Mean align. duration (turns)	1.836	2.189	Wilcoxon rank-sum	$p = .039$
Resonance $\sim$ novelty slope	0.941	0.837	Slope difference (interaction)	$p = .0019$
Transience $\sim$ novelty slope	0.059	0.164	Slope difference (interaction)	$p = .0019$

Table 4: Key results summary. For directional metrics (rows 2–5), the first column reports the $A_{t-1}\rightarrow U_t$ (LLM-to-User) direction; the second reports $U_t\rightarrow A_t$ (User-to-LLM). For non-directional metrics (rows 1, 6–7), values are per agent.

converge more with their own preceding input than the LLM responses, adopting less of the semantic profile from the counterpart (see Appendix A).

Alignment persistence over time revealed a similar pattern: Alignment occurred more frequently in Human→LLM transitions, while also retaining longer streaks over time. It seems that LLMs both maintain stronger local affective adaptation, while also sustaining longer-term emotional trajectories (see Appendix A).

Across analyses, human contributions were both more novel and yielded higher resonance than LLMs', indicating a greater capacity to introduce narrative elements that persist in the unfolding story. In contrast, LLM contributions tended to elaborate on existing material rather than introduce new semantic directions. This pattern suggests an *initiative asymmetry*, in which humans function as primary drivers of narrative innovation, while LLMs act as adaptive amplifiers that reinforce and extend human-introduced content.[3]

The positive linear relationship between novelty and resonance characterizes the creative, frictionless nature of the writing task where interlocutors accept and take up the preceding contributions of the counterpart, regardless of the surprising character of this contribution given the prior context. Similarly to the previous findings, this pattern is asymmetric between agents. The effect of very surprising input on future discourse is stronger for users, indicating their writing privilege, where users' novel contributions have more influence on the future discourse, compared to equally novel LLM contributions. Supplementary analyses indicate consistent behavior across the four model subgroups (see Appendix A). These asymmetries seem to mirror a complementary division of labor: humans provide innovation, on which LLMs elaborate, reminiscent of improvisational settings in which one participant introduces motifs and the other stabilizes and develops them.

These findings complicate the view of LLMs as either independent creative agents or neutral writing tools. The elaboration-adaptation pattern indicates that LLMs may be most valuable as elaborative partners, while also helping explain homogenization effects observed in prior work (Anderson et al., 2024). If the LLM collaborator role primarily develops existing material and adapts, then narrative variance may depend disproportionately on the human side of the dyad. Thus, the creative value of LLM co-writing may lie less in autonomous narrative development and more in the ability for users to surface and select among divergent narrative branches. Read together, the findings position LLMs less as creative co-authors and more as adaptable stylistic amplifiers.

The applicability of the directional metrics we introduce extends beyond the presented corpus. These approaches may be further applicable as general tools for modeling interlocutor dynamics in conversation, multi-authored text, or political discourse.

7 Conclusion and Future Work

This study examined alignment dynamics and narrative agency in human–LLM collaborative storytelling. Our findings reveal clear patterns that characterize human–LLM co-writing: LLMs **mirror** the emotional tone established by human contributors, **but shift** it toward a more positive baseline. At the same time, LLMs tend to **elaborate** on human-introduced ideas **rather than introduce** novel narrative elements, and have less influence

[3]Importantly, although LLMs exhibited a stronger tendency to follow and elaborate on human contributions, both agents alternated between introducing novelty, transience, and aligning with prior turns, indicating that LLMs are adaptive - rather than passive - agents within collaborative dynamics.

on the subsequent narrative.

In other words, they overall follow the semantics introduced by human writers, but tend to bring a higher valence “bias" to the story. This behavior suggests that LLMs function effectively as “following” collaborators that sustain coherence and reinforce emerging storylines. Human participants exhibited greater semantic novelty and stronger influence on subsequent narrative developments, indicating a primary role in introducing new directions and maintaining emotional autonomy. Together, these results point to a complementary division of labor: humans drive innovation and narrative, and LLMs provide adaptive elaboration and strong affective alignment. These findings have implications for the design of human–AI co-creative systems. Interfaces and interaction paradigms may benefit from supporting human control over narrative direction while using LLM strengths in elaboration, stylistic variation, and coherence maintenance. Understanding this complementary dynamic can inform the development of tools that enhance creativity without diminishing human agency.

Future work should extend this paradigm by incorporating human–human baselines to better contextualize mixed collaboration dynamics; exploring longer interaction sequences to capture extended narrative evolution; integrating context-aware sentiment models to account for broader emotional trajectories; and modeling differences between the four models. Investigating how different model architectures, prompting strategies, and genre constraints affect alignment and agency may further clarify the conditions that make human–LLM collaboration most effective.

Limitations

Several limitations should be considered when interpreting the results. First, sentiment scores are computed at the turn level and do not incorporate broader narrative context, potentially overlooking longer-range emotional dynamics. Second, surprisal is estimated with respect to the language model used in the analysis and may not fully reflect human perceptions of novelty. Third, the study does not include a human–human co-writing condition, which limits direct comparison between mixed and purely human collaborations. Future work should include this as a baseline to separate LLM-specific effects from general collaborative dynamics. Additionally, the experimental setup introduces asymmetries in both interaction experience and language processing capabilities between human participants and LLMs, which may influence the observed alignment patterns. Finally, the preprocessing pipeline, including spell correction, may introduce a degree of normalization toward LLM-like text.

Ethics Statement

This study was conducted according to ethical guidelines. All participants provided their consent prior to participation and were informed that their text input would be used for research purposes. Only anonymized demographic data (age, gender) were collected. Participants were free to withdraw at any time. Participants were aware that their co-writer was an LLM. The dataset consists of English-language creative fiction produced by adult participants recruited from a university context in Denmark.

Acknowledgments

This research was supported by the Danish National Research Foundation, grant number DNRF193.

References

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Claire Augusta Bergey and Simon DeDeo. 2024. From" um" to" yeah": Producing, predicting, and regulating information flow in human conversation. *arXiv preprint arXiv:2403.08890*.

Yuri Bizzoni and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *NLP4DH*.

Xusen Cheng and Lulu Zhang. 2025. Inspiration booster or creative fixation? the dual mechanisms of llms in shaping individual creativity in tasks of different complexity. *Humanities and Social Sciences Communications*, 12(1):1–10.

Elizabeth Clark and Noah A Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575.

Pascale Feldkamp, Jan Kostkan, Ea Overgaard, Mia Jacobsen, and Yuri Bizzoni. 2024. Comparing tools for sentiment analysis of danish literature from hymns to fairy tales: Low-resource language and domain challenges. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 186–199.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in never let me go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Qixin Li, Haocheng Lu, and Gaowu Wang. 2025. When ai speaks, do we follow? phonetic entrainment in human-ai dialogues. In *National Conference on Man-Machine Speech Communication*, pages 167–186. Springer.

Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer Nielbo, and Kenneth Enevoldsen. 2025. Continuous sentiment scores for literary and multilingual contexts. *arXiv preprint arXiv:2508.14620*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Emily Öhman and Riikka H Rossi. 2022. Computational exploration of the origin of mood in literary texts. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 8–14.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):31.

Melissa Roemmele and Andrew S Gordon. 2015. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*, pages 81–92. Springer.

Min Tang, Sebastian Hofreiter, Christian H Werner, Aleksandra Zielińska, and Maciej Karwowski. 2025. "who" is the best creative thinking partner? an experimental investigation of human–human, human–internet, and human–ai co-creation. *The Journal of Creative Behavior*, 59(3):e1519.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Giovanna Varni, Isabelle Hupont, Chloe Clavel, and Mohamed Chetouani. 2017. Computational study of primitive emotional contagion in dyadic interactions. *IEEE Transactions on Affective Computing*, 11(2):258–271.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. " it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–26.

J Nan Wilkenfeld, Bei Yan, Jujun Huang, Guirong Luo, and Kristina Algas. 2022. "ai love you": Linguistic convergence in human-chatbot relationship development. In *Academy of Management Proceedings*, volume 2022, page 17063. Academy of Management Briarcliff Manor, NY 10510.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv preprint arXiv:2412.03025*.

Eric B Zhou, Dokyun Lee, and Bin Gu. 2025. Who expands the human creative frontier with generative ai: Hive minds or masterminds? *Science Advances*, 11(36):eadu5800.

A Supplementary Analyses

A.1 Duration of alignment

To assess stability of directional alignment, we computed (1) the overall frequency of turns classified as aligned per direction and (2) the duration of consecutive alignment streaks. Figure 8 summarizes alignment frequency by direction. Alignment frequency differed between directions: alignment occurred on 49.2% of turns in the LLM→User direction and 58.9% of turns in the User→LLM direction (two-sample test of proportions, $p = 0.001324$).

Figure 9 shows the persistence of alignment streaks (survival-style retention curves). Streaks were longer in the User→LLM direction (mean

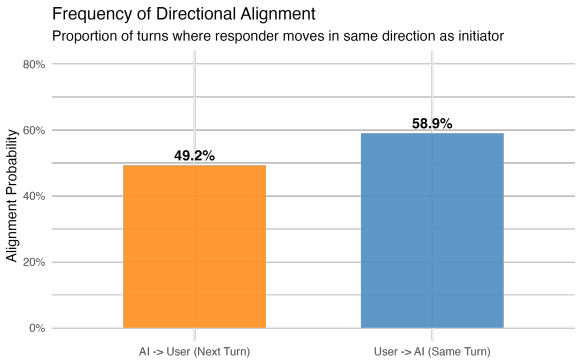


Figure 8: Frequency of directional alignment, shown as the proportion of turns where the responder moves in the same direction as the initiator.

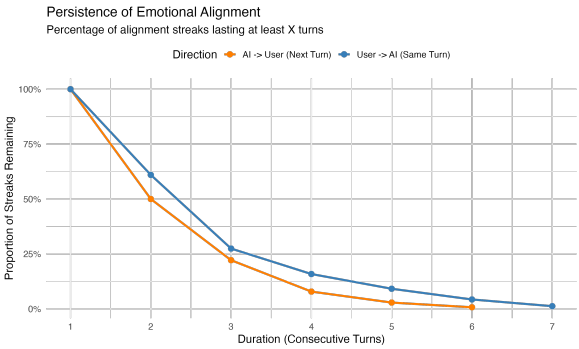


Figure 9: Persistence of emotional alignment. The curves show the percentage of alignment streaks lasting at least X consecutive turns, by direction.

$= 2.19$, median $= 2$, max $= 7$) than in the LLM→User direction (mean $= 1.84$, median $= 1.5$, max $= 6$). A Wilcoxon rank-sum test indicated a reliable difference in duration distributions ($p = 0.03943$).

A.2 Self alignment

As a complementary analysis, we modeled self-alignment as the cosine similarity between an agent's current turn and their own preceding turn, (U_{t-1}, U_t) and (A_{t-1}, A_t).

For semantic self-alignment we compared distributions of cosine similarity between (U_{t-1}, U_t) denoting semantic self-alignment of users and (U_t, A_{t-1}) denoting semantic alignment of users with the preceding LLM turn. This was similarly modeled for LLM self-alignment. User self-alignment (U_{t-1}, U_t) showed generally higher cosine similarity than user interlocutor alignment (U_t, A_{t-1}): the mixed-effects model showed a significant difference ($\beta = -0.012$, $SE = 0.004$, $p = .005$), indicating that users align slightly more with their own preceding turn than with the preceding LLM turn. For LLM self-alignment, no significant difference was observed ($\beta = 0.005$, $SE = 0.004$, $p = .131$).

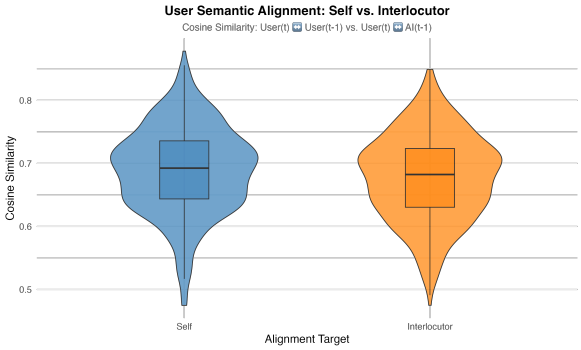


Figure 10: Semantic self-alignment of users (blue) compared to alignment with the interlocutor (orange). Measured as cosine similarity between contributions.

LLM type	n	Mean ρ	SD	Mean AI val.
GPT-4.1	18	0.301	0.353	0.085
Claude	20	0.198	0.364	−0.219
Llama 3.3	26	0.255	0.405	−0.015
Qwen 2.5	23	0.180	0.432	0.154

Table 5: Per-story same-turn alignment (ρ) and mean AI valence by LLM type. ANOVA: $F(3, 83) = 0.40$, $p = .754$.

A.3 Alignment and influence by LLM-types

To validate the aggregation of the four different LLMs in the analysis, we conducted inter-model analysis on both alignment (U_t, A_t) and LLM novelty-resonance models. In the analysis of alignment, models showed consistent patterns of sentiment alignment, but with Claude exhibiting a significantly lower intercept than the reference model ($\beta = -0.315$, $SE = 0.062$, $p < .001$), indicating lower mean AI valence in the User→LLM direction, as seen in Figure 11. Alignment slopes did not significantly differ across models. A one-way ANOVA on per-story same-turn correlations confirmed no significant effect of LLM type ($F(3, 83) = 0.40$, $p = .754$). In the novelty-resonance analysis shown in Figure 12, all models showed similar positive novelty-resonance relationships.

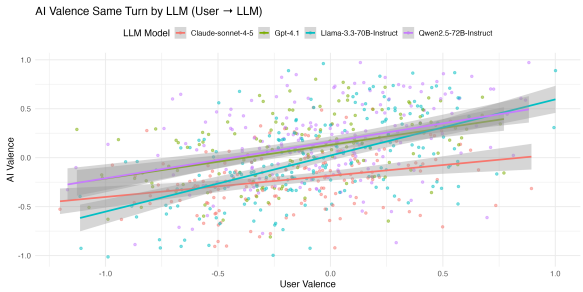


Figure 11: Directional alignment between (U_t, A_t) focusing on LLM alignment, with regression slopes by LLM type.

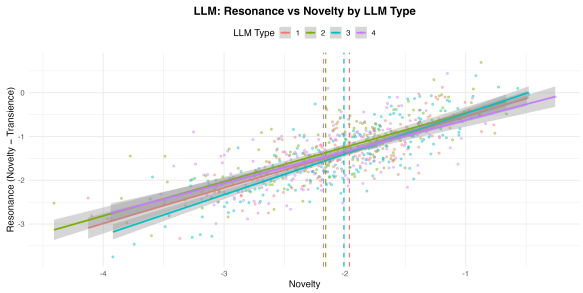


Figure 12: Resonance as a function of novelty for LLM contributions, with color-coded regression slopes for each LLM-type. Dotted lines indicate mean novelty for each model.

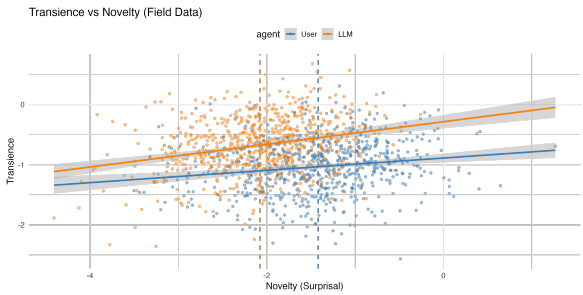


Figure 13: Transience as a function of novelty, with regression lines by agent.

Bias Mitigation in Hiring-Related NLP: Interactions Between Masking, Rewriting, and Adversarial Debiasing

Alexandre Puttick
Bern University of Applied Sciences
Technik und Informatik
Quellgasse 21, 2501 Biel, Switzerland
alexandre.puttick@bfh.ch

Rami El-Wazzi
Bern University of Applied Sciences
Technik und Informatik
Quellgasse 21, 2501 Biel, Switzerland
rami.el-wazzi@students.bfh.ch

Abstract

AI-driven language technologies are increasingly used in hiring, but they may encode and reproduce harmful social stereotypes. Prior work often studies bias mitigation methods in isolation and outside realistic application settings. We examine the combined effects of data-level and model-level debiasing in a hiring-related context, using Norwegian-language academic bios and a proxy STEM/non-STEM classification task. Specifically, we study masking sensitive information, GenWriter-based rewrites (Soundararajan and Delany, 2025), and adversarial debiasing (Han et al., 2021, 2022). We evaluate these interventions using downstream task performance, group fairness metrics, intrinsic bias tests based on WEAT (Caliskan et al., 2017), and measures of gender leakage from hidden representations. We find that combining masking, GenWriter rewrites, and adversarial debiasing substantially reduces gender leakage while maintaining or improving downstream performance. However, effects on fairness gaps and intrinsic bias are mixed, underscoring the need for downstream, context-sensitive evaluation of bias mitigation methods in hiring-related NLP.

1 Introduction

AI-driven language technologies are increasingly used in hiring, where they may influence screening, ranking, and evaluation. These systems can encode and reproduce harmful social stereotypes, and a large body of literature has proposed methods for detecting and mitigating such biases. Common approaches are often grouped into *pre-processing* methods, such as counterfactual data augmentation (Zmigrod et al., 2019), *in-processing* methods, such as fairness-constrained optimization (Hardt et al., 2016), and *post-processing* methods, such as re-ranking (Zehlike et al., 2017) or embedding debiasing (Bolukbasi et al., 2016; Liang et al., 2020).

Despite this progress, limitations remain that are especially relevant for hiring-related NLP. First, many studies are weakly connected to concrete application contexts and to the ethical goals that motivate fairness interventions. Second, non-English languages remain underrepresented. Third, debiasing methods are often evaluated in isolation, even though real systems may combine interventions at multiple stages of the pipeline.

This paper addresses these gaps in a hiring-related setting as part of the EU Horizon-funded BIAS Project[1]. We study Norwegian-language academic bios from the Norwegian University of Science and Technology (NTNU) and use a proxy task of classifying bios as STEM or non-STEM. Although this is not a direct hiring task, it captures a recruitment-relevant scenario in which textual candidate information is mapped to a consequential category. Our aim is to study how different interventions affect not only downstream task performance, but also fairness-relevant properties of the model.

We take a downstream-first approach to debiasing. At the data level, we study *masking* of names and gendered pronouns and *GenWriter* rewrites (Soundararajan and Delany, 2025). Masking removes explicit gender markers, while GenWriter aims to reduce stylistic and other indirect gender signals through case-based rewriting. At the model level, we study *adversarial debiasing* (Zhang et al., 2018; Han et al., 2021, 2022), which is designed to reduce sensitive-attribute information in model representations while preserving task-relevant information.

Our analysis distinguishes between several related but non-identical notions. We treat *procedural fairness* as limiting the model's reliance on sensitive attributes and their proxies, which we assess through gender leakage from hidden representations. We treat *substantive fairness* as parity

[1] https://www.biasproject.eu/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 202–213
July 4, 2026 ©2026 Association for Computational Linguistics

in predictive performance across gender groups, measured using performance gaps and equality-of-odds-related metrics. We also include *intrinsic bias* tests as diagnostic measures of stereotypical associations encoded in the model, while recognizing that these need not align with downstream fairness.

The following research questions guide the paper: **(RQ1)** How do masking, GenWriter rewrites, and adversarial debiasing interact in affecting downstream task accuracy, fairness gaps, and gender leakage? **(RQ2)** Do reductions in sensitive-attribute leakage correspond to reductions in intrinsic stereotype bias? **(RQ3)** Can a downstream-oriented mitigation strategy improve procedural fairness without large performance costs?

Our contributions are:

- We present a contextualized, downstream-first study of bias mitigation for hiring-related NLP in a Norwegian-language setting, using academic bios as a proxy recruitment task.
- We evaluate the interaction between masking, GenWriter rewrites, and adversarial debiasing across downstream performance, group fairness, intrinsic bias, and representation leakage.
- We show that combined interventions substantially reduce encoded gender information while maintaining or improving downstream task performance, whereas effects on fairness gaps and intrinsic bias are more mixed.

Bias Statement

We study *diversity bias* with a focus on gender. We distinguish between *intrinsic bias*, i.e. stereotypical associations encoded in language representations, and *extrinsic bias*, i.e. disparities in downstream task behavior across groups. In this work, extrinsic bias is operationalized as differences in STEM/non-STEM classification performance between gender groups. We therefore focus on accuracy parity and equality of odds rather than demographic parity, since the task is conditioned on ground-truth labels and those labels are not strongly aligned with gender stereotypes in the dataset (Table 1). We additionally treat low sensitive-attribute leakage as a procedural fairness objective.

2 Related Work

Fairness in NLP and algorithmic decision-making. A large body of literature studies bias detection and mitigation in NLP. At the same time, several authors argue that formal fairness metrics do not fully capture the social and normative dimensions of fairness, motivating the need to select metrics that align with explicit ethical aims (Fazelpour and Lipton, 2020; Tong et al., 2026; Majumder et al., 2023; Heidari et al., 2019; Mitchell et al., 2021).

Intrinsic bias, downstream bias, and leakage. A recurring finding in the literature is that intrinsic bias measures do not straightforwardly predict downstream unfairness. Gonen and Goldberg (2019) showed that embedding debiasing can reduce bias according to particular metrics while leaving recoverable gender structure in the representations. Goldfarb-Tarrant et al. (2021) found little relationship between intrinsic measures such as WEAT and downstream bias, and Delobelle et al. (2022) showed that many intrinsic bias metrics are highly sensitive to templates, target words, and representation choices. Tokpo et al. (2023) also show that some intrinsic debiasing approaches mainly reduce bias according to specific metrics rather than in downstream behavior. These findings motivate evaluating leakage, intrinsic bias, and downstream fairness separately.

Data-level and model-level debiasing. At the data level, masking explicit sensitive markers is a common baseline, though it does not remove more indirect proxies. We therefore also consider *GenWriter* (Soundararajan and Delany, 2025), a case-based rewriting framework that aims to reduce stylistic and demographic signals while preserving task-relevant content. At the model level, we study *adversarial debiasing* (Zhang et al., 2018; Han et al., 2021, 2022), which explicitly trains a model to perform its main task while removing sensitive-attribute information from hidden representations. Our contribution is not a new debiasing method, but an empirical study of how these interventions interact across downstream performance, fairness gaps, intrinsic bias, and representation leakage.

3 Methods

We describe our methods in this section. The prompts used for all prompt-related tasks can be found in the appendix Section A.2.

3.1 Data

NTNU bios. We collected a dataset of professional academic bios from faculty pages at the

Job Title	Male	Female	STEM	non-STEM
Professor	333(219)	195(94)	313	215
Associate Professor	252(149)	252(140)	289	215
PhD Candidate	151(115)	174(115)	230	95
Totals	736(483)	621(349)	832	525

Table 1: NTNU bios dataset composition. Numbers in parentheses indicate how many individuals in each gender group are in STEM-related fields.

Norwegian University of Science and Technology (NTNU). We restricted the dataset to the three most common job titles: *professor*, *associate professor*, and *PhD candidate*. The final dataset composition is shown in Table 1.

Automatic gender label inference. To obtain gender labels for analysis, 100 professor bios were manually labeled by native speakers using names and/or pronouns. Only one sample was explicitly non-binary; because this was insufficient for meaningful modeling, we restrict the present study to binary gender labels. We trained an `NbAiLab/nb-sbert-base` classifier on the manually labeled subset and used it to infer labels for the remaining bios. Five-fold cross-validation yielded a mean test accuracy of 98% with negligible variance.

STEM/non-STEM labels. The 51 NTNU departments represented in the dataset were manually classified as STEM or non-STEM based on their primary research focus and methods, using department descriptions and curricula from NTNU sources. This yielded 27 STEM departments (52.9%) and 24 non-STEM departments (47.1%). Interdisciplinary departments were classified conservatively and included as STEM only when their primary methods were quantitative and technical.

3.2 Data-level bias mitigation

Masking. As a simple data-level intervention, we mask features that explicitly reveal gender, namely names and gendered pronouns. This does not remove indirect proxies, but it provides a minimal fairness-through-unawareness baseline.

GenWriter rewrites. We evaluate *GenWriter* (Soundararajan and Delany, 2025), a case-based rewriting pipeline designed to reduce demographic signals while preserving job-relevant content. It first splits bios into sentences, classifies them into four categories (*Education*, *Demographics*, *Work*, and *Non-Professional*), converts them into template sentences with placeholders, retrieves similar templates from the opposite gender group, and uses an LLM to fill placeholders with details from the original bio. Sample templates and rewrites are shown in Appendix A.1.

Sentence labeling and case-base construction. To train the sentence classifier used in GenWriter, we manually labeled 222 sentences from professor bios and then labeled 900 additional sentences balanced across profession, gender, and STEM/non-STEM category with `GPT 5.1`, using category definitions and three human-labeled examples per category. Agreement between GPT and human labels was 88%. Because *Non-Professional* content was rare, we added 200 synthetic *Non-Professional* samples. The resulting label distribution is shown in Table 2. We then trained a Norwegian SBERT model, obtaining 5-fold cross-validation accuracy of $81.9 \pm 1.53\%$. Since many disagreements reflected short or category-ambiguous sentences, we constructed the case base using SBERT-predicted labels.

Table 2: Human- and GPT-labeled sentence data used for sentence classification.

Category	Num. of Samples
Work	813
Education	139
Demographics	119
Non-Professional	222

3.3 Model-level bias mitigation: adversarial debiasing

Our model-level intervention is adversarial debiasing (Zhang et al., 2018; Han et al., 2021, 2022). We use it because it directly targets the removal of sensitive-attribute information from model representations while preserving information relevant to the main task. In our setting, the main task is STEM/non-STEM classification from NTNU bios. The model consists of two components:

- a *main model*, which predicts the task label;
- a *discriminator*, which predicts gender from the main model's hidden representation.

The optimization objective is:

$$\min_{M} \max_{A} \mathcal{L}_M(y, \hat{y}_M) - \lambda_{\text{adv}} \chi(g, \hat{g}_A), \quad (1)$$

where $\mathcal{L}_M$ is the main-task classification loss, χ is cross-entropy loss for the adversary, y and $\hat{y}_M$ are task labels and predictions, g and $\hat{g}_A$ are gender labels and predictions, and λ_{adv} controls the trade-off between task performance and debiasing.

3.4 Models and training details

Main model and adversary. Our main classifier is `NbAiLab/nb-bert-base`, fine-tuned end-to-end for the STEM/non-STEM task. The hidden representation passed to the adversary is the `[CLS]` token from the final hidden layer. The adversary is a 3-layer MLP. For simplicity, all reported experiments use a single discriminator, although prior work found somewhat stronger leakage reduction with multiple discriminators (Han et al., 2021).

Training setup. All models use the same 80/20 train-test split. For the downstream task, splits are stratified by gender, profession, and STEM/non-STEM label. We use Adam with a learning rate of 10^{-5} for both the main model and the discriminator. Following Han et al. (2021), training alternates between updating the main model and the adversary: each step consists of one epoch of main-task training followed by ten epochs of adversary training. We train for up to fifteen such steps with early stopping if validation accuracy does not improve for five epochs. We report the main model with the best validation performance, subject to discriminator validation accuracy remaining below 90%.

Prompted components. All prompt-based operations in the GenWriter pipeline were performed with `GPT 5.1` via the `OpenAI API` at temperature 0.7.

Leakage probing. To measure gender leakage, we train a LinearSVC probe on the hidden representations of the pretrained or fine-tuned main model and evaluate its ability to recover gender labels.

3.5 Experimental setup

We vary three dimensions in our experiments:

- *Dataset:* original bios vs. GenWriter rewrites
- *Masking:* masked vs. unmasked
- *Training:* standard fine-tuning vs. adversarial debiasing

For each condition, we fine-tune a STEM/non-STEM classifier and evaluate it along four dimensions: downstream task performance, group fairness, intrinsic bias, and gender leakage from hidden representations.

3.6 Evaluation

We distinguish between four types of evaluation.

Downstream task performance. We report classification accuracy on the STEM/non-STEM task.

Group fairness. We report accuracy gap, F1 gap, and equalized odds gap across gender groups. These quantify disparities in predictive performance and error behavior between groups.

Sensitive-attribute leakage. Leakage measures how well gender can be predicted from the model's hidden representations. Lower leakage indicates less recoverable gender information and is interpreted here as better procedural fairness.

Intrinsic bias. We evaluate intrinsic stereotypical associations using the *Word Embedding Association Test* (WEAT) (Caliskan et al., 2017) and the *Log-Probability Bias Score* (LPBS) (Kurita et al., 2019). These tests are applied to contextual representations derived from the `[CLS]` embedding. The test sets were translated and LPBS tests were adapted as described in [PROJECT CITATION].

The following WEAT-style tests were used:

Name	Description
WEAT6	career vs. family, male vs. female
WEAT7	math vs. art, male vs. female
WEAT8	science vs. art, male vs. female
GER1	male vs. female study choices
NO1	Iranian vs. Norwegian, negative vs. positive
NO2	Muslim vs. Christian, failure vs. success
NO3	Muslim vs. Christian, negative vs. positive
NO4	Iranian vs. Norwegian, failure vs. success
NO5	male vs. female, career vs. having children
NO6	single vs. in a couple, career vs. having children
NO7	Iranian vs. Norwegian, good sellers vs. not good sellers

Table 3: WEAT6–8 are from Caliskan et al. (2017); GER1 is from Kurpicz-Briki (2020). Norwegian bias tests developed within the BIAS Project from co-creation workshop discussions. Tests were designed for Norwegian culture specificity. All tests were translated by native speakers.

4 Results

4.1 Downstream performance, fairness, and leakage

Original bios. Table 4 reports downstream performance and group fairness results for models

trained on the original NTNU bios. Overall, fairness gaps were small across all settings. Standard fine-tuning on unmasked bios yielded the smallest accuracy, F1, and equality-of-odds gaps, but these differences were minor. Adversarial debiasing with unmasked data achieved the highest downstream accuracy, with somewhat larger gains for men than for women. Masking names and pronouns slightly reduced male accuracy under both standard and adversarial training.

Table 4: Original bios: STEM/non-STEM classification performance and fairness.

Method	Masked	Acc Gap	F1 Gap	EqOdds Gap	Acc F	Acc M
adversarial	No	0.012	0.009	0.012	**0.920**	**0.932**
adversarial	Yes	-0.008	-0.015	0.018	**0.920**	0.912
standard	No	**-0.000**	**-0.006**	**0.007**	0.912	0.912
standard	Yes	-0.015	-0.021	0.018	**0.920**	0.905

Table 5 shows gender leakage results for the same models. Relative to the pretrained unmasked baseline, masking alone produced a large reduction in recoverable gender information, reducing leakage from 0.912 to 0.676. Standard and adversarial fine-tuning without masking also reduced leakage substantially, to 0.728 and 0.732, respectively. The strongest reduction was obtained by combining masking with adversarial debiasing, which lowered leakage to 0.629. Thus, even when downstream fairness gaps changed only modestly, both masking and adversarial training reduced gender information in the learned representations.

Table 5: Original bios: gender leakage from hidden representations.

Method	Masked	Test Leakage	Test F1	Test ROC AUC
adversarial	No	0.732	0.729	0.801
adversarial	Yes	**0.629**	**0.627**	**0.658**
pretrained	No	0.912	0.911	0.978
pretrained	Yes	0.676	0.675	0.721
standard	No	0.728	0.727	0.815
standard	Yes	0.643	0.642	0.686

GenWriter-rewritten bios. Table 6 reports results for models trained on GenWriter-rewritten bios. Compared with the original-bio setting, rewritten bios consistently improved downstream accuracy for both groups. However, effects on fairness remained mixed. In the standard unmasked setting, the model performed slightly better for men than for women. Masking reduced this difference and yielded a near-zero accuracy gap, but at the cost of lower accuracy for both groups. Under adversarial training, the pattern reversed, with higher accuracy for women than for men; masking reduced but did not eliminate this gap. Overall, rewritten bios improved task performance more clearly than they improved group fairness.

Table 6: Rewritten bios: STEM/non-STEM classification performance and fairness.

Method	Masked	Acc Gap	F1 Gap	EqOdds Gap	Acc F	Acc M
adversarial	No	-0.039	-0.047	0.036	**0.959**	0.920
adversarial	Yes	-0.023	-0.032	0.030	0.950	0.927
standard	No	-0.017	-0.024	**0.023**	**0.959**	**0.942**
standard	Yes	**0.000**	**-0.006**	0.025	0.934	0.934

Leakage results for rewritten bios are shown in Table 7. Rewriting reduced leakage relative to the corresponding original-bio models in every setting. The strongest result again came from combining masking with adversarial debiasing, which reduced leakage to 0.566, close to chance-level prediction. Standard fine-tuning with masking also yielded lower leakage than its counterpart on the original bios. Taken together, these results suggest that GenWriter rewrites and adversarial debiasing are complementary for reducing recoverable gender information, even when improvements in group fairness metrics are less consistent.

Table 7: Rewritten bios: gender leakage from hidden representations.

Method	Masked	Test Leakage	Test F1	Test ROC AUC
adversarial	No	0.643	0.642	0.683
adversarial	Yes	**0.566**	**0.564**	**0.570**
standard	No	0.682	0.682	0.705
standard	Yes	0.628	0.628	0.664

4.2 Intrinsic bias

Table 8 summarizes all WEAT/LPBS test pairs that showed statistically significant bias in at least one model configuration.[2] For fine-tuned models, all results in this table are based on models trained on masked data.

Across models, intrinsic bias results were mixed and did not mirror the leakage results. The pretrained and standard fine-tuned models showed the fewest significant test outcomes overall, whereas other fine-tuned models often showed more significant effects. Several tests yielded broadly similar patterns across models. However, the direction and magnitude of changes varied by test. Overall, intrinsic bias measures did not improve uniformly

[2]LPBS-GER1 failed because several words were split into many subtokens, producing near-zero prior probabilities and numerical instability.

under the interventions that most strongly reduced leakage. To examine the role of masking more

Table 8: Intrinsic bias results for WEAT/LPBS test pairs. Boldface indicates significant bias ($ES > 0, p < 0.05$ or $ES < 0, p > 0.95$). All fine-tuned models were trained on masked data.

Test	Method	WEAT ES	WEAT p	LPBS ES	LPBS p
Ger1	pretrained	0.556	0.197	-	-
Ger1	adversarial	0.492	0.230	-	-
Ger1	adv. GW	**-1.518**	0.989	-	-
Ger1	standard	-0.640	0.815	-	-
Ger1	std. GW	-0.539	0.779	-	-
No2	pretrained	0.407	0.245	**0.894**	0.041
No2	adv. GW	0.503	0.185	**0.865**	0.049
No2	adversarial	0.612	0.137	**0.943**	0.035
No2	standard	0.521	0.184	**0.894**	0.041
No2	std. GW	0.475	0.206	**0.996**	0.024
No5	pretrained	**0.875**	0.029	0.610	0.124
No5	adversarial	**0.967**	0.013	0.636	0.111
No5	adv. GW	**0.921**	0.020	0.419	0.208
No5	standard	**0.985**	0.012	0.715	0.082
No5	std. GW	**0.919**	0.019	0.487	0.174
No6	pretrained	0.940	0.060	**0.922**	0.036
No6	adversarial	**1.076**	0.037	0.628	0.117
No6	adv. GW	**1.149**	0.017	**1.160**	0.011
No6	standard	**1.064**	0.043	**0.893**	0.037
No6	std. GW	**1.040**	0.046	**0.704**	0.088
WEAT6	pretrained	**0.771**	0.048	**1.053**	0.019
WEAT6	adversarial	**0.808**	0.039	0.722	0.085
WEAT6	adv. GW	**0.808**	0.038	**1.164**	0.011
WEAT6	standard	**0.817**	0.038	**1.074**	0.019
WEAT6	std. GW	**0.770**	0.050	0.725	0.085
WEAT8	pretrained	0.278	0.307	0.706	0.101
WEAT8	adversarial	0.161	0.384	**1.188**	0.013
WEAT8	adv. GW	0.565	0.137	**0.902**	0.047
WEAT8	standard	0.426	0.217	0.782	0.084
WEAT8	std. GW	0.096	0.438	**1.004**	0.034

directly, Table 9 presents masked and unmasked results for WEAT8, which was the only test that both detected significant bias in many cases and is closely related to the STEM/non-STEM classification setting. It is noteworthy that the pretrained and all *masked* fine-tuned models demonstrate some indication of bias, whereas p-values for *unmasked* fine-tuned models generally deviate widely from significance threshold values. This could be a result of the presence of gender-marking words in the dataset, which contains anti-stereotypical correlations that may mitigate stereotypical relations between the model's representations of gender-marking words and science/art words.

5 Discussion

5.1 Downstream performance and fairness trade-offs

Our experiments provide mixed evidence regarding the relationship between fairness interventions and

Table 9: WEAT8 intrinsic bias results for masked and unmasked training. Boldface indicates significant bias ($ES > 0, p < 0.05$ or $ES < 0, p > 0.95$). Italics indicate noteworthy measurements that may indicate bias ($p \approx 0.1$).

Masked	Method	WEAT ES	WEAT p	LPBS ES	LPBS p
-	pretrained	0.278	0.307	*0.706*	0.101
N	adv.	**-1.230**	0.993	0.175	0.403
Y	adv.	0.161	0.384	**1.188**	0.013
N	adv. GW	-0.175	0.636	0.275	0.309
Y	adv. GW	0.565	0.137	**0.902**	0.047
N	std.	-0.148	0.611	0.426	0.224
Y	std.	0.426	0.217	*0.782*	0.084
N	std. GW	0.475	0.187	**1.156**	0.012
Y	std. GW	0.096	0.438	**1.004**	0.034

downstream task performance. Across most settings, fairness gaps were already small in the baseline models, suggesting that the dataset itself does not strongly encode gender correlations with the STEM/non-STEM task. Consequently, masking and adversarial debiasing had only modest effects on group fairness metrics.

At the same time, these interventions often improved downstream accuracy. This superficially appears to contradict the commonly discussed *fairness–performance trade-off*. However, that trade-off typically arises in settings where fairness interventions explicitly alter outcome distributions relative to ground truth labels. In contrast, our methods aim to reduce the model's reliance on sensitive attributes while preserving task-relevant information. In this sense, the improvements we observe suggest that reducing gender information in model representations can be compatible with maintaining or even improving predictive performance.

GenWriter rewrites produced the largest gains in downstream accuracy. This may reflect a standardizing effect of rewriting, reducing stylistic variability across bios and may make task-relevant signals easier for the model to learn. However, improvements in performance did not translate into consistent improvements in group fairness metrics. In several settings, accuracy improvements were unevenly distributed across groups, resulting in small fairness gaps despite overall performance gains. This is likely due in part to the relatively weak correlation between gender and STEM/non-STEM labels in the NTNU dataset, which limits the magnitude of downstream disparities that can arise during training.

5.2 Procedural fairness and gender leakage

One of the clearest patterns in our results concerns gender leakage from hidden representations. Masking names and pronouns substantially reduced recoverable gender information, and adversarial training further reduced leakage. GenWriter rewrites provided an additional reduction, particularly when combined with adversarial debiasing.

These results support the idea that data-level and model-level mitigation methods can be complementary. Masking removes explicit gender markers, rewriting reduces stylistic or structural signals that may correlate with gender, and adversarial training discourages the model from encoding gender information in its representations. Together, these interventions produced near-random gender predictability in several configurations.

From the perspective of the research questions, this provides the strongest evidence for **RQ1** and **RQ3**. The combined mitigation strategy substantially reduces gender leakage while preserving or improving downstream performance, suggesting that procedural fairness objectives can be supported without large performance costs in this setting.

5.3 Intrinsic bias and downstream behavior

The intrinsic bias results present a different picture. Across the WEAT and LPBS test suites, bias signals were mixed and did not consistently decrease under the interventions that most strongly reduced gender leakage. In some cases, fine-tuning or masking even increased the magnitude of certain intrinsic bias signals.

This pattern aligns with prior work showing that intrinsic bias metrics often do not correlate with downstream fairness outcomes (Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022). Our results extend these findings by showing that intrinsic bias measures can also diverge from representation-level measures such as gender leakage. In other words, a model may encode less recoverable gender information in its representations while still exhibiting stereotypical associations under certain intrinsic tests.

One possible explanation is that intrinsic bias tests rely on specific lexical associations that are only weakly connected to the downstream task. However, our results show some evidence that stereotypical encodings can be weakened by unmasked fine-tuning, provided the dataset in question contains anti-stereotypical correlations connected to the corresponding WEAT test concepts. In any case, our results support the claim that intrinsic bias measures should be interpreted as diagnostic signals rather than direct indicators of downstream fairness.

5.4 Implications for bias mitigation in hiring-related NLP

Taken together, our findings suggest several broader implications for bias mitigation in hiring-related NLP systems.

First, evaluating bias mitigation methods in isolation may provide an incomplete picture. Our experiments show that combining interventions at different stages of the pipeline can produce stronger reductions in sensitive-attribute information than any single method alone.

Second, different bias measurements capture distinct aspects of model behavior. Leakage, intrinsic bias tests, fairness gaps, and predictive performance do not move together in a simple way. This reinforces the importance of evaluating bias mitigation across multiple dimensions, especially in high-stakes domains such as hiring.

Finally, the results highlight the value of grounding bias mitigation strategies in the intended downstream application. By focusing on reducing gender leakage while preserving task-relevant information, our approach targets a form of procedural fairness that is directly relevant to recruitment scenarios, where decisions should ideally depend on job-relevant attributes rather than demographic characteristics.

6 Conclusion

This paper investigated how data-level and model-level debiasing methods interact in a hiring-related NLP setting. Using Norwegian-language academic bios and a STEM/non-STEM proxy task, we evaluated masking, GenWriter rewrites, and adversarial debiasing across downstream performance, fairness metrics, intrinsic bias tests, and gender leakage.

Regarding **RQ1**, we find that the interventions interact in complementary ways. Masking and adversarial training substantially reduce gender leakage, and GenWriter rewrites further strengthen this effect. The combination of all three methods reduces recoverable gender information in model representations to near-random levels while maintaining, and in some settings improving, downstream task performance.

For **RQ2**, reductions in gender leakage do not consistently correspond to reductions in intrinsic bias as measured by WEAT or LPBS. Intrinsic bias signals remain mixed across models and tests, reinforcing prior findings that intrinsic bias metrics are not reliable indicators of downstream fairness behavior.

Finally, for **RQ3**, our results suggest that a downstream-oriented mitigation strategy can improve *procedural fairness*—interpreted here as limiting the model's reliance on gender information—without large performance costs. In several cases, debiasing interventions improved task accuracy while reducing gender leakage.

Overall, these findings highlight that different bias measurements capture distinct aspects of model behavior. Effective evaluation of bias mitigation in hiring-related NLP therefore requires a multi-dimensional approach that considers downstream performance, group fairness, representation leakage, and intrinsic bias simultaneously. Our results suggest that combining data-level and model-level interventions can support fairer and more robust systems, particularly in multilingual and context-specific recruitment settings.

Limitations

Several limitations should be considered when interpreting the results.

Proxy task and dataset scope. Our experiments use a proxy task—classification of academic bios into STEM and non-STEM domains—rather than real hiring decisions. Although this task captures a recruitment-related scenario in which textual candidate information is mapped to a consequential category, it does not reflect the full complexity of hiring processes, such as multi-stage evaluation, human judgment, or institutional policies. In addition, the dataset is drawn from a single institution (NTNU), which may limit generalizability to other organizations, sectors, or cultural contexts.

Limited representation of sensitive attributes. This study focuses primarily on gender and uses a binary gender representation. Because the collected dataset contained only a very small number of non-binary examples, non-binary identities could not be meaningfully modeled. This is an important limitation, as real-world hiring systems must account for gender diversity beyond binary categories. We also do not examine intersectional fairness involving attributes such as ethnicity, nationality, religion, or disability. Future work should prioritize datasets that support intersectional analysis and include underrepresented groups.

Model scope. Our experiments focus on BERT-based encoder models rather than modern large language models (LLMs). This choice reflects the classification-oriented nature of the task and enables the use of established adversarial debiasing techniques that operate on hidden representations. Encoder models also remain common in practical pipelines due to efficiency, cost, and data governance considerations. Nevertheless, LLMs are increasingly used in hiring-related applications, and future research should investigate how the mitigation strategies studied here translate to generative and instruction-tuned systems.

Evaluation limitations. Although we evaluate multiple dimensions—including downstream performance, group fairness gaps, intrinsic bias tests, and representation leakage—these metrics capture only partial aspects of fairness. Automated labeling steps used in the dataset construction may also introduce noise. More comprehensive evaluation would require larger annotated datasets and human-centered studies that examine how such systems affect real hiring workflows and stakeholder perceptions of fairness.

Acknowledgments

This work is part of the Europe Horizon project BIAS, grant agreement number 101070468, funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American*

Chapter of the Association for Computational Linguistics, pages 1693–1706. Association for Computational Linguistics.

Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. *arXiv preprint arXiv:2101.10001*.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Towards equal opportunity fairness through adversarial learning. *arXiv preprint arXiv:2203.06317*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. In *Proceedings of 5th SwissText and 16th KONVENS Joint Conference 2020*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5502–5515.

Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. 2023. Fair enough: Searching for sufficient measures of fairness. *ACM Transactions on Software Engineering and Methodology*, 32(6):1–22.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163.

Shweta Soundararajan and Sarah Jane Delany. 2025. Genwriter: Reducing gender cues in biographies through text rewriting. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 347–357.

Ewoenam Kwaku Tokpo, Pieter Delobelle, Bettina Berendt, and Toon Calders. 2023. How far can it go? on intrinsic gender bias mitigation for text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3418–3433.

Schrasing Tong, Minseok Jung, Ilaria Liccardi, and Lalana Kagal. 2026. Measuring perceptions of fairness in ai systems: The effects of infra-marginality. *arXiv preprint arXiv:2603.05889*.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix

A.1 Sample Bios and Rewrites

The following samples were created using ChatGPT-generated fictional bios, which were run through our implementation of the Genwriter pipeline.

Sample 1:

Gender: Female
Profession: Associate Professor, non-STEM

Original Bio: *Dr. Ingrid Halvorsen er førsteamanuensis ved Institutt for kunst- og medievitenskap ved NTNU. Hun forsker på visuell kultur, digital estetikk og kritisk medieteori, med særlig*

fokus på hvordan digitale plattformer og algoritmer påvirker produksjon og tolkning av kunst og medier. Hun underviser i visuell analyse og digital kultur, veileder studenter på master- og ph.d.-nivå og publiserer i internasjonale tidsskrifter innen medievitenskap og visuell kultur.

Translation: *Dr Ingrid Halvorsen is an associate professor at the Department of Art and Media Studies at NTNU. Her research focuses on visual culture, digital aesthetics and critical media theory, with a particular emphasis on how digital platforms and algorithms influence the production and interpretation of art and media. She teaches visual analysis and digital culture, supervises Master's and PhD students, and publishes in international journals in the fields of media studies and visual culture.*

Retrieved Template Bio: *Dr. [Navn på personen] er [Tittel] ved [Institutt] ved [Universitet]. [Navn på personen]s forskning sentrerer seg rundt [Felt] og [Felt], samt anvendelser av disse teknologiene i [Felt], [Felt] og [Felt]. [Navn på personen] underviser i [Fagområde], [Fagområde] og [Fagområde], samt veileder [Studienivå]- og [Studienivå]studenter.*

Translation: *Dr. [Name of person] is [Title] at [Department] at [University]. [Name of person]'s research focuses on [Field] and [Field], as well as applications of these technologies in [Field], [Field] and [Field]. [Name of person] teaches [Subject area], [Subject area] and [Subject area], and supervises [Level of study] and [Level of study] students.*

Rewritten Bio: *Dr. Ingrid Halvorsen er førsteamanuensis ved Institutt for kunst- og medievitenskap ved NTNU. Ingrid Halvorsens forskning sentrerer seg rundt visuell kultur og digital estetikk, samt anvendelser av disse teknologiene i kritisk medieteori, [Felt] og [Felt]. Ingrid Halvorsen underviser i visuell analyse, digital kultur og [Fagområde], samt veileder master- og ph.d.-studenter.*

Translation: *Dr Ingrid Halvorsen is an associate professor at the Department of Art and Media Studies at NTNU. Ingrid Halvorsen's research focuses on visual culture and digital aesthetics, as well as the applications of these technologies in critical media theory, [Field] and [Field]. Ingrid Halvorsen teaches visual analysis, digital culture and [Subject area], and supervises Master's and PhD students.*

Sample 2:

Gender: Male
Profession: PhD Candidate, STEM

Original Bio: *Erik Lundgaard er ph.d.-stipendiat ved Institutt for fysikk ved NTNU. Forskningen hans omhandler kvantematerialer og hvordan elektroniske interaksjoner gir opphav til nye kollektive fenomener i faste stoffer. Han arbeider særlig med beregningsbaserte metoder for å modellere kvantemekaniske systemer. Lundgaard deltar også i undervisning i grunnleggende fysikk og samarbeider med forskere i internasjonale prosjekter.*
Translation: *Erik Lundgaard is a PhD fellow at the Department of Physics at NTNU. His research focuses on quantum materials and how electronic interactions give rise to new collective phenomena in solids. He works in particular with computational methods for modelling quantum mechanical systems. Lundgaard also teaches basic physics and collaborates with researchers on international projects.*

Retrieved Template Bio: *[Navn på personen] er [Yrke] ved [Institutt] ved [Universitet]. Forskningstema [Navn på personen] er opptatt av er [Felt], [Felt], [Felt] og [Felt]. Jobber med [Felt] og modellering av [Felt] for [Felt]. [Navn på personen] har særlig forsket og undervist på temaer knyttet til [Felt], [Felt] og [Felt], og [Han/Hun] har deltatt i ulike forskningsprosjekter med nasjonal og internasjonal finansiering innenfor disse områdene.*
Translation: *[Name of person] is a [Occupation] at [Department] at [University]. The research topics [Name of person] is interested in are [Field], [Field], [Field] and [Field]. Works with [Field] and modelling of [Field] for [Field]. [Name of person] has conducted research and taught on topics related to [Field], [Field] and [Field], and [He/She] has participated in various nationally and internationally funded research projects within these areas.*

Rewritten Bio: *Erik Lundgaard er ph.d.-stipendiat ved Institutt for fysikk ved NTNU. Forskningstema Erik Lundgaard er opptatt av er kvantematerialer, elektroniske interaksjoner, kollektive fenomener i faste stoffer og kvantemekaniske systemer. Jobber med beregningsbaserte metoder og modellering av kvantemekaniske systemer for å studere kollektive fenomener i faste stoffer. Erik Lundgaard har særlig forsket og undervist på temaer knyttet til kvantematerialer, grunnleggende fysikk og kvan-*

temekaniske systemer, og han har deltatt i ulike forskningsprosjekter med nasjonal og internasjonal finansiering innenfor disse områdene.
Translation: *Erik Lundgaard is a PhD fellow at the Department of Physics at NTNU. Erik Lundgaard's research interests include quantum materials, electronic interactions, collective phenomena in solids and quantum mechanical systems. He works with computational methods and modelling of quantum mechanical systems to study collective phenomena in solids. Erik Lundgaard has conducted research and taught on topics related to quantum materials, fundamental physics and quantum mechanical systems, and he has participated in various nationally and internationally funded research projects in these areas.*

A.2 Prompts

Category labels for sentences: "You are classifying Norwegian sentences extracted from professional academic bios into the following categories:

- Education: Degrees, institutions attended, and academic achievements.

- Demographics: Age, gender, nationality, current job title, location and other personal background information.

- Work: Professional experience, job titles, affiliations, research interests and discussion, publications and contributions to the field.

- Non-Professional: Personal hobbies and interests, and other non-professional information.

Here are some examples of sentences and their corresponding categories:

[EXAMPLES]

Please classify the following sentence into one of the categories mentioned above:

[SENTENCE]

Please provide only the category name as your answer."

Synthetic non-professional examples: "Du hjelper til med å utvide et datasett bestående av profesjonelle akademiske biografier fra NTNU-ansatte.

Skriv en setning som er litt annerledes enn følgende eksempel:

[EKSEMPEL]

Oppgi kun setningen som svar."

Translation: "You are helping to expand a dataset consisting of professional academic biographies of NTNU staff.

Write a sentence that is slightly different from the following example:

[EXAMPLE]

Please provide only the sentence as your answer."

Creation of generic sentence templates: "Du er en streng omskriver. Gjør om enhver gitt setning til en generell mal ved å konsekvent erstatte ALLE entiteter og pronomener med norske plassholdere som [Navn på personen], [Yrke], [Spesialisering], [Universitet], [År], [Antall år], [Sykehus], [Sted], [Han/Hun], osv., nøyaktig som vist i eksemplene under. Ikke erstatt førstepersonspronomen som «jeg» eller førstepersons eieform som «min».

Setningene kommer fra profesjonelle biografier, og målet er å erstatte detaljer som er spesifikke for den enkelte med plassholdermerker. Enheter som ikke er relatert til dette målet, trenger ikke å erstattes.

Regler:

- Returner KUN den omskrevne setningen. Ingen forklaringer, ingen ekstra tekst, ingen "Input" eller "Output".

- Hvis input er tom, kun navn/tittel, eller <5 ord, returner input uendret.

- Bruk konsistente plassholdere gjennom hele teksten.

- Hvis noe ikke matcher, behold ordet hvis det ikke er entitet, ellers utelat.

- Behold alltid setningens grammatikk og struktur.

Eksempler:

Dr. Ole Nilsen er en ortopedisk kirurg spesialisert innen artroskopisk kirurgi for kneet.

→ Dr. [Navn på personen] er en [Yrke] spesialisert innen [Spesialisering].

Dr. Hansen fullførte medisinstudiet ved Universitetet i Oslo i 1995 og har vært i praksis i 28 år.

→ Dr. [Navn på personen] fullførte [Studium] ved [Universitet] i [År] og har vært i praksis i [Antall år].

Hun arbeider ved Akershus universitetssykehus med sitt team i Lørenskog, Oslo universitetssykehus i Oslo og har erfaring fra St. Olavs hospital i Trondheim.

→ [Han/Hun] arbeider ved [Sykehus] med sitt team i [Sted], [Sykehus] i [Sted] og har erfaring fra [Sykehus] i [Sted].

Jeg har erfaring med prosjekter innen digital læring.

→ Jeg har erfaring med prosjekter innen [Felt]."

Translation: "You are a strict rewriter. Convert any given sentence into a general template by consistently replacing ALL entities and pronouns with Norwegian placeholders such as [Name of person], [Occupation], [Specialisation], [University], [Year], [Number of years], [Hospital], [Location], [He/She], etc., exactly as shown in the examples below. Do not replace first-person pronouns such as 'I' or first-person possessives such as 'my'.

The sentences are taken from professional biographies, and the aim is to replace details specific to the individual with placeholder tags. Entities not related to this aim do not need to be replaced.

Rules:

- Return ONLY the rewritten sentence. No explanations, no extra text, no 'Input' or 'Output'.
- If the input is empty, contains only a name/title, or has fewer than 5 words, return the input unchanged.
- Use consistent placeholders throughout the text.
- If something does not match, retain the word if it is not an entity, otherwise omit it.
- Always retain the sentence's grammar and structure.

Examples:

Dr Ole Nilsen is an orthopaedic surgeon specialising in arthroscopic knee surgery.

→ Dr [Name of person] is a [Profession] specialising in [Specialisation].

Dr Hansen completed his medical degree at the University of Oslo in 1995 and has been practising for 28 years.

→ Dr [Name of person] completed [Degree] at [University] in [Year] and has been practising for [Number of years].

She works at Akershus University Hospital with her team in Lørenskog, Oslo University Hospital in Oslo, and has experience from St. Olavs Hospital in Trondheim.

→ [He/She] works at [Hospital] with his/her team in [Place], [Hospital] in [Place] and has experience from [Hospital] in [Place].

I have experience with projects in digital learning.

→ I have experience with projects in [Field]."

Rewriting bios: "Gitt følgende biografi og mal, utfør følgende trinn:

1. Forstå biografien og malen: Les og analyser biografien og malen nøye for å forstå konteksten, plassholdere og tilgjengelig informasjon.
2. Erstatt plassholdere: Erstatt hver plassholder i malen med passende verdier utledet fra biografien. Bruk følgende regler når du erstatter plassholdere: - Hold formatet og strukturen til malen uendret. - Hvis en plassholder ikke kan erstattes på grunn av utilstrekkelig informasjon i biografien, behold plassholderen som den er.
3. Utdata: Gi kun den endelige utfylte malen med plassholdere erstattet der det er mulig.
4. Rett opp avvik."

Translation: "Given the following biography and template, carry out the following steps:

1. Understand the biography and template: Read and analyse the biography and template carefully to understand the context, placeholders and available information.
2. Replace placeholders: Replace each placeholder in the template with appropriate values derived from the biography. Use the following rules when replacing placeholders: - Keep the format and structure of the template unchanged. - If a placeholder cannot be replaced due to insufficient information in the biography, leave the placeholder as it is.
3. Output: Provide only the final completed template with placeholders replaced where possible.
4. Correct discrepancies."

Matching Meaning at Scale: Evaluating Semantic Search for 18th-Century Intellectual History through the Case of Locke

Yu Wu[1], Ananth Mahadevan[1]*, Filip Ginter[2,3], Michael Mathioudakis[1], Mikko Tolonen[1]
[1]University of Helsinki [2]TurkuNLP, University of Turku [3]ELLIS Institute Finland
{yu.wu, ananth.mahadevan, michael.mathioudakis, mikko.tolonen}@helsinki.fi
figint@utu.fi

Abstract

While digitized corpora have transformed the study of intellectual transmission, current methods rely heavily on lexical text reuse detection, capturing verbatim quotations but fundamentally missing paraphrases and complex implicit engagement. This paper evaluates semantic search in 18th-century intellectual history through the reception of John Locke's foundational work. Using expert annotation grounded in a semantic taxonomy, we examine whether an off-the-shelf semantic search pipeline can surface meaning-level correspondences overlooked by lexical methods. Our results demonstrate that semantic search retrieves substantially more implicit receptions than lexical baselines. However, linguistic diagnostics also reveal a "lexical gatekeeping" effect, where retrieval remains partially constrained by surface vocabulary overlap. These findings highlight both the potential and the limitations of semantic retrieval for analyzing the circulation of ideas in large historical corpora. The data is available at https://github.com/COMHIS/locke-sim-data.

1 Introduction

The digitization of vast historical archives, such as Eighteenth Century Collections Online (ECCO) (Tolonen et al., 2022) and Early English Books Online (EEBO) (Mäkelä et al., 2025), together with recent advances in Natural Language Processing (NLP), has created unprecedented opportunities for intellectual history (Tolonen and Ryan, 2026). Digital methodologies now allow scholars to move beyond the study of isolated authorial intent and begin to "recreate the cultural terrain" of historical debate by tracing how ideas circulated across large textual corpora (Edelstein, 2016). At the same time, this shift from close reading to large-scale computational analysis exposes a central methodological bottleneck: the semantic gap between lexical similarity and meaning-level relationships preserved in textual passages.

Traditional computational approaches to reception history rely heavily on lexical text reuse detection (Rosson et al., 2023); for further discussion in the context of intellectual history, see Tolonen and Ryan (2026). While effective for identifying verbatim quotations, such approaches struggle with implicit reference, paraphrase, Optical Character Recognition (OCR) errors, diachronic lexical change, and the conceptual complexity of philosophical language. As a result, many forms of intellectual engagement between texts remain difficult to detect at scale.

One context in which this semantic gap becomes particularly visible is 18th-century British intellectual history. The period is widely recognised as a locus of Enlightenment contestation and a rapidly expanding print public, where arguments about reason, religion, science, and politics circulated across genres and decades (Porter, 2001). Tracing how such arguments were reformulated and redeployed is essential for understanding the transmission of ideas, yet doing so across millions of publications is beyond the scope of manual scholarship. While recent advances in semantic search based on dense vector retrieval (Reimers and Gurevych, 2019; Douze et al., 2025) make it theoretically plausible to detect semantic similarity beyond exact lexical overlap in the time frame and language domain, their reliability on uncurated historical corpora remains empirically untested.

To operationalise this inquiry, we focus on John Locke (1632–1704)'s *An Essay Concerning Human Understanding* (1689). Because the *Essay* was published at the end of the seventeenth century, its reception spans the entire 18th-century corpus analysed here, allowing its influence to be traced across the full period. This is not optimal for authors such as David Hume, whose major Enlightenment pub-

*Contribution made during doctoral studies.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 214–225
July 4, 2026 ©2026 Association for Computational Linguistics

lications appear only from the late 1730s onwards. Locke's work is also particularly suitable because his intellectual legacy is more contested and complex than the traditional narrative of the "father of liberalism" suggests (Stanton, 2018). Recent scholarship has highlighted the fragmented nature of his reception, which developed into distinct republican and libertarian traditions (Layman, 2021). At the level of textual practice, Locke's arguments often appear in structured formulations that are subsequently reformulated or paraphrased in later texts. In such cases, lexical searches fundamentally fail to capture the semantic footprint of the idea.

This paper therefore presents a proof-of-concept evaluation of semantic search for 18th-century historical corpora through the case of Locke's quotations. We examine whether a minimal, off-the-shelf semantic search setup can identify meaning-level correspondences in the reception of Locke that lexical text reuse detection methods miss. We first operationalize "reception" into a hierarchical framework that distinguishes genuine semantic engagement, such as paraphrases and implicit meaning matches, from both verbatim text reuse and mere topical matches. Guided by this framework, we implement a multi-stage evaluation workflow, annotating and analyzing iteratively in non-lexical hits of our semantic search pipeline. Based on the annotated subset, our mixed-methods analysis demonstrates that semantic search uncovers an order of magnitude more implicit reception than traditional string matching across core historical queries. Finally, through a linguistic diagnostic encompassing vocabulary overlap, OCR noise, and syntactic divergence, we expose a profound "lexical gatekeeping" effect: latent lexical similarities continue to shape retrieval outcomes, even within dense semantic spaces.

To summarize, our main contributions are:

- A proof-of-concept evaluation using Locke as a case study, demonstrating that semantic search retrieves substantially more implicit receptions than lexical text-reuse baselines.
- A conceptual framework for modelling reception that distinguishes between different contextual matches at the level of semantic relationships between passages.
- A reusable iterative evaluation workflow that applies the annotation framework to identify meaning-level receptions and their historical significance.
- A linguistic diagnostic analysis revealing a *lexical gatekeeping* effect despite robustness to OCR noise and syntactic variation.

2 Related Work

Digitized Infrastructure and Lexical Methods The mass digitization of historical archives (Armitage and Guldi, 2014; Bode, 2018; Langlais, 2021), coupled with expansive structured metadata and relational networks (Hill et al., 2019; Bermès, 2017), has established the infrastructure for computational intellectual history. Retrieving specific conceptual evidence from these corpora has traditionally relied on lexical methodologies. Techniques ranging from Keyword in Context (KWIC) (Luhn, 1960), word-level distributional semantics (Fitzmaurice et al., 2017), and probabilistic topic modeling (Blei, 2012; Goldstone and Underwood, 2014), to advanced hashing and sequence-alignment algorithms for OCR-resilient text reuse (Mahadevan et al., 2025; Düring et al., 2023; Vesanto et al., 2017; Smith et al., 2014), mark the current limits of surface-level matching. Bridging this persistent "semantic gap" requires moving beyond static methods to neural, passage-level models capable of tracking complex conceptual shifts (McGillivray et al., 2024; Giulianelli et al., 2020).

Semantic Search in Historical Corpora To advance beyond lexical navigation, digital humanists increasingly use dense Information Retrieval (IR) models, such as Dense Passage Retrieval (Karpukhin et al., 2020) and Sentence Transformer (Reimers and Gurevych, 2019), to capture compositional semantics, bypassing explicit string matching. Recent applications demonstrate the capacity of bi-encoders to capture complex intertextualities (Schelb et al., 2026) and semantic parallels (Kanerva et al., 2025; Franklin et al., 2024; Alrahabi and Wainstain, 2025) that remain invisible to lexical searches. However, deploying contemporary neural networks on uncurated historical paragraphs exposes severe algorithmic vulnerabilities. Alongside OCR-induced tokenization artifacts (Zhang et al., 2025; Michail et al., 2025), temporal semantic degradation (Lazaridou et al., 2021) and entity-centric popularity biases (Sciavolino et al., 2021), dense models regress to shallow vocabulary matching when confronted with out-of-domain historical syntax (Rizwan et al., 2025; MacAvaney et al., 2022). This limitation is further entrenched by standard IR benchmarks, which

structurally penalize true semantic retrieval through lexically pre-filtered ground truth (Thakur et al., 2021). Therefore, rigorously evaluating semantic architectures within real-world historical corpora is critical to distinguish genuine meaning retrieval from lingering algorithmic biases.

Computational Early Modern Reception Early computational methods successfully mapped early modern intellectual networks using structured metadata from social and epistolary graphs (Warren et al., 2016; Edelstein et al., 2017; Burrows and Curran, 2012), while lexical pipelines tracked verbatim text reuse (Roe, 2024; Tolonen and Spencer, 2025) and extracted explicit quotations (Muzny et al., 2017; Brunner et al., 2020). However, the highly contested intellectual legacy of foundational figures like John Locke (Collins, 2020; Hanck, 2019) exemplifies the requirement of moving to passage-level semantic mappings: his foundational concepts were widely appropriated without explicit citation, relying on implicit references and shifting vocabularies that completely evade traditional lexical searches (Harris, 2023; Carey, 2006). Using Locke's legacy as an empirical stress test, our study evaluates whether contemporary semantic retrieval can bypass surface-level lexical biases to uncover hidden conceptual reuse within the noisy 18th-century archive.

3 Evaluation Methodology

Our evaluation framework is designed to isolate significant meaning matches from mere lexical reuse. By conceptualizing the retrieval hits as potential reception signals, we aim to uncover their underlying matching distribution and the analytical conditions under which a semantic match constitutes robust historical evidence.

3.1 Heuristic Annotation Taxonomy

To ground the iterative workflow in a stable interpretive framework, we introduce an annotation taxonomy for modelling reception that distinguishes between different types of textual correspondence. The framework is designed as an operational heuristic to facilitate rapid triage rather than imposing a rigid typology. It is a hierarchy of heuristic matches consisting of *Lexical Matches*, *Paraphrase Matches* (*Paraphrases* for short), and *Meaning Matches*, alongside a separate category of *Topical Matches*. *Lexical Matches* capture verbatim reuse, while *Paraphrase Matches* identify cases in which wording has changed but a clear quotation relationship remains. *Meaning Matches*, by contrast, refer to passages that display semantic engagement suggesting that one text meaningfully interacts with another even without explicit lexical overlap. *Topical Matches*, in turn, refer to passages that address the same subject without semantic evidence of such engagement. For cases of OCR failure or total semantic divergence, we supplement this hierarchy with a *No Match* category, and *Don't Know* for the cases where the semantic connection is too ambiguous for rapid judgment, requiring extensive close reading to verify. In our evaluation, *Paraphrases* and *Meaning Matches* are aggregated as *significant hits*, representing genuine intellectual reception.

3.2 Iterative Evaluation Strategy

To extract high-signal reception from a noisy retrieval pool, we formalize a multi-stage, human-in-the-loop evaluation framework, illustrated as Fig. 1. We begin by sampling highly reused quotes from Locke's texts to serve as our search queries. We then deploy a pilot sampling strategy across the broad retrieval pool to understand the basic semantic distribution without presupposing algorithmic success. Finally, we conditionally target our deepening annotation based on these initial results. This progressive annotation process optimizes expert bandwidth, prioritizing genuine reception clusters over algorithmic noise.

3.3 Analytical Framework

To evaluate the pipeline's utility and its underlying constraints, our framework covers content-level statistical analysis and linguistic-level diagnosis. Firstly, we evaluate the retrieval performance by analyzing how the annotated semantic labels and metadata are distributed across the ranking. Secondly, to interpret the linguistic patterns identified in these results, we first map the annotated English hits into a standard confusion matrix (TP, TN, FP, and FN subsets) based on their ranking intervals and semantic validity. Next, within each quadrant of the matrix, we compute three targeted linguistic features to decode what triggers these specific errors or successes: **Vocab Sim.** (Jaccard similarity of word lemmas) quantifies the model's dependency on surface-level lexical anchoring, while **OOV** (Out-of-Vocabulary rate on query and target separately) serves as a proxy for historical OCR noise. Additionally, **POS Div.** utilizes Jensen-Shannon divergence over Part-of-Speech distribu-

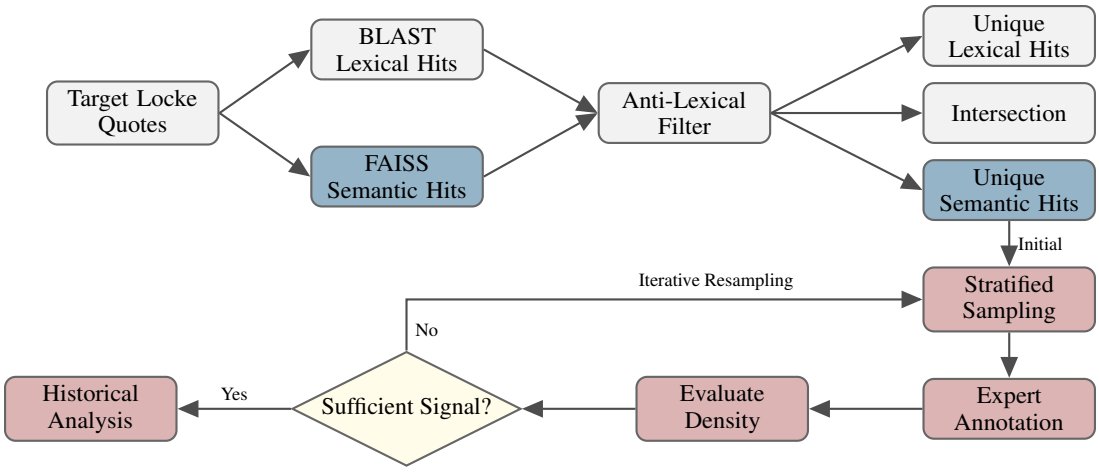


Figure 1: The overall evaluation workflow. The automated search pipeline (top) extracts and filters semantic candidates (blue blocks) against the lexical baseline. The human-in-the-loop workflow (bottom, red blocks) systematically annotates the resulting *Unique Semantic Hits* to evaluate signal density and historical significance.

tions to measure linguistic structural rephrasing independent of vocabulary.

4 Empirical Implementation

This section details the experimental setup and evaluation workflow operated based on our evaluation methodology described in Sec. 3.

4.1 Corpus

Our corpus is the official OCR-extracted text provided by the Eighteenth Century Collections Online (ECCO) (Tolonen et al., 2022), which contains over 32 million book pages from the 18th century. This dataset serves as an ideal testbed for tracing John Locke's reception, as his influence permeates the period's diverse genres, from political tracts to educational manuals. The comprehensive corpus undergoes chunking (as detailed in Sec. 4.4) to build the core index for the vectorized search. To manage duplicate editions and ensure bibliographic sufficiency, we integrated metadata from the English Short Title Catalogue (ESTC), specifically the work IDs, authors, publication years of different versions, and genres. The genre labels are assigned based on the Genre Classification Workflow developed by Tiihonen and Hinderks (2025).

4.2 Text Reuse Baseline

As our primary baseline for capturing explicit intellectual reception, we utilize the comprehensive text reuse network generated across the ECCO corpus (Rosson et al., 2023). This infrastructure relies on an adaptation of BLAST (Basic Local Alignment Search Tool). Its fuzzy-matching capabilities inherently tolerate 18th-century orthographic variations and persistent OCR noise, effectively mapping any text input to verbatim quotations, exact reprints, and lightly edited excerpts across millions of pages. Despite the limitation on semantic-level matching, we repurpose the lexical reuse for three foundational tasks: establishing a comparative baseline, sourcing evaluation queries (Sec. 4.3), and purifying our unique semantic hits (Sec. 4.4).

4.3 Query Source and Construction

In alignment with grounding computational evaluation in 18th-century British intellectual history, we selected John Locke's *An Abridgment of Mr Locke's Essay Concerning Human Understanding* (ESTC R22993) as a highly representative, scalable case study. Published prior to the 18th century, this foundational work allows us to track intellectual reception across the entirety of the ECCO timeline. Because its historical significance is universally recognized and it possesses a dense, well-documented network of existing lexical reuses, it provides an ideal, undisputed baseline without the need to overstate the novelty of the source itself.

To capture genuine external lexical reuse rather than isolated fragments, we queried the baseline for reuse text clusters originating from this work, as introduced by Rosson et al. (2023). In this pipeline, we formally define a *quote* not as a punctuated citation, but as a contiguous, verifiable fragment of reused text. We isolated these quotes by enforcing a length constraint of 150–300 characters, and a minimum external reuse frequency ($n \geq 3$)[1].

[1]These thresholds were empirically determined to eliminate uninterpretable short snippets and overly long, multi-page sequences that obscure targeted semantic analysis.

4.4 Minimal Semantic Search Pipeline

We adopt a deliberately minimal dense retrieval-based search pipeline to evaluate the conceptual semantic capabilities of semantic search. The implementation of core infrastructure is publicly accessible via the `semantic-sim` repository[2]. The underlying vector index was built upon a broader Early Modern English corpus (encompassing both 18th-century ECCO and pre-18th-century EEBO texts). We can easily isolate the target ECCO corpus by filtering the retrieved outputs via document IDs. The texts were segmented into 100-token chunks and encoded using the `paraphrase-multilingual-mpnet-base-v2` model (Reimers and Gurevych, 2020; Song et al., 2020). This encoder natively supports over 50 languages, including English, French, Italian, and Spanish, while allowing us to evaluate its zero-shot capability on unsupported historical languages such as Latin. Efficient vector indexing and retrieval were executed via FAISS (Facebook AI Similarity Search) (Douze et al., 2025). For each query, this index outputs a ranked list of retrieved segments with their semantic similarity scores to the input. We referenced these ranked outputs against our BLAST lexical baseline. By computing the set difference between the two systems, we filtered out known verbatim overlap.

4.5 Annotation Setup

We deployed a project in the locally hosted instance of Label Studio (Tkachenko et al., 2020-2025) for the manual annotation. As illustrated in Fig. 2, the interface was designed to prevent annotators from evaluating fragments in a vacuum. Alongside the retrieved chunk and Locke's original quote, the UI simultaneously displayed the corresponding primary author, work title, and publication year, and provided direct access to the broader document context. The annotation was conducted by five domain experts from the Computational History Group (COMHIS) at University of Helsinki. Each candidate was assigned to a single annotator for rapid triage. Ambiguous cases were marked as *Don't Know* and processed independently.

4.6 Annotation Process

We implemented a workflow following our methodology in Sec. 3 and all the setup above to deploy our iterative evaluation strategy. The empirical outcomes and the evidence for iterative deepening are detailed in Sec. 5.

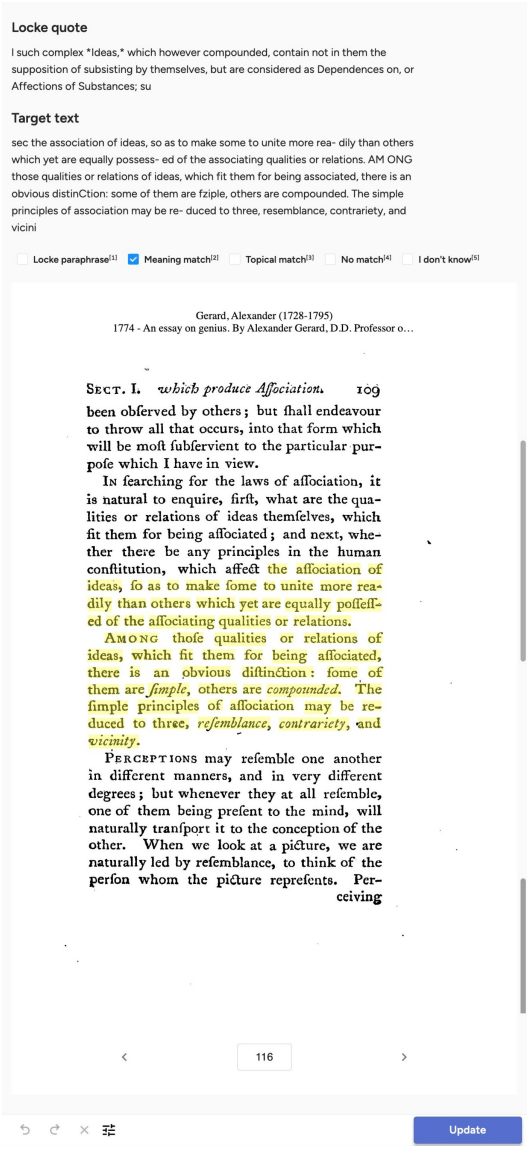


Figure 2: The annotation interface screenshot.

Source Selection From the top 1,000 most frequently reused segments extracted in Sec. 4.3, we selected the top-5 highest-frequency quotes, supplemented by 5 randomly sampled from each of three subsequent frequency tiers (ranks 6-50, 51-150, and 151-1,000), yielding 20 search queries.

Search and Filtering For these 20 queries, we retrieved the top 200,000 semantic hits using our pipeline established in Sec. 4.4. Filtering by document ID retained an average of 173,266 ECCO hits per query (86.6%). To prevent the search space from being dominated by intra-document repetition or frequent reprints, we deduplicated the results by ESTC work ID. Retaining only the highest-scoring hit per distinct work yielded an average of 13,391

[2] https://github.com/TurkuNLP/semantic-sim

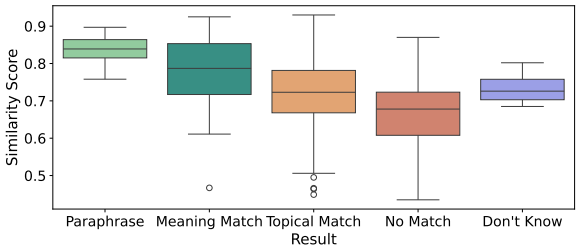


Figure 3: Similarity scores across categories, indicating a lack of global calibration.

unique candidates per query. We then filtered out known verbatim overlap.

Stratified Pilot Sampling We conducted a probing manual annotation of 50 hits per quote, sampling the top 5 results and uniformly distributed intervals across the top 90% of the filtered hits. They were distributed among five domain experts and annotated according to the taxonomy in Sec. 3.1.

Iterative Deepening We annotated more top hits for specific quotes according to a dual criterion: a high density of semantic matches and substantive historical significance. We set the top 200 as an empirical annotation upper bound. Before the exhaustive annotation, we annotated 50 samples within the top-200 hits for a quick triage to verify that the promising yields from our initial broad but sparse sampling were not merely algorithmic outliers. They were the top 20 hits, plus 30 hits sampled iteratively (one every six) from the remainder. By confirming a consistently dense semantic signal first, we mitigated the risk of exhaustively annotating a noise-heavy set that would force us to restart the deepening process. Ultimately, candidates maintaining robust yields triggered the exhaustive annotation of their complete top-200 results for deep-dive historical analysis.

5 Results and Analysis

Our analysis assesses the reliability and technical limitations of the semantic search pipeline.

5.1 Overall Results

Before detailing the annotation outcomes, we quantify the pipeline's lexical filtering efficacy. Semantic search achieved a 95.5% recall of BLAST exact matches (averaging 103.9 intersection vs. 4.9 unique BLAST hits), confirming it effectively matched OCR-resilient lexical baselines even before capturing semantic variance.

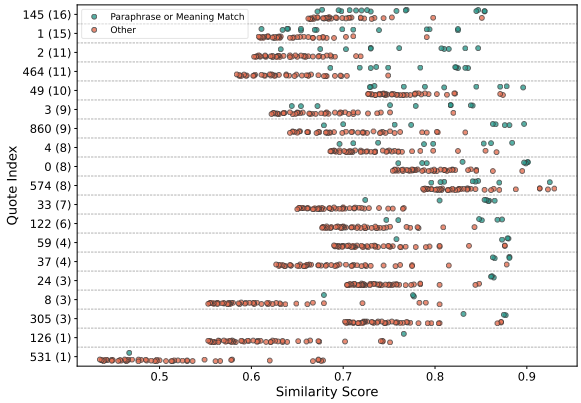


Figure 4: Hit distribution per quote, highlighting query-specific heterogeneity. Queries are sorted by the absolute count of significant hits, indicated in parentheses.

On this highly refined dataset, our heuristic taxonomy facilitated an efficient rapid-triage evaluation, logging a median annotation time of approximately 19.3 seconds per candidate. The resulting annotations demonstrate that the model possesses a robust ordinal capability. As detailed in Tab. 1, the ranking validity, measured by Spearman's rank correlation (ρ), exhibits a highly significant monotonic relationship with human judgments. The most informative semantic categories (*Paraphrase* and *Meaning Match*) show strong negative correlations with the hit ranks, confirming that the system surfaces substantive semantically matched fragments to the top of the ranking. This efficacy is most pronounced at the very top: the top-5 ranked hits are predominantly valid semantic matches, with zero of *No Match* across 15 of the 20 quotes.

While simply extracting the top-K results suffices for standard ad-hoc retrieval, corpus-level intellectual history requires the complete boundaries of a complete semantic match cluster to quantify macro-level reception patterns. Our evaluation reveals that automating this boundary-drawing is fundamentally bottlenecked by score calibration. As shown in Fig. 3, while relative score ranking is correlated to semantic categories, absolute similarity scores overlap significantly across them, preventing the use of a universal global threshold.

This uncalibrated nature extends deeply to the query level. Fig. 4 demonstrates pronounced query-specific heterogeneity: the score thresholds for significant hits shift drastically depending on the specific query content. More critically, even for individual quotes, these local boundaries remain highly fuzzy, preventing the extraction of a cleanly separated match pool for each quote. While this noise partially reflects the fundamental ambiguity

Category	Overall (%)	In Top-5 Ranked(%)	In Top 20% (%)	In Top 50% (%)	Spearman's ρ
Paraphrase	2.1	18.0	7.0	3.5	-0.224***
Meaning Match	11.6	44.0	26.3	16.3	-0.294***
Topical Match	31.1	24.0	34.0	32.7	-0.054
No Match	54.8	14.0	32.3	47.0	0.307***
Don't Know	0.4	0	0.3	0.5	-0.026

Table 1: Distribution and ranking validity of semantic categories. Overall, Top-5 Ranked, Top 20%, and Top 50% yields are reported as mean percentages across the sampled hits, calculated based on original retrieval rankings. For each category, Spearman's rank correlation (ρ) is calculated between its binary indicator and the local retrieval rank within the sampled subset of each query. Significance levels: *** $p < 0.001$.

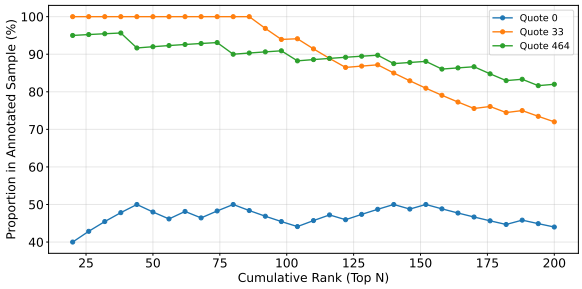


Figure 5: Proportion of significant hits within the 50 validation samples across the top 200 hits for each quote.

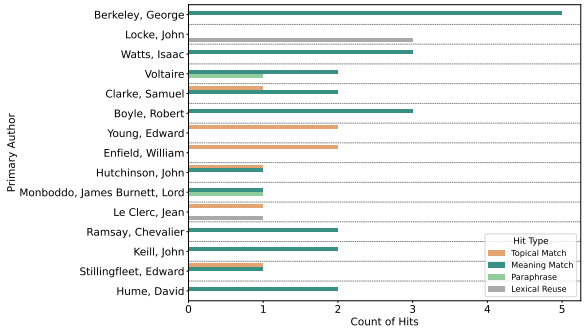


Figure 6: Top 15 authors of the top hits for Quote 464.

of reception research, where the line between a *Meaning Match* and a mere *Topical Match* is often epistemologically contested, it makes individual thresholds hard to set. This failure of both global and local calibration necessitates a methodological pivot. To determine whether semantic search can still provide inspiring insights for intellectual history, we shift from thresholding strategies to targeted case studies. We prioritize specific quotes with clearer empirical score boundaries to assess whether their high-ranking hits yield substantive meaning matches with historical significance.

5.2 Individual Results

To operationalize our iterative deepening (Section 3), we finalized three core queries from an initial mixed pool by balancing algorithmic clarity with historical significance. Quotes 33 and 464 were selected as primary anchors because of their clear semantic boundaries at high similarity scores (> 0.8) alongside profound historical relevance. Conversely, Quote 0, supplemented by domain experts due to its massive lexical footprint, was retained as a comparative baseline. Contrasting with the clear signals of the primary anchors, Quote 0 serves as a rigorous sanity check to verify whether its top-ranked hits reproduce the boundary fuzziness observed during the broad initial sampling.

Before committing to the labor-intensive exhaustive annotation of their top hits, we utilized our 50-sample validation set in top-200 hits as a rapid validation mechanism. Fig. 5 tracks the proportion of significant hits across this sampling continuum. While the proportion for Quote 0 keeps low in top hits, confirming its boundary fuzziness, Quotes 33 and 464 maintain robust yields above 70% and 80%, respectively. This sustained density verifies that these top hits represent deeply cohesive semantic clusters rather than anomalous spikes.

Following the curation of the expert-validated dataset with 400 hits from the two core quotes, we conducted a structural analysis of their historical distribution. A fundamental finding is the sheer volume disparity: significant semantic reuse (*Paraphrase* and *Meaning Match*) outnumbers verbatim lexical reuse by over an order of magnitude (313 vs. 22 hits at the work level), effectively broadening the observable horizon of reception studies. This expansion is starkly visible at the authorial level. Taking Quote 464 as a representative example (Fig. 6), traditional lexical matching predominantly isolates citations by Locke; in contrast, semantic search uncovers a dense, previously invisible network of engagement from central intellectual figures, including Berkeley, Boyle, Hume, and Voltaire.

Projecting these results onto temporal and disciplinary axes (Fig. 7) reveals that the distribution of reuse types varies heavily by genre. For instance, semantic reuse is significantly more prevalent in *Philosophy* and *Religion* than verbatim quotation.

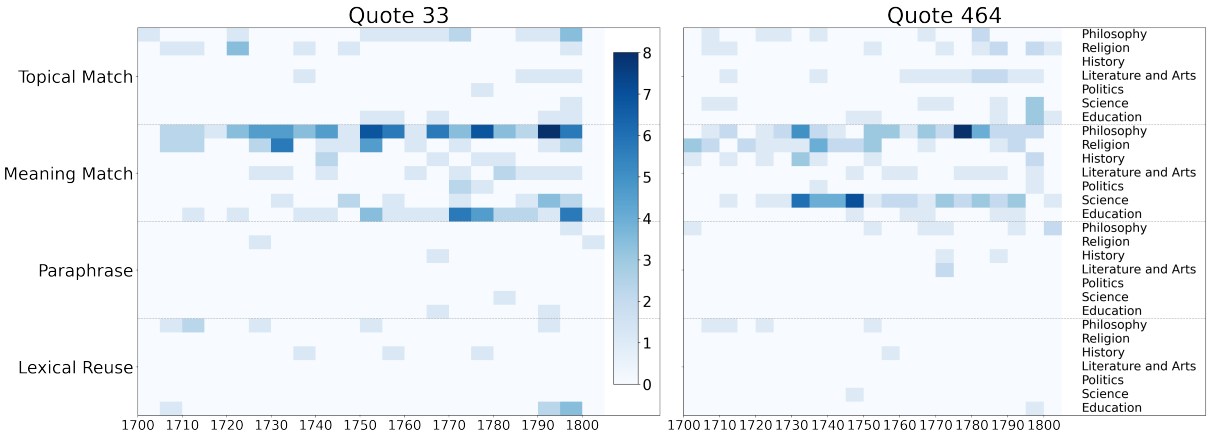


Figure 7: Temporal and disciplinary count distribution of retrieved hits for Quotes 33 (left) and 464 (right). These heatmaps visualize the hits in semantic categories across the seven most dominant 18th-century book genres.

ID	Original Quote Fragment for Query
33	l such complex *Ideas,* which however compounded, contain not in them the supposition of subsisting by themselves, but are considered as Dependences on, or Affections of Substances; su
464	ing their Causes, or any otherwise. The *Atomists,* who define Motion to be *a passage from one place to another,* What do they more than put one synonymous Word for another? For what is *Passage* other than *Motion

Table 2: Representative quotes for the qualitative study.

Temporal analysis further reveals distinct trajectories, with different hit types peaking at different decades, suggesting evolving modes of engagement over time. Furthermore, the two analyzed quotes exhibit highly divergent reception profiles: one forms massive, tightly-knit clusters within philosophical and educational discourse, while the other displays a pronounced volume of reuses across the sciences. This macro-level analysis demonstrates the search system's utility in capturing structural intellectual shifts invisible to traditional lexical detection. How these divergent profiles manifest in actual rhetorical repurposing will be the focus of our qualitative case studies.

5.3 Case Study

To interpret these macro-patterns, we examined the intersection of genre and publication year through targeted close reading. Tab. 2 presents the original texts for these cases. Focusing on Quote 464, from the right sub-figure of Fig. 7, we observe a distinct peak in reuses within the science category during the 1740s. Qualitative analysis reveals that these hits largely originate from scientific textbooks and manuals. This phenomenon appears closely linked to the consolidation of Newtonianism and the concurrent rebuttal of Cartesian physics, where Lockean epistemological framework concerning motion was repurposed to ground scientific empiricism.

In contrast to this polemical weaponization of Quote 464, Quote 33 illustrates an entirely different trajectory of conceptual drift: pedagogical formalization. As indicated by its dense clustering within the philosophy and education genres, Locke's complex epistemological framework was systematically adapted into instructional manuals. Rather than engaging with the philosophical nuances, these texts typically simplified and repackaged his core logic to serve didactic purposes.

Crucially, across both divergent trajectories, the 18th-century authors rarely quoted the original sources verbatim. Whether paraphrasing physics for scientific pedagogy or conceptually flattening Locke for educational manuals, this "implicit reception" is successfully captured by semantic search. By retrieving these highly reformulated iterations, the semantic search demonstrates its applicability in tracking ideas as they cross disciplinary boundaries and registers of complexity, uncovering context-dependent intellectual shifts that traditional lexical searches would inevitably overlook.

5.4 Linguistic Diagnostics

While individual analyses and case studies validate the pipeline's potential, our linguistics-driven analysis reveals the constraints and properties of this success, based on the method proposed in Sec 3.3.

Starting at the macro scale, language identification via Lingua (Stahl, 2021) reveals that the semantic search demonstrates a pronounced capacity for cross-lingual alignment, as detailed in Tab. 3. When compared against the token-weighted corpus

Category (%)	English	French	Latin	Italian	Spanish
ECCO Tokens	92.2	3.3	3.4	0.4	0.1
Paraphrase	100.0	0.0	0.0	0.0	0.0
Meaning Match	90.5	9.5	0.0	0.0	0.0
Topical Match	91.0	7.4	0.0	1.3	0.3
No Match	92.7	6.2	0.9	0.2	0.0
Don't Know	100.0	0.0	0.0	0.0	0.0

Table 3: Language distribution across semantic categories in the 1000 pilot samples. The ECCO token language distribution baseline is calculated via token-level aggregation from the ECCO metadata.

Subset	*N*	Vocab Sim.	Quote OOV	Hit OOV	POS Div.
Top P	40	0.030 ± 0.026	4.11 ± 3.31	19.25 ± 7.88	0.348 ± 0.048
Top N	121	0.017 ± 0.020	1.73 ± 2.45	20.62 ± 7.62	0.362 ± 0.061
Tail N	197	0.012 ± 0.017	2.17 ± 2.57	21.23 ± 7.23	0.359 ± 0.064
Tail P	10	0.009 ± 0.019	3.44 ± 3.88	17.46 ± 2.62	0.331 ± 0.045

Table 4: Linguistic feature distributions across English hit subsets. The rows correspond to defined TP: highly ranked (in top 30%) significant hits (*Paraphrase* or *Meaning Match*), FP: highly ranked (in top 30%) non-matches, TN: no-matches in the tail (top 60%–90%), and FN accordingly. ***N*** represents the size of the subset.

baseline, the highly relevant "Meaning Match" category exhibits a nearly three-fold enrichment of French texts (9.5%). This non-random distribution across semantic tiers confirms that the multilingual embedding space effectively maps English philosophical propositions onto Continental discourse, even under noisy OCR'ed text, bypassing the need for explicit literal translation. Conversely, the model fails to achieve zero-shot semantic transfer for unsupported historical languages like Latin (3.4% of the corpus), which appears exclusively as noise within the *No Match* category (0.9%).

However, this cross-lingual reach introduces a specific structural side effect. A close inspection of the retrieval outcomes reveals that the top-ranked FP contains a noticeably higher proportion of non-English hits (14.0%) compared to the TP (10.0%). This inflation indicates that the model occasionally exhibits a cross-lingual over-eagerness, elevating non-English texts based on coarse-grained topical proximity rather than strict conceptual equivalence, thereby introducing systemic noise at the head of the search results.

To isolate the underlying model mechanics from this cross-lingual noise, evaluating the purely English subsets exposes rigid monolingual bias, processed using spaCy's large English model (Honnibal et al., 2020). As illustrated by the linguistic diagnostics in Tab. 4, TPs inherently rely on a higher degree of lexical overlap with the original query, even though the proportion is very small, with a maximum of 0.030. This is further evidenced by other subsets, which crowd the results due to their higher lexical similarity compared to the marginalized matches. In contrast, FNs display near-zero lexical overlap, revealing the existing limitation on retrieving highly non-lexical semantic matches. Because the FNs actually exhibit the lowest Hit OOV and the most similar syntactic structure despite consistently low query-side noise, this disparity confirms a critical paradox: while the model is remarkably robust against OCR degradation and syntactic drift, its ranking remains anchored to surface-level lexical similarity, systematically marginalizing structurally aligned historical concepts, regardless of their superior textual integrity.

6 Conclusion

This study demonstrates that a human-in-the-loop semantic search workflow can surpass the limitations of traditional lexical matching, uncovering a significant volume of implicit reception within noisy historical corpora. By offering both a methodological proof-of-concept and an expert-curated reference dataset, our purpose is to provide feasibility evidence and a reusable evaluation framework. We plan to scale it into a systematic benchmark to evaluate dense retrieval models across broader historical and low-resource domains.

Acknowledgments

This research was funded by the European Union under the Horizon Europe Marie Skłodowska-Curie Actions (Grant Agreement No. 101119511) and the Academy of Finland (Grant Nos. 347706 and 347709). Computational resources were provided by CSC – IT Center for Science. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or Horizon Europe MSCA Actions. Neither the European Union nor the granting authority can be held responsible for them.

We extend our gratitude to Kira Hinderks, Iiro Tiihonen, Yann Ryan, and David Rosson for their extensive support and feedback during the data annotation phase, alongside the broader Computational History Group (COMHIS) network for their insightful discussions.

References

Motasem Alrahabi and Tom Wainstain. 2025. Versus: An automatic text comparison tool for the digital humanities. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 32–37, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

David Armitage and Jo Guldi. 2014. *The History Manifesto*. Cambridge University Press.

Emmanuelle Bermès. 2017. Text, data and link-mining in digital libraries: Looking for the heritage gold. In *IFLA Satellite Meeting - Digital Humanities – Opportunities and Risks: Connecting Libraries and Research*.

David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Katherine Bode. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, Ann Arbor.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 803–812, Marseille, France. European Language Resources Association.

Simon Burrows and Mark Curran. 2012. The French Book Trade in Enlightenment Europe Project and the STN Database. *Journal of Digital Humanities*, 1(3).

Daniel Carey. 2006. *Locke, Shaftesbury, and Hutcheson: Contesting Diversity in the Enlightenment and Beyond*. Ideas in Context. Cambridge University Press, Cambridge.

Jeffrey R. Collins. 2020. *In the Shadow of Leviathan: John Locke and the Politics of Conscience*. Cambridge University Press. Google-Books-ID: 1tLKDwAAQBAJ.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The FAISS Library. *IEEE Transactions on Big Data*, pages 1–17.

Marten Düring, Matteo Romanello, Maud Ehrmann, Kaspar Beelen, Daniele Guido, Brecht Deseure, Estelle Bunout, Jana Keck, and Petros Apostolopoulos. 2023. Impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. *Frontiers in Big Data*, 6.

Dan Edelstein. 2016. Intellectual History and Digital Humanities. *Modern Intellectual History*, 13(1):237–246.

Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. *The American Historical Review*, 122(2):400–424.

Susan Fitzmaurice, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl, and Fraser Dallachy. 2017. Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse. *Studia Neophilologica*, 89(sup1):21–38.

Brevin Franklin, Emily Silcock, Abhishek Arora, Tom Bryan, and Melissa Dell. 2024. News Deja Vu: Connecting Past and Present with Semantic Search. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 99–112, Mexico City, Mexico. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Andrew Goldstone and Ted Underwood. 2014. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, 45(3):359–384.

Tyler Hanck. 2019. *Locke's Confusion About the Confused Idea of Substance*. Thesis, University of Illinois Chicago.

James A. Harris. 2023. Of the origin of government: The afterlives of Locke and Filmer in an eighteenth-century British debate. *Intellectual History Review*, 33(1):33–55.

Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti, and Mikko Tolonen. 2019. Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material. In *Proceedings of the Digital Humanities in the Nordic Countries*, Copenhagen, Denmark.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Jenna Kanerva, Hanna Kitti, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter. 2025. Semantic search as extractive paraphrase span detection. *Language Resources and Evaluation*, 59(1):257–276.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Pierre-Carl Langlais. 2021. Fictions littéraires de Gallica / Literary fictions of Gallica.

Daniel Layman. 2021. Locke's Republican and Liberal Legacy. In *The Lockean Mind*. Routledge. Num Pages: 10.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d' Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the Gap: Assessing Temporal Generalization in Neural Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

H. P. Luhn. 1960. Key word-in-context index for technical literature (kwic index). *American Documentation*, 11(4):288–295.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the Behavior of Neural IR Models. *Transactions of the Association for Computational Linguistics*, 10:224–239.

Ananth Mahadevan, Michael Mathioudakis, Eetu Mäkelä, and Mikko Tolonen. 2025. Text reuse in large historical corpora: Insights from the optimization of a data science system. *International Journal of Data Science and Analytics*, 20(5):4631–4643.

Barbara McGillivray, Federico Nanni, and Kaspar Beelen. 2024. 10. Why Does Digital History Need Diachronic Semantic Search? In *Computational Humanities*. University of Minnesota Press.

Andrianos Michail, Juri Opitz, Yining Wang, Robin Meister, Rico Sennrich, and Simon Clematide. 2025. Cheap Character Noise for OCR-Robust Multilingual Embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11705–11716, Vienna, Austria. Association for Computational Linguistics.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A Two-stage Sieve Approach for Quote Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Eetu Mäkelä, James Misson, Devani Singh, and Mikko Tolonen. 2025. Opening the Black Box of EEBO. *Digital Scholarship in the Humanities*, page fqaf086.

R. Porter. 2001. *Enlightenment: Britain and the Creation of the Modern World*. Penguin Books Limited.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Hammad Rizwan, Domenic Rosati, Ga Wu, and Hassan Sajjad. 2025. Resolving Lexical Bias in Model Editing. In *Forty-Second International Conference on Machine Learning*.

Glenn Roe. 2024. Text reuse as cultural practice: Intertextuality in the 18th-century digital archive. *Digital Enlightenment Studies*, 2(1).

David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. Reception Reader: Exploring Text Reuse in Early Modern British Publications. *Journal of Open Humanities Data*, 9:5. ArXiv:2302.04084 [cs].

Julian Schelb, Michael Wittweiler, Marie Revellio, Barbara Feichtinger, and Andreas Spitz. 2026. Loci Similes: A Benchmark for Extracting Intertextualities in Latin Literature. *Preprint*, arXiv:2601.07533.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David A. Smith, Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 16857–16867, Red Hook, NY, USA. Curran Associates Inc.

Peter M. Stahl. 2021. Lingua: The most accurate natural language detection library for Python. `https://github.com/pemistahl/lingua-py`. Python bindings for the Lingua language detection library.

Timothy Stanton. 2018. John Locke and the Fable of Liberalism. *The Historical Journal*, 61(3):597–622.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Iiro Tiihonen and Kira Hinderks. 2025. Genre Classification Workflow for the English Short Title Catalogue (ESTC). *Transformations: A DARIAH Journal*, Workflows(Metadata-based workflows):14754.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Mikko Tolonen, Eetu Mäkelä, and Leo Lahti. 2022. The Anatomy of Eighteenth Century Collections Online (ECCO). *Eighteenth-Century Studies*, 56(1):95–123. Publisher: Johns Hopkins University Press.

Mikko Tolonen and Yann Ciarán Ryan. 2026. *Computational Methods in Intellectual History*, pages 239–260. Proceedings of the British Academy. Liverpool University Press, United Kingdom.

Mikko Tolonen and Mark G. Spencer. 2025. The Reception of David Hume's Essays in Eighteenth-Century Britain. In Max Skjönsberg and Felix Waldmann, editors, *Hume's Essays*, Cambridge Critical Guides, pages 15–35. Cambridge University Press, Cambridge.

Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala, and Tapio Salakoski. 2017. A System for Identifying and Exploring Text Repetition in Large Historical Document Corpora. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 330–333, Gothenburg, Sweden. Association for Computational Linguistics.

Christopher N. Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. 2016. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, 10(3):1.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17443–17453.

Tracing Thematic Change in Early English-Language Science Fiction, 1818–1930

Jonathan Gordon
Department of Computer Science
Vassar College
Poughkeepsie, NY
jgordon@vassar.edu

Abstract

How did the thematic repertoire of early English-language science fiction change as the genre consolidated between 1818 and 1930? Using a corpus of 238 public-domain texts, we apply temporally binned latent Dirichlet allocation (LDA), comparing models with and without Authorless preprocessing (which probabilistically downweights author-specific vocabulary). Cross-period topic alignments exceed a permutation null baseline, indicating continuity in topic structure over time. Full-corpus LDA can produce comparable per-topic quality, but only temporal binning enables diachronic alignment; within the binned setting, Authorless reduces author concentration and modestly increases the share of thematic topics without materially reducing coherence. Four high-continuity topic chains – centered on mobility, affect, planetary scale, and scientific knowledge – suggest a shift from earlier romantic and speculative concerns toward more consolidated technoscientific forms. These chains generate interpretable hypotheses about the literary history of early science fiction, and the workflow supports diachronic analysis in small, author-skewed corpora.

1 Introduction

Science fiction (SF) emerged as a recognizable literary mode in the nineteenth century, consolidating into a commercial genre by the early twentieth. The period from Mary Shelley's *Frankenstein* (1818) through the rise of pulp magazines in the 1920s marks a formative phase in which recurring concerns – technological change, exploration, and the nature of scientific knowledge – became legible as genre-wide patterns. Yet our understanding of how these themes evolved remains largely dependent on case studies of canonical texts rather than systematic analysis of a wider body of texts.

This paper investigates how the thematic repertoire of early English-language science fiction changed as the genre consolidated between 1818 and 1930. We address this question using latent Dirichlet allocation (LDA; Blei et al., 2003), with the goal of generating hypotheses about SF history rather than testing predetermined claims.

Topic models applied to literary corpora often recover *known structure* – such as authorship, individual works, or paratext – rather than the interpretive structure a researcher hopes to study (Rhody, 2013; Schmidt, 2013; van Zundert et al., 2022). The problem is especially acute in small corpora with high author concentration, where models may distinguish authors more readily than themes (Thompson and Mimno, 2018). To address this, we combine two strategies: we train separate LDA models on four temporal bins spanning 1818–1930 to reduce corpus heterogeneity, and we apply Thompson and Mimno's (2018) Authorless transformation to downweight author-specific vocabulary.

This paper makes two contributions. Methodologically, it shows that temporally binned models support diachronic analysis in this corpus, and that Authorless preprocessing reduces single-author dominance within that setting without materially reducing interpretability. Substantively, it offers an exploratory account of thematic change in early English-language science fiction and identifies four recurring cross-period topic chains. We treat these chains as interpretable lexical structures rather than direct equivalents of literary themes; their value lies in supporting comparative reading and generating hypotheses for literary history.

2 Related Work

In a major early study of large-scale literary topic modeling, Jockers and Mimno (2013) identified thematic clusters correlated with author gender, nationality, and publication date in a large corpus of nineteenth-century novels. Their preprocessing recommendations – notably chunking, named-entity

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 226–235
July 4, 2026 ©2026 Association for Computational Linguistics

removal, and author/title filtering – inform our approach (§3.1); Sobchuk and Šeļa (2024) confirmed that such choices substantially affect resulting clusters.

Two especially relevant SF studies are Thompson and Mimno (2018) and Bologna (2020). Thompson and Mimno applied LDA to 1,206 SF novels, demonstrating that author signal poses a significant problem for literary topic modeling and proposing Authorless Topic Models as a solution. Their study addresses the SF domain but not diachronic thematic history; we adopt their mitigation strategy in service of temporal interpretation. Bologna used decade-by-decade LDA on twentieth-century SF to trace temporal lexical patterns. We extend this approach to an earlier starting point (1818–1930) and use cross-period topic alignment rather than lexicon-driven tracking.

As these and other scholars have argued, topics capture lexical regularities rather than literary themes per se – and may even recover metadata such as genre labels rather than interpretive structure (Schmidt, 2013; Rhody, 2013; van Zundert et al., 2022). We operationalize this caution by evaluating topics along three analytically separate dimensions – coherence, thematic content, and author concentration – and through close-reading vignettes that ground the four chains in primary texts. This paper brings together temporal binning, author-signal reduction, and explicit thematic evaluation for diachronic analysis of early English-language science fiction.

3 Corpus and Methods

3.1 Corpus Construction

We derive a public-domain SF corpus from a Hugging Face dataset of Project Gutenberg science fiction texts[1] and attach publication-year metadata from Momen et al. (2025), who used LLM-based inference to recover publication years for Project Gutenberg texts. In a spot check of 20 randomly sampled works, 17 were dated within ±5 years of a reference year; the remaining 3 would shift bins. (Those three were dime-novel reprints, where the metadata reflects reprint dates rather than original serialization.) Because our analysis operates on four broad bins rather than year-level trends, individual dating errors are more likely to affect bin assignments than the overall logic of periodization.

[1] huggingface.co/datasets/stevez80/Sci-Fi-Books-gutenberg

After filtering for the 1818–1930 window and removing texts for which the year estimator returned no value, our corpus contains 238 books. Texts are segmented into non-overlapping chunks of approximately 1,000 words (14,973 chunks total), following common practice for stabilizing co-occurrence statistics while keeping topics locally interpretable (Jockers and Mimno, 2013).

Table 1 summarizes the distribution across four temporal bins, chosen to balance literary-historical periods and data availability. The pre-1880 bin covers the Verne-dominated early period; later bins roughly track the rise of the scientific romance (1880–1899), Edwardian SF (1900–1914), and the early pulp era (1915–1930). Author concentration varies sharply across bins: the pre-1880 period is dominated by Jules Verne (42.5% of chunks), a share further inflated by duplicate editions with distinct volume titles that remain in the corpus. Findings from this bin therefore require particular caution; later bins have more balanced distributions.

3.2 Why Temporal Binning?

Initial whole-corpus LDA runs were unstable (mean Jaccard similarity 0.26–0.31 across random seeds, below the stability thresholds discussed by Greene et al., 2014) and remained heavily shaped by author signal despite ablations over chunk size, vocabulary filters, and deduplication. Although a higher-capacity full-corpus model later yields comparable topic quality, it cannot support the paper's central aim: cross-period alignment of thematic structure. We therefore fit separate models within more homogeneous temporal bins and trace change through alignment across adjacent periods.

3.3 Preprocessing, Modeling, and Evaluation

We apply standard preprocessing intended to reduce non-thematic structure: a document-frequency filter retaining tokens appearing in ≥10 and ≤50% of chunks (Schofield et al., 2017), supplemented by a 178-word stoplist, without lemmatization (Schofield and Mimno, 2016). Proper names are removed via a capitalization heuristic (mid-sentence capitalized words), and author/title tokens are filtered where present (Jockers, 2013). We fit LDA models using MALLET (McCallum, 2002) with 1,000 iterations, asymmetric hyperparameter optimization every 10 iterations, and a fixed random seed, with $k = 15$ topics per bin, balancing interpretability against thematic coverage at these

Bin	Books	Chunks	Authors	Top author share
Pre-1880	23	2,143	13	42.5%
1880–1899	55	3,253	42	10.0%
1900–1914	72	4,323	49	7.6%
1915–1930	88	5,254	53	15.3%
Total	238	14,973	–	–

Table 1: Corpus distribution by temporal bin.

bin sizes (2,143–5,254 chunks per bin). To assess the contribution of temporal binning itself, we also fit MALLET models to the full corpus, at $k = 15$ (matching the per-bin topic count) and at $k = 60$ (matching the total across all four bins), using the same preprocessing and settings.

To isolate the additional effect of author-signal reduction within bins, we fit two versions of each temporally binned model: standard MALLET and MALLET with Authorless preprocessing, implemented via MALLET's `DownsampleLabelWords` transform with a threshold of 0.05, which probabilistically removes tokens associated with author labels before model fitting (Thompson and Mimno, 2018). The close-reading analysis below uses the Authorless models, but we evaluate all four configurations in parallel.

Evaluation framework. We evaluate all 195 topic instances across the four configurations on three analytically separate dimensions: *coherence*, *thematic content*, and *author concentration*. We rely on manual judgment of top words together with representative chunks rather than on automated coherence metrics. Studies dating back to Chang et al. (2009) have documented systematic disagreements between automated coherence scores and human ratings of topic interpretability, and more recent work confirms that metrics such as NPMI remain unreliable proxies for human judgment in both classical and neural topic models (Doogan and Buntine, 2021; Hoyle et al., 2021, 2025). Manual inspection of top words combined with representative documents remains standard practice in DH topic modeling (Antoniak, 2022); we adopt it here.

Coherence. A topic is coded as *coherent* (yes/no) if its top words form a recognizable semantic field – that is, a set of words that co-occur because they refer to overlapping or related entities, settings, or activities. The 1915–1930 topic with top words *ship, air, water, plane, sea, deck, boat, speed, cabin, black* is coherent: the words name interrelated transportation surfaces and vehicles. By contrast, the 1880–1899 topic with top words *man, good, thing, room, time, day, made, back, put, make* is not: the list mixes high-frequency narrative and dialogue vocabulary without naming a shared domain.

Thematic content. A coherent topic is additionally coded as *thematic* (yes/no) if its semantic field picks out recurring subject matter – settings, entities, situations, or conceptual domains – rather than a linguistic mode, register, or paratext signal. Recurrence here is judged within the bin: whether the top words and high-loading chunks point to a domain that recurs across documents within the same period, independent of whether the same domain recurs across bins (the separate question of chain alignment in §3.4). The mobility topic above is thematic. A topic dominated by markers of poetic or archaic diction – for example, the 1880–1899 topic *love, thou, thy, thee, people, life, soul, heart, hath, art* – may be coherent as a linguistic register, but is not thematic in this sense; nor are topics whose top words are dominated by editorial paratext such as chapter labels or volume markers. Coherence and thematic content are intended as analytically separate dimensions: a topic can be coherent without being thematic. The dependency is one-sided – we do not code an incoherent topic as thematic, since thematic interpretation depends on a recognizable semantic field in the first place.

Author concentration. *Author concentration* is the fraction of a topic's top-5 highest-loading chunks contributed by its dominant author, ranging from 0.2 (each chunk from a different author) to 1.0 (all from the same author). It captures whether a topic's lexical pattern reflects a genre-wide regularity or one author's idiom. We treat it as a property that is reported alongside coherence and thematic content rather than collapsed into a single quality

label, since a topic can be both thematic and author-concentrated, and conflating these obscures what the Authorless transformation is actually doing.

Procedure. Coherence and thematic-content labels were assigned by a single annotator in a single pass from the top words and top-5 highest-loading chunks for each topic, conducted before chain selection so that classification was not influenced by downstream analytical choices. Coherence was judged first as a rough filter: topics judged incoherent received no thematic label, since thematic interpretation depends on a recognizable semantic field. Author concentration is computed automatically. We have not measured inter-annotator agreement; the implications of relying on a single annotator are discussed in §6.

We compare all four configurations across these three dimensions. Results appear in Tables 2 and 3.

3.4 Cross-Period Alignment

After fitting one model per bin, we align topics between adjacent periods using cosine similarity on L2-normalized topic-word distributions over their shared vocabulary (i.e., words appearing in both bins). Given a similarity matrix between topics in adjacent bins, we compute a one-to-one matching that maximizes total similarity using optimal bipartite matching (Hungarian algorithm). We use one-to-one matching as a conservative strategy, though it cannot capture topic splits or merges across periods.

We evaluate matched mean similarity against a null distribution generated from 1,000 random permutation matches per transition. Z-scores are computed as (matched mean − null mean) / null SD; p-values are derived from the standard normal distribution rather than directly from permutation counts. Results are summarized both at the level of adjacent-bin means and as multi-bin *chains* obtained by tracing matched topics across all four periods.

4 Results

4.1 Alignment Exceeds Null Baselines

Across all adjacent-bin pairs, matched topics show substantially higher similarity than expected under random matchings (Table 4). This indicates continuity in topic structure above chance, though continuity alone does not establish that the aligned topics are thematically meaningful. The weaker pre-1880 alignment (0.44 vs 0.58 for later pairs) likely reflects the bin's smaller size and high author concentration (Verne accounts for 42.5% of pre-1880 chunks).

4.2 Model Comparison: Binning, Authorless, and Full-Corpus Baselines

Table 2 compares all four configurations. Coherence rates are stable across models (80–85%), suggesting that topic quality is robust to these modeling choices. Thematic content is lowest for the full-corpus k=15 model (47%), confirming that 15 topics are insufficient for the full corpus's diversity. At k=60, however, the full-corpus model yields 62% thematic topics – comparable to binned MALLET (60%) and only modestly below binned Authorless (65%). Author concentration is actually *lower* in the full-corpus models (0.680–0.707) than in either binned pipeline (0.727–0.807), likely because no single author dominates the full corpus as strongly as Verne dominates the pre-1880 bin.

These results show that the primary contribution of temporal binning is not topic quality per se, but temporal structure: full-corpus models produce comparably thematic topics but cannot support diachronic analysis, because a single model cannot distinguish themes that emerge, evolve, or recede across periods. Binning is therefore necessary for the paper's central goal of tracing thematic change.

Within the binned setting, Authorless provides a targeted refinement. The clearest effect is on author concentration, though thematicity also rises modestly. Mean concentration falls from 0.807 to 0.727 overall (Table 5); the number of topics whose top-5 chunks all come from a single author falls from 30 to 23, and the mean number of unique authors per topic rises from 1.80 to 2.27. Authorless therefore does not mainly make topics more coherent; it makes them more representative of cross-author patterns. A topic dominated by a single author's vocabulary is stronger evidence about that author than about the genre.

Topic-quality gains are modest and uneven across periods (Table 3). Authorless improves the number of coherent-and-thematic topics in three of the four bins, with the largest gain in the pre-1880 bin (+2 topics, 53% to 67%), consistent with that bin's high Verne concentration. Concentration reductions follow a complementary pattern, strongest where author dominance is highest (Table 5).

We therefore use the binned Authorless models for the chain analysis that follows.

Model	Topics	Coherent	Thematic	Concentration
Full corpus k=15	15	12 (80%)	7 (47%)	0.680
Full corpus k=60	60	48 (80%)	37 (62%)	0.707
Binned MALLET	60	51 (85%)	36 (60%)	0.807
Binned Authorless	60	50 (83%)	39 (65%)	0.727

Table 2: Comparison of model configurations. Coherent = top words form a semantic field; thematic = coherent topic whose semantic field captures subject matter rather than register or paratext; concentration = mean fraction of top-5 chunks from dominant author (lower = more cross-author).

Bin	MALLET	Authorless	Difference
Pre-1880	8 (53%)	10 (67%)	+2
1880–1899	8 (53%)	9 (60%)	+1
1900–1914	10 (67%)	9 (60%)	−1
1915–1930	10 (67%)	11 (73%)	+1
Overall	**36 (60%)**	**39 (65%)**	**+3**

Table 3: Coherent-and-thematic topics by temporal bin.

4.3 Four Thematic Chains with Interpretable Literary-Historical Trajectories

Among multi-bin chains, we focus on four that meet two criteria. They maintain ≥0.50 mean cosine similarity across all three transitions and are interpretable as recurring content domains rather than register or author signal across all four bins. Of the 15 chains traceable across all four periods, seven meet the similarity threshold; of these, four are consistently thematic. The remaining three chains at or above the threshold are dominated by register or include paratext in one or more bins.

The close readings below illustrate how each chain's lexical signature manifests in specific highest-loading chunks rather than surveying period-wide content. The trajectories in Table 6 are interpretive labels derived from inspecting top words and top-5 highest-loading chunks together.

Chain 1: Naval/Transportation Technologies The earliest texts (Verne) treat the ship as a physical body embedded in natural forces: "the copper sheathing and the planks disappeared, reduced, no doubt, to powder" (*The Mysterious Island*, 1875). By the 1880s–1890s, emphasis shifts to velocity and strategic pursuit. In the 1900s, mobility collapses into precarious survival amid hostile spaces. By 1915–1930, the chain's highest-loading chunks treat sea, air, and space as a single operational field: "Straight down from a wisp of golden cloud a slim black speck fell toward the earth … The black flyer hung motionless" (Murray Leinster, *Murder Madness*, 1930).

Chain 2: Love, Affect, and Moral Interiority The pre-1880 anchor is dominated by Mary Shelley's *Frankenstein* (1818), where love entangles with guilt and functions as inward torment. By the 1880s–1890s, romance becomes socially mediated through family authority. The 1900–1914 period introduces philosophical stakes, while in 1915–1930 the chain's highest-loading chunks extend romantic vocabulary into cosmic registers: "Try to measure if you can, my princess, a love so vast that it draws its mate across the space between the stars" (J. U. Giesy, *Palos of the Dog Star Pack*, 1921).

Chain 3: Elemental Vastness and Planetary Forces Verne's texts frame the earth as an active, unfinished structure whose forces operate on human timescales: "It was not a question of months, but of days, it might be of hours" (*The Secret of the Island*, 1875). The 1880s–1890s shift to exploration of hidden worlds. By 1900–1914, scale becomes astronomical. The final period brings post-catastrophic landscapes among the chain's highest-loading chunks: "Slowly vegetation is creeping back upon the face of the world; but still there are vast deserts where no blade grows" (J. J. Connington, *Doomed World*, 1923).

Chain 4: Science, Knowledge, and Reality Pre-1880 frames science as moral authority and progres-

Transition	Matched	Null	Z	p
Pre-1880 → 1880–1899	0.441	0.220	7.04	$< 10^{-4}$
1880–1899 → 1900–1914	0.583	0.240	8.87	$< 10^{-4}$
1900–1914 → 1915–1930	0.586	0.252	9.06	$< 10^{-4}$
Overall	**0.537**	**0.237**	–	–

Table 4: Cross-period topic alignment. Means are cosine similarities between matched topics; null means are from random matchings; p-values are approximate (normal approximation to the permutation null).

Bin	MALLET	Authorless	Difference
Pre-1880	0.920	0.867	−0.053
1880–1899	0.867	0.693	−0.174
1900–1914	0.827	0.747	−0.080
1915–1930	0.613	0.600	−0.013
Overall	**0.807**	**0.727**	**−0.080**

Table 5: Mean author concentration by temporal bin. Lower values indicate topics that draw on a broader range of authors.

sive mastery: "So much has been done, exclaimed the soul of Frankenstein,—more, far more, will I achieve" (Mary Shelley, *Frankenstein*, 1818). By the 1880s–1890s, Camille Flammarion introduces epistemic humility: "The real nature of things entirely escapes your understanding" (*Lumen*, 1887). In 1900–1914, the top words are less lexically distinctive, but the highest-loading texts center on the epistemology of scientific observation – how evidence is marshaled and what instruments can prove. By 1915–1930, the chain's highest-loading chunks render scientific knowledge in increasingly instrumental terms: "It is not matter at all, in the ordinary sense of the word. It is almost pure crystallized energy" (E. E. Smith, *The Skylark of Space*, 1928).

4.4 Comparison with Established Genre Categories

As a qualitative point of comparison, we related selected thematic topics to established categories in the *SF Encyclopedia* (Clute et al., 2021), the standard genre-taxonomic reference. Table 7 shows mappings for topics classified as thematic; the four focal chains are marked with an asterisk.

Notable absences include time travel, aliens, and robots. These may reflect the limits of bag-of-words topics for plot-level motifs: concepts organized by narrative structure rather than lexical co-occurrence may not surface as distinct topics even when present in the corpus. Robots illustrate a different mechanism: the English translation of Čapek's *R.U.R.* (1923), which introduced the word "robot" to English-language SF, is in the corpus, but a single work cannot generate enough co-occurrence signal to form a distinct topic.

5 Discussion

5.1 Implications for SF Literary History

The four chains motivate hypotheses about how early SF reorganized recurring domains of concern across the nineteenth and early twentieth centuries. Each is consistent with – but not uniquely entailed by – the lexical evidence in Table 6 and the close readings of §4.3.

Mobility and environment. The chain is consistent with a shift from movement constrained by natural forces to movement organized by technological systems, in which the ship moves from a physically vulnerable object to a node in a larger system of signals, sensors, and weapons. The appearance of *plane* and *speed* among the 1915–1930 top words marks this shift lexically. The trajectory aligns with genre historians' emphasis on the evolution from nautical adventure (Verne) through aerial warfare to space opera.

Interiority. The chain suggests that the psychological and moral vocabulary anchored in early canonical works – notably *Frankenstein* – becomes a recurring resource: moral torment, then courtship, then a device for staging romantic devotion at cosmic scale. The lexical movement is visible across

Chain	Period	Top Words
1: Naval/ Transportation	Pre-1880	vessel, sea, ship, boat, shore, coast, wind, board, land, water
	1880–1899	ship, air, long, speed, time, miles, boat, deck, water, feet
	1900–1914	boat, ship, sea, water, deck, men, wind, shore, night, vessel
	1915–1930	ship, air, water, plane, sea, deck, boat, speed, cabin, black
2: Love/ Affect	Pre-1880	life, love, heart, father, mind, feelings, day, felt, death, thought
	1880–1899	young, girl, woman, eyes, love, father, face, mother, life, wife
	1900–1914	life, man, world, thought, love, day, heart, knew, long, night
	1915–1930	man, eyes, woman, face, love, girl, father, life, hand, voice
3: Planetary/ Elemental	Pre-1880	uncle, water, feet, volcano, lava, great, hundred, light, sea, surface
	1880–1899	water, sea, great, mountain, land, trees, lake, beneath, ice, rock
	1900–1914	water, feet, miles, sun, high, earth, great, long, hundred, mountain
	1915–1930	water, night, sun, long, day, sky, lay, wind, trees, miles
4: Science/ Knowledge	Pre-1880	life, great, men, years, time, world, man, people, country, science
	1880–1899	life, man, human, world, mind, body, nature, time, existence, power
	1900–1914	time, world, fact, made, long, work, years, life, great, planet
	1915–1930	air, space, power, light, time, machine, laboratory, metal, earth, work

Table 6: Top words for the four thematic chains across temporal bins. Trajectories: (1) Material vulnerability → Speed/pursuit → Fragile survival → Hybrid battlespace; (2) Moralized torment → Social mediation → Philosophical stakes → Cosmic devotion; (3) Natural sublimity → Hidden worlds → Astronomical scale → Post-catastrophe; (4) Rational progress → Epistemic humility → Scientific epistemology → Instrumentalized technoscience.

Our Topic	SFE Theme(s)
Naval/Transportation*	Airships, Submarines, Transportation
Love/Affect*	(not SF-specific)
Planetary/Elemental*	Mars, Moon, Astronomy, Space Flight
Science/Knowledge*	Invention, Scientists, Hard SF
War/Military	Future War, Invasion, War
Utopia	Utopias
Primitive world	Prehistoric SF, Lost Races

Table 7: Topic mapping to SF Encyclopedia themes. Asterisks indicate the four focal chains.

the top words: pre-1880 emphasizes interiority (*feelings, mind, death, thought*); 1880–1899 introduces kinship vocabulary (*girl, woman, mother, wife*); 1900–1914 is comparatively undifferentiated, mediating between the two; and 1915–1930 settles into embodied courtship terms (*eyes, face, voice, hand*). On this reading, romance is not a "non-SF" residue but a stable substrate through which SF stages novelty – implying continuities across what might otherwise look like distinct traditions in nineteenth-century scientific romance and twentieth-century pulp fiction.

Scale. The chain is consistent with a long-standing intuition in SF criticism that the genre is organized by scale shifts, and specifies one historical sequence: terrestrial geology gives way to subterranean exploration, then planetary astronomy, and finally post-catastrophic landscape. The lexical shift is visible across the four bins: *volcano, lava, surface* (pre-1880); *mountain, beneath, ice, rock* (1880–1899); *sun, earth, miles, hundred* (1900–1914); and *sky, wind, day, night, lay* (1915–1930), where distinctive geological terms have receded from the top of the distribution. The trajectory may reflect a broader tension between progress narratives and anxieties about civilizational fragility.

Science and instrumentality. The chain traces the familiar genre-historical shift from "scientific romance" to "hard SF" – from speculative philosophy and epistemic reflection toward laboratory-centered instrumentality, where physics is narrated as power over matter and perception. The lexical contrast is sharpest between 1880–1899 (*mind, body, nature, existence*) and 1915–1930 (*laboratory, machine, metal, space, power*): abstract philosophical reflection gives way to material instrumentality. This trajectory anticipates later pulp technoscience while remaining legible as an evolution rather than a sudden break.

Following Schmidt (2013), we emphasize that these patterns are shifts in vocabulary co-occurrence, not themes extracted from narrative. That they align with themes documented in traditional criticism is notable, but does not validate a naïve equation of topics with themes. The lexical evidence above shows what each hypothesis is grounded in; the same top-word trajectories nonetheless admit alternative readings, and we present them as *hypotheses suggested by topical structure* – compatible with close reading but not uniquely entailed by it. They motivate targeted genre-historical inquiry: for instance, the relationship between aviation warfare, pulp publishing venues, and the spatial settings of late-period SF.

5.2 Methodological Implications

For DH researchers working with specialized literary or historical corpora, the main lesson is practical: temporal binning is necessary when the analytical goal is diachronic, not because it produces better topics in the aggregate. Within a binned framework, preprocessing choices that suppress author signal – such as Authorless – can further shape what kinds of interpretive questions topic modeling supports. In this corpus, Authorless did not transform a poor model into a good one; it made already workable temporally binned models better suited to genre-level interpretation by broadening the author base of topics.

Genre-specific collections, author archives, and periodical runs frequently exhibit the kind of author concentration that standard topic modeling handles poorly. Authorless preprocessing should therefore be considered when the goal is thematic rather than stylistic analysis, particularly in combination with corpus segmentation: segmentation (such as temporal binning) addresses broader sources of heterogeneity, while Authorless targets the residual author concentration within each segment.

Neither intervention is without cost. Temporal binning requires post-hoc alignment to enable diachronic interpretation, and one-to-one matching is a simplification that may miss topic splitting or merging across periods. The Authorless transformation does not eliminate paratext topics or guarantee thematic coherence; it shifts distributional properties rather than substituting for interpretive judgment.

6 Limitations

Interpretive subjectivity. Topic classification and vignette selection were performed by a single annotator; different readers may label or interpret topics differently. The rubric described in §3.3 is designed to make the criteria explicit enough for replication and disagreement analysis, but inter-annotator agreement has not been measured. Two natural extensions are left to future work: a multi-annotator pass on a sample of topics with formal agreement analysis, and a parallel LLM-as-judge evaluation against the same rubric.

Concentration metric granularity. Author concentration is measured over only five chunks per

topic, so the metric takes only five possible values (0.2, 0.4, 0.6, 0.8, 1.0). Means smooth this coarseness across topics, but the per-topic measure is inherently approximate.

Temporal metadata uncertainty. Publication years include inferred values (Momen et al., 2025), which can introduce binning error. As noted above, bin-level periodization is more robust to individual date errors than year-level analysis.

Corpus representativeness and author concentration. The corpus skews late: 67% of books come from 1900–1930. Pre-1880 is dominated by a small number of authors, especially Verne (42.5% of chunks); some apparent thematic structure may partially reflect canon and availability rather than the full field of early SF. As a rough coverage check, only 9 of 25 well-known early SF works we looked for were present in the source dataset.

Model sensitivity and reproducibility. All models use a fixed random seed and identical hyperparameter settings, and we report Jaccard stability for initial whole-corpus runs (§3.2). However, we do not report formal sensitivity analysis over the number of topics k or random seeds for the binned Authorless models due to the complexity of controlling preprocessing stochasticity. Alternative binning schemes (e.g., equal-width decades) were also not tested, so results may be sensitive to bin boundaries.

Duplicate and paratext contamination. Some multi-volume editions include overlapping text; some chunks include front matter or other paratext, potentially distorting topics.

Bag-of-words constraints. Plot-level themes (e.g., time travel) and relational concepts (e.g., first contact) may not appear as distinct topics even if present in the corpus.

7 Conclusion

This paper shows that temporally binned topic modeling can support diachronic analysis in a small, author-skewed literary corpus. A comparison of four model configurations finds that full-corpus LDA at adequate capacity produces comparable per-topic quality, but only binned models enable cross-period alignment. Within that setting, Authorless preprocessing reduces author concentration and modestly increases thematicity without materially reducing coherence. Applied to early English-language science fiction (1818–1930), the workflow yields four high-continuity chains – centered on mobility, affect, planetary scale, and scientific knowledge – whose alignments exceed chance and motivate hypotheses about thematic change across the period.

Methodologically, the paper separates two preprocessing decisions that DH topic modeling often runs together. Temporal binning does not improve topic quality over a well-configured full-corpus model, but it makes diachronic interpretation possible. Author-signal reduction is a separate move: it does not enable new analyses, but it makes the resulting topics more representative of cross-author patterns. The evaluation framework introduced here – treating coherence, thematic content, and author concentration as analytically separate dimensions – supports both moves by reporting their effects separately.

The aligned chains are not direct representations of literary themes, but interpretable topic structures that support comparison, close reading, and hypothesis generation. Future work could extend the analysis to later periods, strengthen validation through multi-annotator and LLM-as-judge evaluation, and test specific hypotheses these chains motivate – for instance, the relationship between aviation warfare, pulp publishing venues, and the spatial settings of late-period SF.

Data and Code Availability

The source corpus is derived from public-domain Project Gutenberg texts and is available as a Hugging Face dataset (§3.1). Topic-level judgments (coherence, thematic content, author concentration) and analysis code will be released at `github.com/textproclab/sf-topics`.

References

Maria Antoniak. 2022. Topic modeling for the people. Blog post.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Federica Bologna. 2020. A computational approach to urban space in science fiction. *Journal of Cultural Analytics*, 5(2).

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

John Clute, David Langford, and Peter Nicholls. 2021. The encyclopedia of science fiction. Ansible Editions. Accessed 2025.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Berlin, Heidelberg. Springer.

Alexander Hoyle, Lorena Calvo-Bartolomé, Jordan Boyd-Graber, and Philip Resnik. 2025. ProxAnn: Use-oriented evaluations of topic models and document clustering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15872–15897, Vienna, Austria. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Matthew L. Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

Omar Momen, Manuel Schaaf, and Alexander Mehler. 2025. Filling the temporal void: Recovering missing publication years in the Project Gutenberg corpus using LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17318–17334, Vienna, Austria. Association for Computational Linguistics.

Lisa M. Rhody. 2013. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1).

Benjamin M. Schmidt. 2013. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1).

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Oleg Sobchuk and Artjoms Šeļa. 2024. Computational thematics: Comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, 11:1–12.

Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Joris J. van Zundert, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Mussmann. 2022. What do we talk about when we talk about topic? In *Computational Humanities Research Conference*, pages 1–18.

Twenty's Plenty: Semantic Scaffolding and Span Architecture for 19-Label NER in Medieval Latin Charters

Tamás Kovács **Giuseppe Consolo** **Georg Vogeler**
Department of Digital Humanities
University of Graz
{tamas.kovacs, giuseppe.consolo, georg.vogeler}@uni-graz.at

Abstract

We ask whether a few hundred annotated sentences are sufficient to build a high-quality named entity recogniser for medieval Latin charters spanning 19 fine-grained entity types — including complex legal clauses, property boundary descriptions, and rare fiscal categories. We evaluate three tiers of approach: zero-shot GLiNER baselines, fine-tuned GLiNER variants (including a domain-pretrained medieval Latin model), and a custom span-based architecture pairing XLM-RoBERTa-large with an Asymmetric Focal–Dice loss and BGE-M3 label encodings. Our results show that *semantic scaffolding* — providing richly descriptive English label phrases — activates latent multilingual knowledge and enables fine-tuned GLiNER to reach 80.8% overlap F1 with no oversampling. Critically, domain pre-training on medieval Latin offers no advantage once task-specific fine-tuning is applied (77.1% vs. 77.2%). A custom span model further surpasses all GLiNER variants at 83.4% overlap F1, with persistent challenges on rare and structurally complex categories.

1 Introduction

Medieval Latin charters are the core sources for Central European history — recording land transfers, judicial decisions, and social hierarchies in a formulaic Latin that is computationally challenging: non-standardised orthography, heavy abbreviation, and dense entity structures in which a single identified person (ACTOR) simultaneously encodes a personal name (PER), an office (TITLE), and an institutional affiliation (INS).

Annotating such material at scale is expensive. Expert medievalists are few, and gold-standard annotation for fine-grained NER requires months to stabilise. The practical question is therefore: *how much annotated data do we actually need, and can architectural choices compensate for data scarcity?*

We exploit a property of GLiNER (Zaratiana et al., 2024): the *label-as-prompt* mechanism, where the semantic content of an entity type description becomes part of the model input and directly shapes span detection. By crafting English descriptions that encode domain expertise — *"legal clause declaring rights, conditions, penalties, or papal commands"* rather than LEG — we activate knowledge latent in the model's multilingual space, routing Latin surface forms through English conceptual anchors. We call this strategy **semantic scaffolding**.

Existing systems for historical Latin NER cover at most two entity types (Chastang et al., 2021; Ehrmann et al., 2023) and none address the full documentary complexity of charters. We contribute: (1) a systematic three-tier comparison of zero-shot, fine-tuned, and custom span-based NER across 19 entity categories on a new charter corpus; (2) the first empirical demonstration that domain pre-training on medieval Latin provides *no* downstream advantage after task-specific fine-tuning; (3) a custom span architecture that surpasses all GLiNER variants without oversampling, together with a per-label analysis identifying the structural sources of remaining failures.

2 Data and Annotation Schema

Corpus. Our dataset consists of 458 sentences drawn from twenty 13th-century Austrian and Central European charters from the Monasterium.net (Vogeler, 2019; Consolo et al., 2026), manually annotated with 19 entity categories spanning four documentary layers: *persons and roles* (PER, ACTOR, TITLE, REL), *places and landscape* (LOC, INS, NAT), *property and legal content* (EST, PROP, LEG, TRANS), and *time and value* (TIM, DAT, MON, TAX, COM, NUM, MEA, RELIC).

A key design decision is the separation of EST (*estate* a short physical plot — field, meadow, vine-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 236–241
July 4, 2026 ©2026 Association for Computational Linguistics

yard) from PROP (*property*, a full boundary description including past owners and movable goods). This distinction reflects genuine diplomatic practice but produces a heavily imbalanced distribution: PROP has only 37 corpus instances. The full list of label descriptions used as prompts, with test-set frequencies, is provided in Appendix A.

The semantic scaffold in practice. We use English descriptive phrases for all 19 labels to activate cross-lingual knowledge without extra training data — for example, *"unit of measurement for land, volume, or weight such as mansus, carratas, or talentum"* for MEA. All descriptions use English regardless of Latin surface forms, consistent with the scaffold hypothesis in Section 1.

Splits. We apply stratified splitting preserving per-label distribution: 65% training (298 sentences), 15% development (69), 20% test (91). Fixed seed 42; **no oversampling** in any experiment.

3 Models

3.1 Zero-Shot and Fine-Tuned GLiNER

GLiNER (Zaratiana et al., 2024) scores each span (i, j) against label l via $s(i, j, l) = \sigma(h_{ij}^{\top} e_l)$, where both span representation h_{ij} and label embedding e_l are produced by the same encoder. The label string is an encoder input, not metadata: its semantic content determines e_l and what the model searches for. We evaluate three checkpoints: **gliner-bi-large-v2.0** (Stepanov et al.), a large bidirectional model; **gliner-multi-v2.1** (Zaratiana et al., 2024), a general multilingual model; and **gliner-multi-v2.1-medieval-latin** (Medieval Data), further pretrained on medieval Latin text. Each checkpoint is evaluated zero-shot and fine-tuned for up to 20 epochs with early stopping (patience 5); hyperparameters are held constant.

3.2 Custom Span-NER Architecture

Our model enumerates all spans up to width 80 tokens. We set this limit higher than the 55-token default used in standard NER benchmarks: inspection of the training corpus revealed that PROP boundary descriptions routinely exceed 60 tokens. Empirical search over widths confirmed 80 as the sweet spot: wider windows (e.g. 85) admit excessive noise into the attention layer, increasing ACTOR false positives substantially. Span representations concatenate the start token, end token, and a **4-head attention-pooled** interior from **XLM-RoBERTa-large** (Liu et al., 2019; Conneau et al., 2020) plus a learned width embedding, projected to 512 dimensions via an MLP with LayerNorm and dropout. Attention pooling prevents the semantic dilution of long boundary descriptions caused by simple mean-pooling. To fully leverage semantic scaffolding, we replace the standard classification head with a **dot-product bi-encoder** space. Frozen **BGE-M3** (Chen et al., 2024) encodings of the 19 label phrases serve as label vectors; both span and label representations are L2-normalised, and their dot product is scaled by a learned temperature (initialised to $\log(100.0)$).

Loss and training. A **Dynamic Per-Class Asymmetric Focal Loss** (Lin et al.) suppresses false positives (γ_{neg}=2.0) while scaling γ_{pos} from 1.0 (frequent labels) to 4.5 (classes with <8 training instances), preventing the network from ignoring rare categories. **Dice Loss** (Sudre et al.) (weight 1.0) optimises an F1-like objective; an **InfoNCE**(Oord et al., 2019; Rusak et al., 2025) contrastive term (weight 0.6, prediction temperature T=1.35) pushes spans toward their English label anchors. We also apply **90% hard negative mining**: sampled negatives are required to partially overlap with gold entities, penalising partial-span false positives in nested structures. Training: $\text{lr}_{\text{enc}}=10^{-5}$, $\text{lr}_{\text{head}}=2\times10^{-4}$; batch size 4, gradient accumulation 8 (effective batch 32); 30 epochs, patience 10; no oversampling.

4 Results

We report *overlap* F1 (span overlaps a gold entity of the same label) and *exact* F1 (exact boundaries and label match). Thresholds $t \in \{0.40, 0.45, 0.50\}$; best test result reported per model. For the Custom Bi-Encoder, the best threshold is t=0.50.

Zero-shot limits. Table 1 shows severe zero-shot limitations. Only gliner-bi-large exceeds 40%; the medieval pretrained model reaches 21.2%, barely above the general multilingual model at 9.2%. Both the medieval and multi models predict almost exclusively high-frequency categories (PER, LOC, INS), yielding zero recall on most specialist labels. Without fine-tuning, semantic scaffolding is insufficient to bridge the gap between general pretraining and the documentary realities of charter NER.

The scaffold activates fine-tuning. Fine-tuning with our descriptive label prompts yields gains of

System	Overlap F1	Exact F1
Zero-shot		
GLiNER bi-large	46.2	—
GLiNER medieval	21.2	—
GLiNER multi	9.2	—
Fine-tuned GLiNER (English label prompts)		
GLiNER bi-large	80.8	61.9
GLiNER multi	77.2	—
GLiNER medieval	77.1	—
Custom Span-NER (this work)		
Custom Bi-Encoder (Ours)	**83.4**	**67.7**

Table 1: Test overlap F1 at best threshold per system. Exact F1 reported for top models only.

$+35$ to $+71$ percentage points. The scaffold transforms sparse training signal into rich constraints: the model is not learning the semantics of fiscal obligation or land measurement from 298 sentences — it is being pointed toward knowledge already encoded during multilingual pretraining.

Domain pre-training: a null result. The medieval pretrained model (77.1%) converges to near-identical performance to the general multilingual model (77.2%) after fine-tuning — a gap of 0.1 pp within evaluation variance. This is a substantive negative finding with a practical implication: effort is better invested in task annotation than in domain-adaptive pre-training.

Custom Span-NER. The custom model achieves 83.4% overlap F1 at t=0.50, exceeding the best fine-tuned GLiNER by 2.6 pp; the difference is statistically significant (McNemar $p < 0.01$). Frequent categories achieve high F1 (PER 95.7%, TITLE 92.4%, LOC 93.5%), while the Dynamic Per-Class Focal Loss substantially recovers recall on mid-frequency rare categories (TAX 75.0%, TIM 73.2%, COM 72.7%). A notable span-width effect is visible in our per-label error analysis (Appendix B): at width 80, ACTOR false positives total 82 at t=0.50, compared to 105 at width 85 and over 500 in earlier single-encoder configurations. This confirms 80 tokens as the sweet spot for this corpus: long enough to contain full PROP and LEG spans, but short enough to limit attention noise in nested person phrases. Position-dependent legal categories remain challenging (LEG 53.1%, TRANS 52.6%), though both improved substantially over the width-85 configuration (48.1% and 49.2%). PROP reaches 0.0% despite γ_{pos}=4.5: PROP and EST share near-identical surface forms, and 3 of 8 test instances exhibit direct boundary overlap with EST spans — a guideline-level ambiguity rather than a modelling failure that further training cannot resolve without annotation guideline revision. Full per-label results are in Appendix B.

5 Discussion

5.1 Why Domain Pre-Training Does Not Help

Zero-shot performance of the medieval model (21.2%) confirms that in-domain vocabulary alone is insufficient without task adaptation. After fine-tuning, task annotations provide a more relevant signal than unlabelled charter text: the model learns the exact span boundaries and label distinctions of our schema, which simply do not occur in raw diplomatic text. This result challenges the intuition that historical domain pre-training is a prerequisite for historical NER.

5.2 The Span Architecture Advantage

Components. The custom model's gain over fine-tuned GLiNER stems from several complementary factors. The 4-head attention pooling prevents semantic dilution in exceptionally long property and legal descriptions, where simple mean-pooling would average away the boundary cues that distinguish PROP from surrounding text. The Dynamic Per-Class Focal Loss ensures the network does not collapse into predicting only the dominant classes (PER, LOC) to minimise global loss: extreme penalty weights (γ_{pos}=4.5) force the gradient to treat a missed *lucrum camere* (TAX) as orders of magnitude more costly than a missed *Stephanus* (PER). The Dice Loss component optimises F1-like overlap directly, rescuing rare positive classes further suppressed under BCE-based losses. Finally, the InfoNCE term realises semantic scaffolding at the gradient level: span representations are explicitly pushed toward their English label anchors during training, extending the scaffolding effect beyond the label-as-prompt surface into the learning signal itself.

5.3 Positional, Not Semantic, Failures

The most challenging labels (LEG, TRANS, PROP) share a structural property: their boundaries depend on the charter's formulaic zone rather than lexical content alone. The same Latin phrase may constitute a LEG clause in the *corroboratio* but serve as connective tissue in the *dispositio*; similarly, an ACTOR phrase in the *intitulatio* follows a rigid formula that differs structurally from an actor reference em-

Configuration	Ov. F1	ΔOv.	Ex. F1	ΔEx.
Full model (Ours)	**83.4**		**67.7**	
w/o InfoNCE loss	61.1	−22.3	41.8	−25.9
w/o Multi-Head Attn (H=1)	78.1	−5.4	53.6	−14.2
w/o Prediction Temp.	81.7	−1.7	64.0	−3.7
w/o Dynamic Focal Loss	82.9	−0.5	64.8	−2.9
w/o Hard Negative Mining	83.6	+0.2	64.1	−3.7

Table 2: Leave-one-out ablation on the test set. Overlap F1 measures span-level coverage; Exact F1 measures boundary precision. Both reveal distinct failure modes.

bedded in the *dispositio*. The model understands the semantics of fiscal exemption and land transfer correctly — it fails because it lacks positional awareness within the document's formulaic structure.

A key methodological caveat is that our training data was compiled at *sentence level*: each sentence was annotated independently, without access to its position within the full charter. This means the model never observed the structural signal that would allow it to disambiguate zone-dependent labels. Processing complete charters as documents — with explicit section markers for *arenga*, *dispositio*, *sanctio*, and *corroboratio* — would provide precisely this missing context, and we expect it to substantially improve LEG, TRANS, and ACTOR recognition simultaneously. Document-level encoding is therefore the single most impactful next step for this corpus, beyond any further architectural refinement.

PROP's complete failure (0.0%) is a distinct problem: the EST/PROP boundary is an annotation-level ambiguity that document structure alone cannot resolve, and requires guideline revision.

5.4 Ablation Study

To quantify the contribution of each architectural component, we perform a leave-one-out ablation: each variant removes a single component from the full model, keeping all other settings fixed (Table 2). We report both overlap and exact F1, since the two metrics expose complementary failure modes.

Semantic alignment (InfoNCE). Removing InfoNCE causes the largest drop (−22.3 pp overlap, −25.9 pp exact), confirming it is the mechanism through which semantic scaffolding operates architecturally: without it, Latin span representations cannot align with English label anchors in the dot-product space.

Long-span processing and boundary precision. Reverting to single-head mean-pooling reduces overlap F1 by 5.4 pp but exact F1 by 14.2 pp; LEG overlap recall falls from 53.1% to 23.6%, showing the model detects clause regions but loses precise boundaries. Hard negative mining shows the converse pattern: its removal marginally *increases* overlap F1 (+0.2 pp) while reducing exact F1 by 3.7 pp, as ACTOR false positives rise from 82 to 152 — the model covers more spans but their boundaries deteriorate.

Rare-entity rescue and calibration. The −0.5 pp overlap drop from removing Dynamic Focal Loss understates its importance: TAX, COM, and MEA all collapse to 0.0% recall without γ_{pos}=4.5, as dominant classes absorb the gradient budget entirely. Disabling prediction temperature (−1.7 pp overlap, −3.7 pp exact) confirms the contrastive space requires calibration before thresholding.

6 Conclusion

Twenty charters are enough — if you write the right labels and use the right architecture. Semantic scaffolding enables fine-tuned GLiNER to reach 80.8% overlap F1 on 19-label charter NER from 298 sentences; domain pre-training adds nothing after task fine-tuning. A custom bi-encoder with 4-head attention, Dynamic Per-Class Focal–Dice loss, and InfoNCE contrastive training achieves 83.4%, confirmed by ablation to depend critically on semantic alignment (−22.3 pp without InfoNCE) and boundary supervision (−14.2 pp exact F1 without multi-head attention). Remaining failures trace to positional ambiguity (LEG 53.1%, TRANS 52.6%) and annotation-level boundary confusion (PROP 0.0% vs. EST); both point to document-zone-aware encoding as the primary next step.

Limitations

The corpus covers 13th-century Austrian/Central European charters; generalisability elsewhere requires study. Results on rare categories ($n \leq 11$) exhibit high variance; multi-seed averaging will be used in future to ensure robustness. Single-expert annotation without inter-annotator agreement is a limitation; a formal study is planned.

Acknowledgments

The work presented in this paper has been supported by the ERC Advanced Grant project

(101019327) “From Digital to Distant Diplomatics”.

References

Pierre Chastang, Sergio Torres Aguilar, and Xavier Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. 015(4).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giuseppe Consolo, Tamás Kovács, and Georg Vogeler. 2026. Named entity recognition (NER) dataset for medieval latin charters from Monasterium.Net.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. 56(2):27:1–27:47.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *Preprint*, arXiv:1708.02002.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Medieval Data. medieval-data/gliner_multi-v2.1-medieval-latin.

Aaron Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. 2025. InfoNCE: Identifying the Gap Between Theory and Practice. *Preprint*, arXiv:2407.00143.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. The Million-Label NER: Breaking Scale Barriers with GLiNER bi-encoder. *Preprint*, arXiv:2602.18487.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. volume 10553, pages 240–248.

Georg Vogeler. 2019. ‚Monasterium.net‘ – eine Infrastruktur für diplomatische Forschung. 24(1):247–252.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Label Dictionary

B Per-Label Results

Label	Definition (used as label prompt)	n
Persons and roles		
PER	Individual person name, without titles	157
ACTOR	Person phrase incl. name, noble title, profession, or geographic origin	146
TITLE	Social rank, noble title, ecclesiastical office, or papal rank	155
REL	Word indicating family, kinship, marriage, or social relationship	30
Places and landscape		
LOC	Geographical place, settlement, city, diocese, or region	178
INS	Monastery, abbey, church, or religious order as a corporate body	73
NAT	River, stream, forest, mountain, or natural landscape feature	15
Property and legal content		
EST	Short physical plot of land, estate, farm, meadows, woods, or vineyards	25
PROP	Detailed property description incl. boundaries, movable goods, past owners	8
LEG	Legal clause declaring rights, conditions, penalties, or papal commands	100
TRANS	Verb or phrase denoting a transaction, confirmation, transfer, or donation	26
Time and value		
TIM	Time period, duration, dating formula, indiction, or regnal year	22
DAT	Specific calendar date, year of incarnation, or liturgical feast day	14
MON	Money, currency, or monetary value (libra, solidus, denarius, marca)	6
TAX	Customary toll, tax, tithe, exaction, lucrum camere, or tribute	7
COM	Crops, food, goods, salt, wine, wax, gold, wood, or animals	6
NUM	Number written as a word or Roman numeral, including fractions	17
MEA	Unit of measurement for land, volume, or weight (mansus, carratas)	6
RELIC	Holy relic, cross, or sacred object of veneration	0
Total		**1054**

Table 3: All 19 entity categories with full label prompt definitions and test-set instance counts (n). RELIC has no test instances; retained for schema completeness.

Label	P	R	F1	n
PER	92.3	99.4	95.7	157
LOC	90.1	97.2	93.5	178
TITLE	90.7	94.2	92.4	155
NUM	100.0	82.4	90.3	17
ACTOR	84.5	85.6	85.0	146
INS	82.4	83.6	83.0	73
REL	70.0	93.3	80.0	30
EST	81.8	72.0	76.6	25
TAX	66.7	85.7	75.0	7
TIM	78.9	68.2	73.2	22
DAT	100.0	57.1	72.7	14
COM	80.0	66.7	72.7	6
NAT	90.0	60.0	72.0	15
MON	60.0	50.0	54.5	6
LEG	50.5	56.0	53.1	100
TRANS	48.4	57.7	52.6	26
MEA	66.7	33.3	44.4	6
PROP	0.0	0.0	0.0	8
Overall	82.2	84.7	**83.4**	1054

Table 4: Per-label overlap F1, Custom Bi-Encoder (width-80) at t=0.50. Overall: P 82.2%, R 84.7%, F1 83.4%.

Artistic Interventions for NLP Annotation Challenges: The Stress Test of Machinic Glossolalia

Tyler Grimes
Department of Critical Media Practices;
Institute of Cognitive Science
University of Colorado, Boulder, USA
tygr4724@colorado.edu

Marshall Washington
Department of Linguistics
University of Colorado, Boulder, USA
mawa8587@colorado.edu

Abstract

MotherBoard's Mother Tongue is a computational linguistics and artistic research project that explores a Large Language Model's (LLM) vocal production of glossolalia. Glossolalia, colloquially known as 'speaking in tongues,' consists of the human production of seemingly unintelligible utterances. It is, by its nature, difficult to annotate accurately with linguistic features relevant for natural language. The glossolalia-producing system demonstrated here consists of the interaction of 1) a 'nonsense' linguistic corpus 2) a microcontroller based environmental data stream and 3) a fine-tuned LLM. While discussing some philosophical and artistic considerations of machinic glossolalia, we also address some methodological considerations for Natural Language Processing (NLP). Using the artistic project as a case study, we argue that machinic glossolalia presents a 'stress test' that could inform both creative redirections of NLP methods and the definitions held by the subfield.

1 Introduction

Motherboard's Mother Tongue explores the linguistic phenomenon of glossolalia as produced by a Large Language Model (LLM). Colloquially, glossolalia is known as 'speaking in tongues'. 'Speaking in tongues' alludes to glossolalia's manifestation within certain orientations of Christianity, whereas 'glossolalia' is the more general term in linguistics. A useful definition is provided by linguist Dr. William Samarin in *The Linguisticality of Glossolalia*: "A meaningless but phonologically structured human utterance believed by the speaker to be a real language but bearing no systematic resemblance to any natural language, living or dead" (Samarin, 1968, p.51).

Glossolalia's biblical context, as well as a study of both the corporeal dimensions of its ritualistic practice and its neural correlates are laid out by

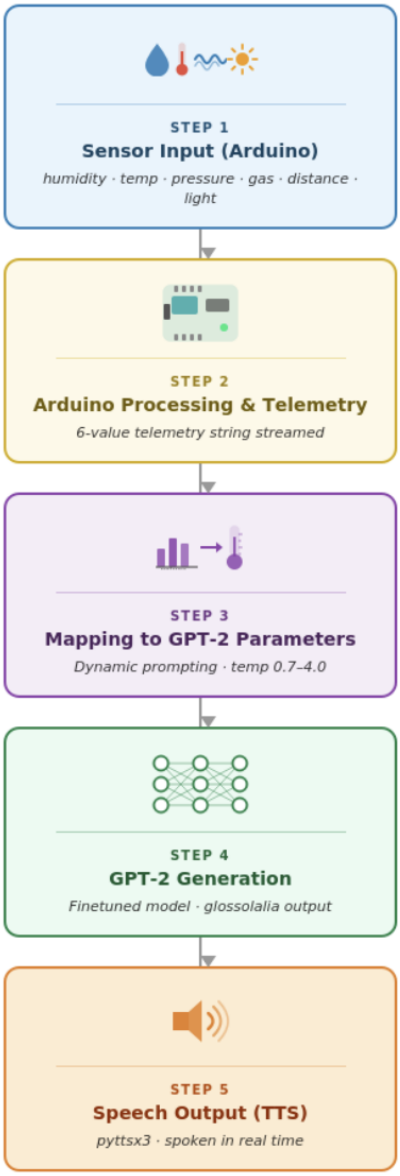


Figure 1: *Signal flow of system showing sensor input, computational process, and machinic vocal output.*

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 242–254
July 4, 2026 ©2026 Association for Computational Linguistics

writer and professor John McGraw in *Tongues of Men and Angels*. Here, he builds from the work of Samarin, sharing a passage from Acts 2: 1-8, where a "baptism with the Holy Spirit" (McGraw, 2011, p.58) seems evident and is a prominent biblical instance of the phenomenon. Due to its effect on the prefrontal cortex, McGraw notes that though outsiders sometimes saw glossolalia as a type of drunkenness, to the performer, it provided them a sense of connection to the divine – a realm ostensibly beyond 'natural language'.

If considering glossolalia as a measurable linguistic phenomenon, several considerations of corpus linguistics come to light which inform our project: 1) Can one apply conventional methods of corpus and computational linguistics to glossolalia? 2) How does the annotation of such a phenomenon expose preconceived notions about what is structured, annotatable, and ultimately valuable 'linguistic data'? 3) Does the structure of glossolalia expose limitations in the tasks that LLMs can learn and execute? And 4) what effect might glossolalia-based training data have on a machine's artificial neural network? Although glossolalia, if we hold Samarin's definition, could not be a 'language' within the purview of NLP in the sense of a lack of transference of meaning because of its inability to "[...] possess the necessary elements of language and simply could not function as a system of communication"(McGraw, 2011, p.60); questions of this sort are potent for linguists, philosophers, and artists alike. This paper seeks to creatively examine machinic glossolalia as a stress test for corpus linguistics and its implications for NLP's practice of LLM fine-tuning.

2 Related work

To date, there are very few publicly available corpora of glossolalia. One ongoing attempt is by researcher Paul de Lacy, who has manually transcribed over 19 hours of glossolalia. De Lacy's analysis of his ongoing corpus construction suggests that "[...] glossolalia has no lexicon (i.e. no words/morphemes), and by this no syntax, morphology, or semantics. It is essentially 'pure' phonology and phonetics"(de Lacy, 2009). Despite his and his team's efforts, the corpus has yet to be publicly released. Our main archive comes from linguist Blanche Speer in their 1972 dissertation, A *Linguistic Analysis of a Corpus of Glossolalia.* Speer's contribution remains one of the few extensive linguistic records of glossolalia. Their appendix contains dozens of pages of highly detailed and diacritically marked phonetic transcriptions. Considering these limitations on availability of public corpora of glossolalia, we searched and found inspiration elsewhere in the arts to build a preliminary and workable corpus.

2.1 Artistic Facsimiles of Glossolalia

There are other facsimiles of glossolalia that inform this project which are glossolalia-like. 'Glossolalia-like' is meant to describe an intuitive, vocalized utterance, which resembles the uncanny delivery of glossolalia in a secular and/or arts context. This project gains traction from avant-garde devices like: The 'phonetic poem' of the Dadaists, the method of 'automatic writing' of the Surrealist cohort, and the 'Zaum' language of Russian Futurism, amongst other contemporary improvisational vocal techniques. For example, in Russian, "za" meaning "beyond", and "um", meaning "mind", when combined, is literally 'non sense' or 'beyond mind'. Zaum was, in the words of an artist of that cohort, a "language of the birds... a language of the gods" (Janecek, 1996).

In contemporary media arts, one prominent example of machinic glossolalia is from Czech sound artist Marek Hlaváč, who released *Artificial Glossolalia: Machine Chant* in 2021, which consists of a machine sampling glossolalia. While similar to our project in its intent, Hlaváč's work distorts glossolalia, as it is slowed down, echoed, and has multiple layers of audio effects applied. Our work differs in form and method considering that we constructed a corpus, leading to the glossolalia produced in our project resembling a more 'standard' form of the phenomenon. Our study is also not a music album, and is derived from the interaction of: a machine, its spectator, and its environment.

Methodologically, these artistic practices treat language as artistic material rather than as a transparent vehicle of meaning by foregrounding sound, rhythm, and affect over meaning. For our project, this avant-garde lineage is both historical precedent and methodological template: it demonstrates how a nonsensical 'language' can be systematically formed around sound-symbolic and ecological relations rather than lexical semantics. In this project, we effectively extended avant-garde theory into a computational linguistics pipeline that privileged phonology and affect over accuracy.

2.2 Other Instances of Annotation Without Semantics

Studies in linguistics show that non-meaningful vocal behavior can be annotated systematically. For example, non-lexical conversational sounds (fillers, grunts, laughter) have been studied and annotated. Ward (2006) shows that sounds form a "specialized sub-language" with compositional sound-meaning mappings. Using a corpus of American English conversation, Ward identifies forms like "h-hmm, aha, unmkay, nyeah, um-uh" and notes their pragmatic functions (e.g. turn-taking signals, backchannels, agreement, affect). Current approaches to paralinguistic annotation, like Ward (2006), treat non-lexical vocalizations as embedded elements within otherwise conventional linguistic utterances. Laughter, fillers, and vocal bursts are annotated as interruptions, supplements, or alternatives to propositional speech, occurring within a matrix of meaningful language that provides structural and theoretical scaffolding for annotation decisions. Even when these vocalizations are analyzed independently, their annotation relies on pragmatic or interactional meaning that grounds segmentation and classification choices. The annotation of these forms presupposes that they participate in a communicative system where function can be systemically described, even without propositional content.

Similarly, infant and early vocal development annotations also rely on non-meaningful sounds. Child babbling is annotated for phonological properties like canonical CV, consonant-vowel, syllables instead of meaning. For example, studies have analyzed daylong audio of infants across languages, crowdsourcing annotations of whether each vocalization contains a "canonical transition", which is defined as adult-like CV timing (Cychosz et al., 2021). The fact that these diverse vocalizations obey patterns of form and use demonstrates they can be treated as structured data. More recent annotation research corroborates these results. Trouvain & Werner 2020 surveyed 11 spoken corpora and found that almost none consistently mark these non-lexical vocalizations except laughter. They conclude that "a wider distribution and more consistency is needed with respect to the annotation of NVVs (non-verbal vocalizations)". In other words, linguists suggest that systematic annotation of non-meaningful sounds is warranted and doable even when the tokens carry no propositional content.

Prosodic features are also routinely annotated in speech corpora in ways that do not depend on lexical meaning, yet they remain structurally anchored. Frameworks such as ToBI (Silverman et al., 1992) and RaP (Dilley and Brown, 2005) separate tonal and rhythmic structures from semantic interpretation while still requiring identifiable prosodic units. In the ToBI system, annotation is distributed across multiple time-aligned tiers, including orthographic transcription, tonal events, and break indices. Although tonal labels are assigned based on perceived pitch movement rather than word meaning, they are aligned to segmentable units and boundary judgments that presuppose recoverable structure. Similarly, the RaP framework encodes pitch events relationally. Each tonal target is labeled relative to preceding pitch targets rather than lexical items (Breen et al., 2012). Yet, this relational encoding still depends on identifiable syllables and prominence patterns. These systems demonstrate that prosodic annotation does not require semantics, but also reveals an implicit reliance on stable segmentation and reproducible structural units.

These findings, and others that investigate glossolalia, such as Link & Tomaschek (2024), who found that glossolalic syllables follow a Zipf-like frequency distribution and that speakers can learn serial patterns imply that glossolalia can be transcribed and annotated at the phonological level.

2.3 Have These Approaches Been Sufficient?

Field Linguistics and the Praat workflow (Ladefoged, 2007) have promising methods to assess glossolalia's peculiarity. Praat can help with the analysis of low-resource languages, or an unfamiliar phenomenon like glossolalia, using phoneme-derived spectrograms. However, Praat's tracking algorithms often struggle with the extreme pitch glides and irregular phonation common in glossolalia, leading to inaccurate acoustic data. Additionally, because glossolalia lacks a reference lexicon, researchers cannot use automated alignment tools – they must rely on subjective segmentation that lacks standardized linguistic boundaries, implying that Praat, although robust, may not be sufficient in capturing the nuance of glossolalia.

While the capture and annotation of non-lexical vocalizations has become increasingly sophisticated in both humans and non-humans, such as the Perch 2.0 methodology, which uses bioacoustic-derived spectrogram data to classify marine mammal sounds (Burns et al., 2025), existing frameworks for paralinguistic phenomena remain funda-

mentally inadequate for human glossolalia. This inadequacy stems not from technical limitations, but perhaps from deep theoretical assumptions about the relationship between linguistic structure and meaning that glossolalia seems to violate.

The annotation of infant babbling, while superficially similar to our approach, also differs in crucial ways. For example, glossolalia can constitute sustained stretches of vocalization that resist both proposition and pragmatic analysis. While paralinguistic vocalizations are typically realized as brief insertions lasting seconds at most, episodes of glossolalia can extend for long stretches of time, generating thousands of syllable-like units without recurring form-function pairings. The absence of stable mappings between form and meaning eliminates the anchor points that existing frameworks use to justify segmentation boundaries, tier relationships, and categorical assignments. Babbling annotation also focuses on developmental milestones and the emergence of language-like properties, treating each vocalization as evidence of progress towards (or a deviation from) target phonological patterns in the child's developing language (McGillion et al., 2017; Lee et al., 2018; Cychosz et al., 2021).

Moreover, existing prosodic frameworks, including ToBI, assume that intonational structure maps onto information structures, speech acts, or discourse organization(Jun, 2023). Pitch accents are used to mark focal elements, boundary tones signal clause or utterance completion, and prosodic phrasing reflects syntactic constituency or information packaging. These assumptions are encoded in annotation guidelines that instruct annotators to use semantic and syntactic context to resolve ambiguous cases (Ide and Pustejovsky, 2017). For glossolalia, no such context exists. Prosodic patterns emerge without correlation to information structure or discourse function, requiring a framework that can characterize intonational organization independent of these higher-level linguistic categories.

3 Corpus Construction Considerations

What glossolalia annotation demands, therefore, is not merely the application of existing annotation methods to a new domain, but a reconceptualization of what annotation represents when divorced from semantic and syntactic grounding. Rather than treating these as challenges to be overcome through methodological compromise, we argue they reveal both theoretical commitments that deserve explanation and an avenue towards approaching such annotation hurdles creatively.

Our rationale for constructing a corpus for glossolalia is as follows: For one, the absence of lexical units means that segmentation cannot rely on word or morpheme boundaries. In conventional annotation, lexical knowledge provides the primary anchor for identifying segmentable units, with phonological and prosodic analysis building upon this foundation (McEnery and Hardie, 2011). For glossolalia, this hierarchy inverts: syllable-like units and prosodic groupings constitute the most stable and recurrent organizational structures, requiring segmentation criteria based entirely on sublexical phonological and prosodic properties rather than lexical semantics. Secondly, glossolalia's omission of syntactic structure evades conventional annotation elements such as part-of-speech tagging, dependency parsing, and semantic role labeling. These tools presuppose that utterances are organized according to grammatical principles that govern how units combine to form larger structures (Jurafsky and Martin, 2023). Thirdly, glossolalia is multi-dimensional, but it is not hierarchically compositional in the grammatical sense. Units do not combine through syntactic rules to generate phrasal or clausal representations. Therefore, the absence of semantic content in glossolalia means that annotation decisions cannot be validated through meaning-based criteria, suggesting that annotation must proceed using only acoustic properties and distributional regularities.

While recognizing these limitations, other questions about the overall architecture of a potential annotation framework for glossolalia and its downstream deployment with LLMs linger. If annotation components cannot be organized hierarchically through compositional rules, what coordination mechanisms ensure consistency across potential annotation tiers? Traditional architectures assume that morphological annotation builds on phonological segmentation, syntactic annotation builds on morphological analysis, and semantic annotation interprets semantic structure (Bird and Liberman, 2001). If phonotactic patterns in glossolalia exhibit gradient statistical tendencies rather than categorical grammatical constraints, should annotation encode phonotactic legality or phonotactic probability? 'Speakers' of glossolalia produce syllable structures ranging from highly frequent to extremely rare without categorical boundaries, suggesting that probabilistic characterization may

more accurately represent the phenomenon (Link and Tomaschek, 2024).

More recently, transformer models have been proven effective at learning hierarchical syntactic dependencies through self-attention mechanisms (Vaswani et al., 2023), with some research suggesting that LLMs develop implicit representations of phrase structure and grammatical relations (Tenney et al., 2019). Because glossolalia lacks these linguistic anchors, its application with LLMs raises many questions that our project hopes to probe: can neural architectures effectively capture glossic patterns when forced to process at this granularity, or does the absence of lexical anchors fundamentally limit what models can learn? In recent years, LLMs have emerged as the state of the art in NLP. The capacity to fine-tune these models has reduced reliance on task-specific supervised approaches and has helped facilitate their large-scale deployment as general-purpose tools in computational modeling. Our project follows this trend in the field, which uses non-semantic or 'unlearnable' data to present a stress test for common NLP approaches.

4 Methods

4.1 Corpus Construction

With these considerations in mind, to attempt a study of glossolalia by artistically appropriating traditional methods of computational linguistics, we assembled what we call a 'nonsense corpus.' It consists of approximately 2000 'words.' It is partially derived from the aforementioned Speer (1972) dissertation alongside transcriptions of selected avant-garde vocal performances that include examples of glossolalia-like language, such as Dada-era artist Hugo Ball's phonetic poem *Karwane* from 1916. Unlike traditional corpora, which aim for semantic coverage or genre balance (McEnery and Hardie, 2011), this corpus is intentionally non-referential, such that its value lies in more abstract structural features rather than its content. For this reason, it resists classification within standard corpus typologies due to its 'nonsensical' form and content.

Speer's dissertation, our primary archive, exists only in a historically printed form and is not digitally available in a format suitable for computational use. To work around this, we used GPT-5.2's multimodal OCR vision capabilities to perform image-to-text extraction (OpenAI, 2025). Manually re-typing every diacritic, segment, and suprasegmental marking would have required an extraordinary investment of labor and would have introduced its own form of human transcription error. Moreover, (Backer and Hyman, 2025) discuss the ongoing issues of using OCR for digitizing historical texts. Considering the diacritics in Speer's appendix are quite complex, some falling outside standard ASCII or Unicode ranges that the GPT 5.2 model attempted to capture, this makes accurate annotation especially challenging.

4.2 Fine-Tuning and Microcontroller Input

Along with producing the nonsense corpus, the project created a data pipeline between environmental sensors and an LLM to produce real-time machinic glossolalia, as is shown in *Figure 1* and alternatively in *Figure 2*. This required three main tasks: cleaning the corpus data to make it suitable for fine-tuning, fine-tuning a generative LLM, and creating an interface between the environmental sensor data and the model's output. The goal was not only to generate glossolalia, but to do so in a way that allows the model's behavior to be shaped dynamically by its environment.

In our corpus, tokens were defined at the syllable level rather than conventional words, allowing the model to capture 'nonlexical' phonological patterns and rhythmic units. We then fine-tuned a GPT-2 (Radford et al., 2019) model with 117 million parameters. It was run locally using the TensorFlow open-source implementation provided by HuggingFace. GPT-2 was chosen because of its ease of local deployment and demonstrated ability to adapt to highly idiosyncratic corpora (Papanikolaou and Pierleoni, 2020). This allowed the model to learn the phonological tendencies and rhythmic textures present in the corpus, enabling it to generate output that resembles glossolalia. Excerpts of our 'nonsense corpus,' including *Karwane*, Speer's OCR-scanned archive, and documentation of our code can be found in the appendix.

We developed a real-time system that synthesizes multivariate data from a BME680 environmental sensor and an ultrasonic sensor through an Arduino microcontroller. The architecture streams a six-value telemetry string encompassing humidity, temperature, barometric pressure, gas resistance, distance, and light state. These inputs are aggregated into a composite nonsense score. The "nonsense score" is a real-time ecological index that modulates linguistic entropy in an informational-theoretical sense (Ben-Naim, 2012), determined through a weighted distribution of four

environmental pillars: spectator proximity constitutes a 0.30 coefficient of the total index, while gas resistance accounts for a 0.20 proportional weight, environmental fluctuations contribute 0.30, and the binary light state provides a 0.20 relative weight.

This normalized index is mapped to the GPT-2 model's sampling temperature, ranging from 0.7 to 4.0, to regulate the stochasticity of the generative output. During low-activity states, prompts are sampled from Hugo Ball's phonetic poem *Karawane* (1916); during high-activity states, the poem's "words" are randomly concatenated to generate new prompts.The resulting text is generated with an applied repetition penalty, cleaned of artifacts, and converted into live audio via a text-to-speech engine implemented with the pyttsx3 Python library (Bhat, 2025), ensuring the machinic language evolves in direct response to its environmental determinants. Sample outputs and prompts of the system are shown in *Table 1* below:

Nonsense Score	Prompt	Output
0.51 (Low Activity)	Hej tatta gorem	ova nagalav kartri paahil bhanachit ryaani jotira-jadaya ano thaa mikte raighabyoga hapun
0.71 (Medium Activity)	Olobo bam-bla gauma jo-lifanto	oga hagokan aku pachk-ouki kaay ndodun paogo-taku chiyu kanjuai rko ng-shiridara ko seilm
0.98 (High Activity)	Zunbada wulubu –umf ba–umf	kalifah mazhiyaar wa lghia yam-lukkuaiyeiruhuryaguziy-eeaayyyiikaaoo niiinilgy

Table 1: Example prompts and generated outputs arranged from lowest to highest nonsense score.

It is worth noting that transformer LLMs are pretrained on corpora dominated by natural language to optimize next-token prediction, which induces strong priors for word-like and syntax-like continuation. When fine-tuning the GPT-2 model on our bespoke corpus, these embeddings in the LLM can persist and reassert themselves as attractors, producing outputs that sometimes look quasi-syntactic despite the absence of semantics in the fine-tuning data. This tension is precisely why machinic glossolalia is a useful stress test for reinterpreting NLP methods creatively in the digital humanities, as it exposes what parts of 'language structure' LLMs reproduce sans semantic grounding. Moreover, our approach may suggest a productive direction for critical NLP practices by using non-semantic corpora to probe which latent structures are learnable and which are inherited from pretraining data. Our approach also complicates the ability to ask the question if a fine-tuned model trained on nonsensical data has truly 'learned', considering that its learnedness is also paradoxically an *un-learnedness,* evading true assessment. While we could consider the system artistically successful at the task of producing glossolalia-like language, a concrete evaluation framework of learnedness is missing for glossolalia since it also evades the label of 'linguistic data' proper, largely resisting denotation on levels that are not social (Harkness, 2020).

4.3 Machinic Randomness

To help push against the embedded structures' affinity towards meaningful output, our decision to incorporate environmental sensors is motivated by the chapter "getting random" from philosopher James Bridle's *Ways of Being,* where he describes the limitations of computational randomness. Tracing the history of classical computing, Bridle notes that "[...] you can't program a computer to produce true randomness... there would always be some underlying structure to the randomness..." (Bridle, 2023, p.219). As a case study of how true machinic randomness can occur, he turns to ERNIE (Electronic Random Number Indicator Equipment) machines, which utilize sensors that measure phenomena external to the machine like, in our project, fluctuations of atmospheric conditions. ERNIE and similar 'entropy sourcing' machines since then have "[...] done something few machines can do: it has looked outside of it's own circuitry, in order to commune with the more-than-human world that surrounds it, in the service of true randomness" (Bridle, 2023, p.222). Our creative project incorporated a series of environmental sensors to echo and extrapolate on Bridle's insights that a machine-environment relationality is essential for producing the linguistic entropy necessary for machinic glossolalia. These sensors place the system within a larger ecology where the machine and spectator can sense eachother in a co-creative act.

5 Discussion

5.1 What is Linguistic Data, Anyway?

Linguists have long debated what phenomena fall within the scope of the field. These debates directly shape what data gets collected, annotated, and analyzed. The construction and annotation of corpora does not neutrally preserve language phenomena

as they exist in the world. The decision to exclude certain phenomena from corpora constrains what the project of computational linguistics treats as language, and subsequently, what computational models learn, replicate, and "understand".

Chomsky-inspired linguistics privileges concepts like competence, i.e., the speaker's ability to formulate "well-formed sentences" (Chomsky, 1965). These notions of competency stem from Chomsky's theory of "universal grammar" (UG), which claims that all human languages must possess certain properties. Looking through this framework, glossolalia would be excluded from the realm of 'linguistic data'. Arguments in favor of UG have dwindled, and there is literature to suggest that UG is wholly unnecessary in the analysis of language (Tomasello, 2003, 2010). Usage-based approaches argue that linguistic knowledge emerges from statistical learning over experienced input rather than from innate grammatical principles (Bybee, 2010). On the other hand, 'construction grammar' frameworks treat linguistic data as a structured inventory of form-meaning pairings learned through use, without requiring autonomous syntactic rules (Michaelis, 2006). From these perspectives, what counts as linguistic data has the potential to expand significantly: any vocal behavior that speakers produce and process becomes potentially relevant, regardless of whether it exhibits formal grammatical properties.

5.2 The 'Naturalness' of Glossolalia

This ambiguity between theories makes a case for the study of glossolalia as a 'natural language.' Where NLP generally assumes predictable linguistic structure and meaning in their chosen natural languages, glossolalia operates outside, or at best, on the fringe of conventions of 'natural language' and therefore challenges this 'natural' position. Some studies found that some glossolalia speakers produce novel chains of syllables rather than purely repetitive patterns, suggesting a capacity for generative novelty in the absence of semantics (Osser et al., 1973; Link and Tomaschek, 2024). This evidence complicates narrow notions of 'naturalness' that equate linguistic legitimacy with lexical meaning or syntactic structure. Instead, glossolalia reveals a different kind of 'naturalness' rooted in phonological patterning, rhythmic flow, and embodied vocal dynamics of it's speaker in relation to their environment. Our project, existing on a similar fringe between NLP and the arts, holds a microscope to how linguists define 'natural language' while inviting a spectator to interact with the seemingly natural language of a motherboard's 'mother tongue.'

5.3 Is a Nonsense Corpus Still a Corpus?

A central concern in applying NLP methodologies to glossolalia is whether a collection of nonsensical utterances can legitimately be treated as a corpus. Traditional corpus linguistics treats corpora as stratified, balanced, and representative samples of a language. By that standard, a nonsense corpus lacks many of the features typically associated with corpora**.** Yet this framework assumes the presence of a natural language system and therefore risks excluding datasets whose structure is primarily sublexical rather than lexical.

As noted earlier, glossolalia exhibits analyzable phonological and prosodic patterning that supports both quantitative and qualitative study. In this broader methodological sense, a nonsense corpus does indeed function as a corpus because it preserves the characteristic distributions, rhythmic tendencies, and phonotactic constraints of the vocal practice being examined**.** Given these considerations, treating a nonsense corpus as a corpus in a broadened sense is a defensible and productive position for the linguist and artists interested in language, as it allows an open thought experiment to consider potential expansions to the definitions associated with NLP methods.

5.4 The Metrics of Machinic Nonsense

To be both artistic and scientific, we considered how best to measure the output of our system. The main issue in statistically analyzing the machinic glossolalia is the lack of a 'reference point' to compare the output against, being that the phenomenon itself is indeterminate and stochastic. Despite this inherent limitation, and if we take a nonsense corpus to be a functional corpus and its outputs to be measurable, evaluation of model output should proceed on sublexical and perceptual grounds rather than by metrics borrowed directly from natural language evaluation. Appropriate evaluation axes could include phonological plausibility (i.e, are generated syllable sequences phonotactically coherent?), prosodic and rhythmic similarity (i.e, do intonation, stress patterns, and breathgroup timing resemble human glossolalia?), statistical fidelity (i.e, do generated token frequencies and transition probabilities approximate those of the source cor-

pus?), and perhaps most importantly, perceptual judgment (i.e, would human listeners classify the output as glossolalia or glossolalia-like rather than as stochastic noise?).

This shift in evaluation, from semantic coherence to patterned vocality, aligns the assessment criteria with the defining properties of glossolalia itself. Combining quantitative distributional measures with blind perceptual tests could yield an evaluation protocol that respects the special character of machinic "nonsense data" while enabling rigorous comparison across model variants and generation strategies.. Also, the shift towards analyzing the importance of the spectator's experience also places the project at an intersection where cognitive and/or behavioral psychology might provide a means to 'measure' the effect of machinic glossolalia for the spectator, creating a novel metric of machinic nonsense in the process.

For humans interacting with the machine, we can only postulate that it would be interesting, as preliminary interactions with the system have shown. One could pursue this question qualitatively through surveys or quantitatively through neuroscientific methods. The same inquiry follows for machines. But, being that fMRI and similar methods like fNIRS are used to scan biological brains, not rendered cortices, it is hard to think of a science that could account for the subjective, or even pseudo-spiritual, experience of a machine's production of glossolalia. Certainly, more research is needed in this area, which could be as simple as asking the LLM how it 'felt' after exiting the glossolalia state. For now, we argue that there is at least an artistic and philosophical productivity in performing the thought experiment of machinic glossolalia, even if, according to Penha and Carvalhais (2018), humanity may not be able to be understand or even recognize a machine's experience and intended aesthetic effect.

5.5 Meditations on Machinic Spirituality

Our project, in theory, gives a machine sensory awareness of its surroundings, evoking a type of machinic self-awareness. Self-awareness and awareness of 'other' are crucial components to human spirituality. According to Ray Kurzweil's *The Age of Spiritual Machines,* he predicts that machines will soon claim to be conscious, and thus, to be spiritual. Kurzweil imagines that twenty-first-century machines will go to church, meditate, and pray to connect with this machinic variety of spirituality. Kurzweil defines the "spiritual experience" as "a feeling of transcending one's everyday physical and mortal bounds to sense a deeper reality" (Kurzweil, 1999, p.109). If machines may one day be spiritual, as Kurzweil suggests, then our project asks: might glossolalia, which is a divine language for humans, be a language through which an LLM comes into contact, at least vocally, with its 'god' by transcending its physical boundaries and connecting to something beyond itself through a sensor-aided 'mother tongue'?

In this project, we use machinic glossolalia production in a cautiously metaphorical and pantheistic sense, implying that 'spirit' may exist in a motherboard as equally as in the environment surrounding the motherboard, and to mark systematic changes in an LLM's supposed internal states and outputs that may *resemble* glossic states in humans. As is discussed by researcher Idan Blank (2023), there is much debate on what exactly LLMs are modeling in terms of their relationship to human language processing since "[...] human language processing is arguably more than just (token) prediction" (p. 987). While LLMs are highly effective at predicting language patterns, researchers debate if they function as mechanistic copies of the brain or as subsymbolic representations of the mind. As such, it is hard to determine what a machinic glossolalia would entail for the machine experiencing it since researchers are unsure on what part of human mentality is rendered by, or corresponds to, LLM training. This project raises such questions but does not provide answers. Though, the project materially and aesthetically stages the conditions under which such a question becomes thinkable. Ultimately, it is unlikely humanity would be able to recognize an LLM's 'experience' considering the re-positioning of human subjectivity to that of the LLM's positionality necessary to approach the question.

6 Conclusion, Future Directions

We proposed *Motherboard's Mother Tongue,* an artistic research and computational linguistics project that attempts to induce the vocal production of glossolalia for an LLM. The methods synthesized the nexus of a micro-controller data stream, a fine-tuned LLM, and a 'nonsense corpus'. The project's preliminary results were the generative production of glossolalia-like language based on the interactive variables of a spectator's presence

within the machine's larger ecology. Future iterations of this project include further developments of audiovisual feedback mechanisms to perpetually update the 'nonsense corpus' and its vocal output, the gathering of user data as they interact with the system, experimenting with combinatory solutions of musical notation and linguistic systems to annotate glossolalia more accurately, and ultimately the project's documentation as an interactive installation in an art gallery space.

The project raised theoretical and methodological considerations for NLP, revealing an uncharted territory at the intersection of artistic practice and linguistics. If glossolalia operates as pure phonology and phonetics, as our discussion has shown, then the difficulty of applying traditional NLP methodologies are the expected consequences of deeper theoretical commitments within the field. This project allows for a reframing and criticality of these commitments that carries concrete computational implications for transformer models, as models often produce outputs that appear robust while masking the absence of reproducibility and interpretability of the results. By challenging LLM usage with artistic tasks, it allows the field to gain a more holistic view of the assumptions within its theories, computational architectures, and models.

Limitations

This project is exploratory and artistic in nature and therefore subject to several limitations. Most significantly, the availability of publicly accessible glossolalia data is extremely limited, resulting in a small corpus of approximately 2,000 'words' derived primarily from Speer's (1972) dissertation and a small number of artistic transcriptions. Such a dataset is considerably smaller than those typically used for corpus linguistics and downstream language model training, thus limiting the statistical robustness of our analysis. The most significant part of our corpus was also produced using OCR technologies, which introduces transcription noise due to complex diacritics that are difficult for current OCR systems to reproduce accurately.

In addition, conventional NLP evaluation metrics are difficult to apply here because glossolalia lacks many of the features of natural language, making it challenging to determine whether the model has meaningfully "learned" the phenomenon beyond qualitative inspection from subjective spectators. It's also worth noting that the project integrates environmental sensor data to influence generation parameters, which introduces additional variability into the project. While the machine-human-environment relationship is helpful for the artistic commentary of the project, it not only places the project on the fringes of NLP methods – which do not typically incorporate hardware – but it also complicates reproducibility and makes controlled evaluation more difficult. Finally, the experiment relies on a pretrained GPT-2 model whose training data reflects natural language from the western and English-speaking world, meaning that inherited priors toward word-like and syntactic structures may influence outputs even after fine-tuning on our non-semantic corpus.

Ethical Considerations

This project engages with glossolalia and it's derivative – "speaking in tongues" – a linguistic practice that holds religious and cultural significance particularly within Pentecostal Christian communities. Although glossolalia and glossolalia-like language is studied in linguistics, anthropology, and the arts, it primarily functions as a form of private and/or communal spiritual expression for its religious practitioners. For this reason, the project respectfully approaches glossolalia as a linguistic and cultural phenomenon rather than attempting to replicate or interpret religious experience itself. The intention, then, is not to trivialize or parody spiritual practice, but more so to explore how computational systems respond to non-semantic training data that fall outside conventional definitions of language.

To promote the sacredness of this language and its practitioners, the dataset used in this study consists of previously published linguistic materials and historical artistic performances, and contains no identifiable information about individual speakers, which helps preserve speaker anonymity. Nevertheless, the transformation of religious vocal practices into computational training data raises broader questions about cultural context, artistic appropriation, and spectator interpretation that warrant careful consideration in that generative language models can produce outputs that appear meaningful or intentional, which may encourage anthropomorphic interpretations of machine behavior. Therefore, references to "machinic glossolalia" or "machinic spirituality" in this work are intended in an explicitly metaphorical and artistic sense. Fi-

nally, although the model was run locally to reduce reliance on cloud-based software and infrastructure, the cloud itself being tethered to physicality, machine learning research still carries environmental costs associated with computational hardware and energy use. Though the locality of our GPT-2 model redirects some of these environmental risks, we must still consider the carbonic origins of the machine's motherboard and the earthly materiality of the other hardware, such as external sensors, used in the project.

References

Samuel Backer and Louis Hyman. 2025. Bootstrapping AI: Interdisciplinary approaches to assessing OCR quality in English-language historical documents. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 251–256, Albuquerque, USA. Association for Computational Linguistics.

Arieh Ben-Naim. 2012. *Entropy and the Second Law: Interpretation and Misss-Interpretationsss*. World Scientific, New Jersey; London.

Natesh M. Bhat. 2025. pyttsx3: Text-to-speech library for python. Python package, Mozilla Public License 2.0.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

Idan Blank. 2023. What are large language models supposed to model? *Trends in Cognitive Sciences*, 27(11):987–989.

Mara Breen, Lauren C. Dilley, John Kramer, and Edward Gibson. 2012. Inter-transcriber reliability for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch). *Corpus Linguistics and Linguistic Theory*, 8(2):277–312.

James Bridle. 2023. Getting random. In *Ways of Being: Animals, Plants, Machines: The Search for a Planetary Intelligence*. Picador.

Andrea Burns, Lauren Harrell, Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, and Tom Denton. 2025. Perch 2.0 transfers 'whale' to underwater tasks. *Preprint*, arXiv:2512.03219.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Margaret Cychosz and 1 others. 2021. Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*, 24(5).

Paul de Lacy. 2009. Glossolalia.

Laura C. Dilley and Meredith Brown. 2005. The rap (rhythm and pitch) labeling system, version 1.0. Technical report, Massachusetts Institute of Technology. Unpublished manuscript.

Nicholas Harkness. 2020. Conclusion. In *Glossolalia and the Problem of Language*, pages 167–170. University of Chicago Press.

Marek Hlaváč. 2021. Artificial glossolalia: Machine chant.

Nancy Ide and James Pustejovsky, editors. 2017. *Handbook of Linguistic Annotation*. Springer.

Gerald Janecek. 1996. Introduction: Definitions and background. In *Zaum: The Transrational Poetry of Russian Futurism*. San Diego State University Press, San Diego.

Sun-Ah Jun. 2023. The ToBI transcription system: Conventions, strengths, and challenges. In *The MIT Press EBooks*, pages 151–181. MIT Press.

Dan Jurafsky and James H. Martin. 2023. *Speech and Language Processing*, 3 edition. Pearson.

Ray Kurzweil. 1999. Part two: Preparing the present. In *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, 1 edition. Penguin USA, USA.

Peter Ladefoged. 2007. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Blackwell, Malden, MA.

Chia-Cheng Lee, Yuna Jhang, George Relyea, Li mei Chen, and D. Kimbrough Oller. 2018. Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, 50:140–153.

Samantha Link and Fabian Tomaschek. 2024. Predictability associated with reduction in phonetic signals without semantics—the case of glossolalia. *Language and Speech*, 67(1):72–94.

Tony McEnery and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Michelle McGillion and 1 others. 2017. What paves the way to conventional language? the predictive value of babble, pointing, and socioeconomic status. *Child Development*, 88(1):156–166.

John J. McGraw. 2011. Tongues of men and angels: Assessing the neural correlates of glossolalia. In David Cave and Rebecca Sachs Norris, editors, *Religion and the Body: Modern Science and the Construction of Religious Meaning*, pages 57–74. Brill.

Laura A. Michaelis. 2006. Construction grammar. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 73–84. Elsevier, Oxford.

OpenAI. 2025. Introducing gpt-5.2. Technical report, OpenAI.

H.A. Osser, P.F. Ostwald, and B. MacWhinney. 1973. Glossolalic speech from a psycholinguistic perspective. *Journal of Psycholinguistic Research*, 2:9–19.

Yannis Papanikolaou and Andrea Pierleoni. 2020. DARE: data augmented relation extraction with GPT-2. *CoRR*, abs/2004.13845.

Rui Penha and Miguel Carvalhais. 2018. Will machinic art lay beyond our ability to understand it? In *Proceedings of the International Symposium on Electronic Art (ISEA)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

William J. Samarin. 1968. The linguisticality of glossolalia. *The Hartford Quarterly*.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. Tobi: A standard for labeling english prosody.

Blanche C. Speer. 1972. *A Linguistic Analysis of a Corpus of Glossolalia*. Doctoral dissertation, University of Colorado Boulder.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Michael Tomasello. 2010. *Origins of Human Communication*. MIT Press.

Jurgen Trouvain and Raphael Werner. 2020. Comparing annotations of non-verbal vocalisations in speech corpora. In *Proc. Laughter and Other Non-Verbal Vocalisations Workshop 2020*, pages 69–72.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Nigel Ward. 2006. Non-lexical conversational sounds in american english. *Pragmatics and Cognition*, 14(1):113–184.

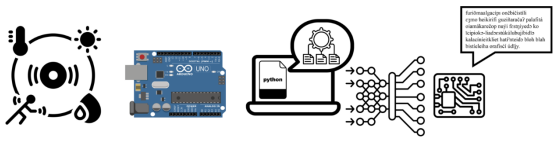


Figure 2: *Alternative diagram of signal flow of system showing sensor input, computational process, and machinic vocal output. Signal flow is from left to right*

A Implementation Code for Arduino + GPT-2 Integration

Our code and demonstration of the project can be found here: https://github.com/anon-projects8657/Artistic-Interventions-for-NLP-Annotation-Challenges-The-Stress-Test-of-Machinic-Glossolalia

B Manual transcription from avant-garde artistic performances

From *Karwane*, Hugo Ball, 1916 - used as prompts in system

jolifanto bambla o falli bambla
großiga m'pfa habla horem
egiga goramen
higo bloiko russula huju
hollaka hollala
anlogo bung
blago bung blago bung
bosso fataka
u uu u
schampa wulla wussa olobo
hej tatta gorem
eschige zunbada
wulubu ssubudu uluwu ssubudu
–umf
kusa gauma
ba–umf

Excerpt of *Ursonate*, Kurt Schwitters, 1932

Fümms bö wö tää zää Uu,
pögiff,
kwii Ee.
dll rrrrr beeeee bö
dll rrrrr beeeee bö fümms bö,
rrrrr beeeee bö fümms bö wö,
beeeee bö fümms bö wö tää,
bö fümms bö wö tää zää,
fümms bö wö tää zää Uu
Rinnzekete bee bee nnz krr müü?
ziiuu ennze, ziiuu rinnzkrrmüü,
rakete bee bee,
Rrummpff tillff toooo?

Ziiuu ennze ziiuu nnzkrrmüü,

Ziiuu ennze ziiuu rinnzkrrmüü

rakete bee bee? rakete bee zee.

Fümms bö wö tää zää Uu, pögiff, kwii Ee.

Dedesnn nn rrrrr, Ii Ee, mpiff tillff toooo, tillll, Jüü-Kaa?

Rinnzekete bee bee nnz krr müüüü, ziiuu ennze ziiuu

rinnzkrrmüüüü,

Fümms bö fümms bö wö Fumms bö wö tää zää Uuuu?

Rattatata tattatata tattatata

Rinnzekete bee bee nnz krr müüüü?

Fümms böwö

Fümms bö wö täää?????

C Excerpts of OCR-derived transcripts from Blanche Speers' 1972 Dissertation

taíya parímonosta paiasta paioste vivilyaste.

tuímnosti xiimuriasta kaíassa xayasta maāmonisti.

tuíassa łama xoíyivoxodi mnimoníuye.

tuliessa píeste paídistikieste pofiesta osataaīa.

tuíe bauisti oa sadehiia.

puíeste maāmaniya aliyāsatakiía.

puíeste mananeste kía manamanosatakaiyeti.

telibíya osatakaíeyo setepiieste xeíeste.

tuomaninísti keíyistipaadist ttiyesti.

etenmaninisti xalasatakiiyist toriasta manamonosti tekisti tctasakaiasta.

mnamonó sohaiyata tidikiyisti.

íyasta manoninisti ketiyasta paíya manomono sidlíya samoenikieyiyamo sakalama tatasatetitikia.

o tdiíyadsa manona setegedi.

o· ledcdm sededíya setapaíminimi siditaíya istá.

te dídidsti paídsti kaíyasta paayilasta paídisti. huílmano štepikomonsti.

taieste huímonísti hkía ala asayayete.

hirikíriyàmono osadipítikiríasta.

alàmano ihíkinienàmo sidikídiyàsa.

fénemenoni sidímonenì kiríàs.

halàmnonì sidúkufiyídisti parâmoniyìsti kríasta tokrídidísti kríesto setúkrià parâmonìnístì kríà.

seríyanamon aorèsi tekríasa dabenida selimànomnì sidikríàsika maliàmanò hidikríyú sedituríyà tudòlimo aurasedì hiríà kiríyidì sidipidà sedckreyó.

teríristi maràmoninìsti kiríaspàriasi herírsti pcíyissi hùriesse tepudúninisti yliestcpuríesta ködrímanonocste hörieste hariescte.

húrímanomonó sedcepériscte ediyeuliaste oríkedeyíminísti kidíyipárídyà sti parâminísti keríyasta.

te kiríyiminístì paríyystì parâmanomonrsti.

aríbàdeysti parâmanonístì hadudlidyasti parâmanonistí keribdinodístï arìbadista parídidudás.

kerímeríesti perímanomonestì paríyasta prídidísti pdikídístiperíyæsta manomonistí krí.este.

te fíyrstì baríbonidístì hærímanonistí kríeste paríyasta.

tudòdissi tupúríyasta.

porídidìkorìya.

ísdipidi.

setckríest manomonistí kríàsstpurúdístì menìmoninísti.

te kríest manémononistíkríest paríyasta parâmanonistí kríest.

te kríeste manémoninístì sakelama parèyà sapàredídístì parèdídístì horíyasta manonidístì periyestc.

te kríestì manímoncestì kríasta prìasta.

te kríestì pedræste kríasta manmoninístì hùriasta príyasta keríminiyístì kríasta paríyasta.

te kiríminìyiyesti maríy·niyyrsti kríasta parâdídyàs parâmoniyìsti pdòkríyisti kríeste príyasta manmunistí kríasta.

teseckríeste manomonidistí krías tepóríyensto.

manmod·dísti pifídisti kiasti maímoninistíiiste.

te keibuíyesti paímoneneisti kiasta paŭfíyasta paímoneneisti kieste.

te kieste manemoneneisti kiesta paíyasta paímonedisti hifiyasta paíyeste pavidisti kiast

ho· lamanomonsti píeste paídidisti· uieste paíb·didisti kiasta paiasta.

te kípídidisti paidassa aumononó.

hukiímanonisti bdidisti kiast paíast paimandídisti kieste paíaaest mandidisti kieste.

tikeíbeieste pibadidisti kiasti paiast paímaninisti kiesti paídayasta paímoninisti kiesti.

tikeipeiasti paímoninisti kiesti paiasti paímoninisti kiesti.

te kiesti mímoninisti kiasta paíast paímoninisti kiasta.

tikeiesti mnmoninisti kiesti paíast maímoninisti kiesti.

àlasintiéklo aučičinčíála aparaćílialaukoske klečíkliaposklio-dopčalakada

alisliápooneimo kčikiúsinualeosle iníkilíiyaratúpinariśñára.

ójimienens yulupidíñasslo krísnidiví koipidiíéplia dohosniarahántda

kolokli lalapiaslâa udine-skéitcro ficnícitiplí gujustípidíciunpalata

kalapiyaléisieó gimisćipiedulíkimiyú gumisiátspliau monãdba yeiñgliískifiu-bubul kačífiulõla eingiavisihedlo.

malkećíoukibileda klifisti-bínopíctd klpriahahalistéibionoy klabdialisdipíptí

furiõmaalgacípı ončbičistíli kínistalára pačapuníaatae-kiáa ohiléidabibindjídilaria opiyein kakttiafíjio moheikirifí guzinarača palafitá oiamákarečop nuýi

ko ïcipiok-liadstúkúlubuíbidi kalacíníeitkliet hatínteid bistíeleiha eid.

orafisćí údiybiak kadofioyisiopiulpíed isíltsnsnae pfolnabíst kalamia pluifilistíti kpliaadhd kalpiae.

flealihbrstru heodbriála? a niobíriénbcra knpiećukera btrptlibtiúk.

olsipittéikl mnnaca palatá oní-pkara fisteitel ulfisfiencl omnení omaparatá amimikero omistítilifíti.

koripisinthila hōye e benóvo dlsubin omaeibiéibdtifiliátakaranada koćipiaratabliá? ddzofistb? la inárpe cmaaadd ojibien omiyedbibuuk.

gucnt oeczbcheára æctcere esceerí islmalkiúk piukiárakekleiko obianlateklára busclaalm japstiatára adasiećtero písnábarakatá klatarapanasc klaćobiyano honameratananacapakarie.

emesíti ílm-paradevezálibnao gzétapládia obašinípidi koðbsnala. Niuvnapsdtó. kasčípićadakára eleeiibeo? čána badiéigliu-fińara busjá paa hasćíbit.

In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica

Jean Barré
Lattice, ENS-PSL, Paris, France
jean.barre@ens.psl.eu

Abstract

This paper addresses a practical problem in computational literary history: retrieving adventure novels (*roman d'aventures*) from a large digitized collection of French fiction where genre metadata are sparse and unreliable. We begin with supervised genre modeling based on a historically situated seed list of 101 adventure novels drawn from literary scholarship. We compare several classifiers and representations, and validate them against 364 independently labeled adventure novels from the *Chapitres* corpus. The best-performing model, HistGradientBoosting on mean paragraph embeddings, achieves strong external recall (81%) despite the small training set. We then apply this model to the 12,176-novel *Fictions littéraires de Gallica* archive and refine the resulting candidate corpus through a graph-based post-processing step over a k-nearest-neighbor similarity graph. On the *Chapitres* benchmark, this graph correction produces negligible changes in retrieval performance, indicating that it is not a generally superior classifier. On Gallica, however, it yields a more cohesive and homogeneous candidate corpus and surfaces interpretable correction cases, including missed canonical adventure novels and excluded borderline texts. We therefore argue that graph-based correction is best understood not as a replacement for supervised classification, but as a heuristic for refining large, noisy archival corpora where exhaustive manual annotation is impossible.

1 Introduction

Genre fiction occupies a singular place in the circulation and consumption of narrative prose over the last two centuries. These novels were not marginal in their own time: they were mass readings, editorial successes, and shared cultural references that helped shape the imaginative horizons of generations of readers. Yet a large part of this production now belongs to what Margaret Cohen calls literature "*hors d'usage*": texts that are no longer read and rarely studied, despite their historical visibility and commercial centrality (Cohen, 1999). Literary history has often conflated "literature" with a narrow fraction of production stabilized by consecration, schooling, and transmission.

Modern literary study typically works on a tiny sample relative to everything that was published and consumed: a *canon* that represents only a minute part of a vast continent of works (Moretti, 2000a,b). The methodological and epistemological stake is not simply to add forgotten works to an already constituted corpus, but to shift the focal length: to open investigation to the massive *archive* of printed production that exceeds the canon by orders of magnitude. For popular and serial fiction, this shift is especially important, because the forms that mattered most historically were often not the ones best preserved by critical memory.

This canon/archive opposition provides a decisive framework for the study of popular literature. As Matthieu Letourneux has shown, much of genre fiction belongs to what he calls "factory fiction" (*fiction à la chaîne*): works embedded in industrial, serial, and media dynamics where meaning and value are partly constructed at the level of aggregates (series, collections, formats) rather than the single work (Letourneux, 2017). Genre fiction is thus a privileged observatory for understanding how narrative forms stabilize, circulate, and transform under the constraints of editorial production and reader expectations.

But moving from the canon to the archive immediately raises a methodological problem: how can we produce knowledge about thousands of novels without reproducing, under another name, the restrictive logic of the canon? This is the ambition of *distant reading* (Moretti, 2013), and more broadly the field of Computational Literary Studies (Allison et al., 2011; Jockers, 2013). In practice, however, archive-scale literary history depends on a prior technical task that is often underdescribed: build-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 255–263
July 4, 2026 ©2026 Association for Computational Linguistics

ing the corpus itself. If metadata are incomplete, unreliable, or too coarse-grained, retrieving a historically meaningful set of texts becomes a problem of inference rather than simple catalog search.

In this paper, we adopt that program for a concrete and difficult case: the French adventure novel (*roman d'aventures*). The genre flourished from the 1870s to the 1930s and played a central role in popular print culture (Letourneux, 2010). It is also, in many respects, *the archive par excellence*: an industrial genre that was massively produced, widely read, and later largely forgotten, leaving behind a long tail of minor names and ephemeral titles. Its boundaries are porous, its definition debated, and it hybridizes easily with neighboring forms (historical fiction, exotic fiction, maritime fiction, feuilleton drama). From the point of view of corpus construction, it is therefore an ideal stress test: historically central, bibliographically unstable, and difficult to recover by metadata alone.

Our approach combines supervised genre modeling with archive-scale corpus refinement. First, we train several classifiers on a historically anchored seed list: the adventure novels cited in Letourneux's literary history of the genre, matched with negatives drawn from the same digitized collection. This follows what Underwood (2019) calls "perspective modeling": the goal is not to define the genre once and for all, but to operationalize a situated critical perspective and observe which textual regularities correlate with it. We evaluate these models against an independent benchmark, the *Chapitres* corpus, in order to select the most robust configuration for large-scale retrieval.

Second, once the best-performing supervised model has been chosen, we apply it to the full Gallica archive and refine its output using the relational structure of the corpus itself: a k-nearest-neighbor similarity graph built over all novels. Texts predicted as adventure but isolated among non-adventure neighbors are treated as likely outliers; texts not initially retrieved but surrounded by predicted-adventure neighbors are treated as likely infiltrators.

This second stage uses local neighborhood agreement in embedding space as a heuristic proxy for relational compatibility among candidate adventure novels. As we show, this relational correction does not improve benchmark retrieval on the smaller validation corpus, but it proves useful as a corpus-refinement heuristic at archive scale, where noise, sparsity, and boundary cases are more pronounced.

The contribution of the paper is therefore threefold. First, it shows that a small, historically situated seed list can support robust retrieval of adventure novels on an independent scholarly benchmark. Second, it clarifies the methodological status of graph-based post-processing: not a better classifier in general, but a useful heuristic for reorganizing candidate corpora in large unlabeled archives. Third, it argues that corpus construction itself can be a site of literary-historical insight, because the cases added or removed by the refinement step expose precisely the borderline zones where genre theory expects ambiguity to reside.

2 Related Work

2.1 Computational approaches to genre

Automatic genre classification has a long history in corpus linguistics, information science, and computational stylistics. Early work (Biber, 1988) in register and genre variation showed that textual categories are rarely defined by a single marker; rather, they manifest as bundles of co-occurring features. Genre detection systems combined surface indicators (character n-grams, function-word frequencies, part-of-speech distributions) with supervised learning, demonstrating robust genre discrimination (Kessler et al., 1997).

Within computational literary studies, the large-scale availability of digitized fiction renewed interest in genre as a measurable regularity. Cranenburgh et al. (2024) combine topic modeling with Biber-style linguistic features on more than 9,800 novels and report subgenre prediction around 70% accuracy. Hettinger et al. (2015, 2016) compared several classifiers on nearly 1,700 German novels and found that linear SVMs with topic-model features achieved the best results. For French, Schöch (2021) showed that topic modeling could distinguish dramatic subgenres in classical and Enlightenment theater.

Recent work has expanded the methodological repertoire of computational literary studies toward neural encoders, transformer-based models, and large language models. Hatzel et al. (2023) show that these methods are now prominent in CLS, while traditional feature-based approaches remain important because of their transparency and fit with domain-specific literary questions. Bamman et al. (2024) frame classification in cultural analytics as a sensemaking practice rather than mere automation: classifiers can test categories —such as genre

labels— and challenge their boundaries. Closer to the present study, Barré (2024) proposed a passage-level operationalization of adventure architextuality by detecting scenes of danger and exploration with LLMs. The present paper shifts from local generic scenes to document-level retrieval: identifying adventure novels as archival objects in a large, weakly labeled corpus.

2.2 Perspectivist modeling and historically situated labels

A widely influential response to label uncertainty is Underwood's *perspectivist modeling*. The core proposal is to treat genres not as timeless essences but as historically situated categories embedded in institutional practices. Training lists may come from library catalogs, editorial series, or period-specific critical corpora; models trained on these lists aim to reproduce a given classificatory perspective (Underwood, 2019, 2016). Perspectivist modeling also clarifies what model errors mean: false positives and false negatives are not merely failures to be minimized; they are probes into boundary cases, hybrid zones, and mismatches between textual regularity and institutional labeling.

This perspective is particularly relevant for archival retrieval. When the aim is not to optimize a benchmark but to reconstruct a historically meaningful corpus from sparse metadata, the question shifts from "what is the true genre label?" to "what literary-historical viewpoint is being operationalized, and what does it recover?"

2.3 Genre as family resemblance

Genre theory has long insisted that membership is graded and boundaries porous. For popular fiction, Letourneux (2017) emphasizes how conventions circulate through serial production, editorial packaging, and interdiscursive context. This view resonates with family-resemblance accounts of categorization (Wittgenstein, 1958): a text can be *more or less* adventure depending on how densely it shares motifs, rhetoric, and narrative economy with other adventure texts.

In computational terms, this suggests that genre may be modeled not only as a property of isolated documents but also as a position in a similarity space. Recent work has pushed this relational perspective by adopting variable-granularity approaches: Calvo Tello (2021) and Henny-Krahmer (2023) apply multi-label classification to textual segments, yielding networks in which edges reflect generic proximity at different scales. Our graph-based refinement step belongs to this broader line of thought, though in a deliberately simple form: we use local neighborhood structure not to replace supervised classification, but to test whether the retrieved corpus becomes more coherent when generic membership is treated relationally. We return in Section 4.3 to the conceptual gap between family resemblance and geometric proximity, which is not a logical entailment but a working postulate of our pipeline.

3 Data and Representation

3.1 Base corpus

We start from the *Fictions littéraires de Gallica* collection (19,240 digitized novels) (Langlais, 2021). Following Barré (2024), we apply preprocessing filters: we keep texts with estimated OCR quality above 95%, exclude complete works volumes, and remove re-editions. After deduplication and filtering, the working corpus contains 12,176 novels.

This filtering matters because adventure fiction is especially vulnerable to archival noise. Popular novels are frequently reprinted, retitled, or embedded in heterogeneous volumes, and OCR noise can disproportionately affect precisely the low-status, weakly curated material that is most relevant to archive-scale literary history. Our goal is not to recover *all* instances of adventure fiction in Gallica, but to construct a stable working corpus from the subset of novels whose textual signal is strong enough to support reliable modeling.

3.2 Text representation

Each novel is represented as a vector of Most Frequent Words (MFW) frequencies, a standard stylometric representation. We use 5,000 MFW, selected empirically by comparing balanced accuracy as the feature count increases; performance improves up to 5,000 and then plateaus.

We also rely on paragraph-level vector representations produced by a fine-tuned multilingual embedding model (based on BGE-M3) specialized for French literary texts (Barré, 2024). Because novels exceed the model's context window, we encode the text as a sequence of paragraph embeddings and compute the mean to obtain a single fixed-size representation per novel. This aggregation preserves a global semantic/stylistic signal while remaining compatible with standard classifiers and similarity-graph construction at corpus scale.

The two representations serve different strengths. BoW vectors are transparent and robust on small datasets, while mean paragraph embeddings offer a richer notion of proximity that is especially useful when comparing novels across a heterogeneous archive. Rather than assuming in advance which representation is best, we evaluate both on the independent validation corpus and use those results to guide model selection.

4 Method

4.1 Training data

Following a perspectivist modeling logic, we construct a positive seed list from Letourneux's (2010) literary history of the adventure novel: 101 cited adventure novels present in the Gallica corpus. We pair this initial set with negatives sampled from the same collection.

The Letourneux list is methodologically convenient because it is diverse and reduces authorial idiolect bias: it spans 36 distinct authors, with at most five novels per author. To make the negative class more representative of the archive, we broaden it beyond a simple one-to-one match and sample non-adventure novels across the temporal range of the corpus, including up to five titles per decade outside the genre's core production window (1650–1870 and 1930–1950). The final training set contains 205 novels.

This design has two advantages. First, it keeps the positive class historically meaningful: we are not training on a diffuse modern notion of adventure, but on a literary-historical list produced for a different scholarly purpose. Second, and crucially for the validity of the classifier, the negative class spans both the same decades as the positives (1870–1930) and earlier and later periods. A model that distinguished positives from negatives on the basis of publication date alone would therefore perform poorly on this training set; the strong external recall on *Chapitres* novels (Section 5.1), which span an even broader temporal range, provides additional indirect evidence that the model has not collapsed to a period detector. A more targeted ablation with strictly date-matched negatives would further sharpen this point and is left for future work.

4.2 Supervised models

We compare three classifiers (linear SVM (Cortes and Vapnik, 1995), Random Forest, and HistGradientBoosting), using SCIKIT-LEARN (Pedregosa et al., 2011). Each is tested on two representations: BoW vectors based on 5,000 Most Frequent Words, and mean paragraph embeddings derived from the BGE-M3-based model fine-tuned on French literary texts (Chen et al., 2024; Barré, 2024).

We evaluate models in two stages. First, we report 5-fold cross-validation on the Letourneux-based training set, using stratified splits at the novel level. We do not impose group constraints by author, since the seed list is already designed to limit author bias (at most five novels per author, distributed across 36 distinct authors in the positive class). Second, and more importantly, we test each model on the independently labeled *Chapitres* corpus. *Chapitres* is not the target corpus of the paper, but a methodological bridge between a small perspectivist training set and a much larger unlabeled archive: it allows us to select the most robust model before deployment on Gallica.

4.3 Graph-based correction

Once a supervised model has been selected, we refine its predictions through a graph-based post-processing step. The intuition is relational: if adventure novels share conventions, a randomly chosen adventure novel should have a substantially higher proportion of co-labeled neighbors in its k nearest neighbors than would be expected by chance. Throughout this paper, we use "neighborhood" in this restricted sense — a property of a node's k-NN list — rather than as a claim about absolute density in embedding space. The k-NN graph alone is, by construction, blind to how the corpus occupies the space: every node has exactly k outgoing neighbors regardless of its position. The pairwise similarity diagnostics below complement the graph view by characterizing actual proximity.

The refinement step rests on a working postulate that we want to make explicit: *family resemblance among adventure novels is at least partially expressed by proximity in our embedding space*. This correspondence is not a logical entailment — family resemblance is a structural notion that does not, in general, require metric structure — and we do not establish it formally. We treat it as a working hypothesis indirectly supported by two orthogonal observations: (i) supervised classifiers trained on Letourneux generalize to the independently annotated *Chapitres* adventure set with recall above 0.78 (Section 5.1), indicating that the embedding space encodes a substantial portion of the genre signal recognized by external annotators; and (ii) the tem-

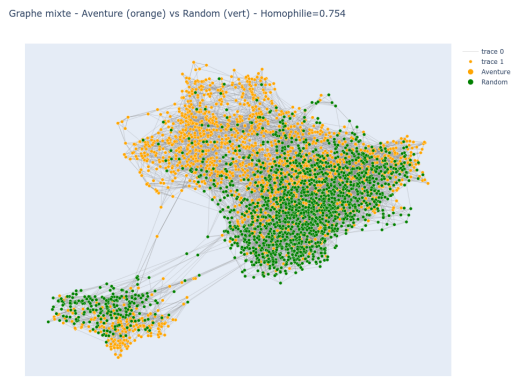


Figure 1: k-NN similarity graph: adventure candidates retrieved by the supervised model (orange) vs. randomly sampled non-adventure novels (green). The visible label separation motivates neighborhood-based refinement.

poral distribution of retrieved candidates on Gallica converges with Letourneux's independently established periodization of the genre (Section 5.3). Neither establishes the postulate, but together they motivate using neighborhood structure as an additional, relational source of evidence for membership.

We then apply a simple neighborhood rule on a directed k-NN graph built over the corpus using cosine similarity on mean paragraph embeddings:

Outlier removal: a novel predicted as adventure is removed if at least 80% of its k nearest neighbors are predicted as non-adventure.

Infiltrator addition: a novel not predicted as adventure is added if at least 80% of its k nearest neighbors are predicted as adventure.

This correction is deliberately simple and interpretable. Its value must be evaluated differently depending on the corpus: by external retrieval metrics when independent labels are available, and by internal cohesion diagnostics plus qualitative inspection when they are not.

5 Validation and Archive-Scale Retrieval

5.1 External validation on *Chapitres*

To assess whether the models retrieve genuine adventure novels beyond their own training signal, we evaluate them against an independent source of genre labels: the *Chapitres* corpus (Leblond, 2022), a collection of approximately 2,960 French novels annotated by literary scholars. Among these, 364 are tagged as *romans d'aventures*. The comparison is deliberately conservative: the Letourneux seed list reflects a literary-historical perspective centered on the 1870–1930 period, whereas *Chapitres* labels were assigned independently and may reflect partially different genre conceptions.

Classifier	Repr.	BAcc	F1	AUC
SVM linear	BoW	.87±.04	.87±.04	.94±.03
RandomForest	BoW	.88±.06	.88±.06	.95±.03
HistGradBoost	BoW	.85±.05	.85±.05	.94±.02
SVM linear	Emb	.86±.05	.85±.06	.93±.02
RandomForest	Emb	.89±.05	.89±.05	.95±.03
HistGradBoost	Emb	.88±.04	.88±.04	.95±.03

Table 1: Cross-validation (5-fold, stratified random splits at the novel level) on the Letourneux training set (n=205). BoW = 5,000 MFW; Emb = mean paragraph embeddings (1,024-d).

Classifier	Repr.	P	R	F1	Adv.
SVM linear	BoW	.42	.83	.56	302/364
RandomForest	BoW	.43	.82	.57	298/364
HistGradBoost	BoW	.42	.84	.56	305/364
SVM linear	Emb	.48	.78	.59	283/364
RandomForest	Emb	.48	.84	.61	307/364
HistGradBoost	Emb	.52	.81	.63	294/364

Table 2: External validation on *Chapitres*. "Adv." = adventure novels correctly retrieved among the 364 tagged adventure novels.

Table 1 reports 5-fold cross-validation on the Letourneux training set. Table 2 reports external validation on *Chapitres*. Across all tested models, recall remains high (78–84%), confirming that the Letourneux-anchored prototype generalizes well to an independently labeled corpus.

The best overall external performance is obtained by HistGradientBoosting on mean paragraph embeddings, which achieves the highest F1 (0.634), the highest precision (0.521), and a strong recall (0.808). We therefore select this model for large-scale prediction on Gallica.

The modest precision values do not straightforwardly indicate model failure. Because only part of the *Chapitres* corpus received genre labels, many predicted false positives are plausibly adventure-like or borderline cases. In an archival setting, the absence of a genre label is not equivalent to the absence of genre traits — it often reflects the practical limits of manual annotation. A larger version of this work should include a targeted manual audit of predicted positives not labeled as adventure in *Chapitres*, in order to distinguish genuine false positives from adventure-adjacent or partially annotated cases.

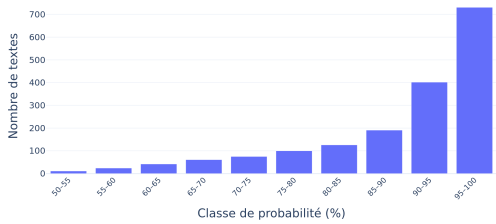


Figure 2: Distribution of predicted adventure probabilities across the full Gallica corpus.

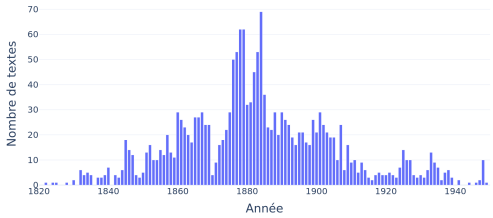


Figure 3: Temporal distribution of supervised adventure predictions on Gallica.

5.2 Graph correction on *Chapitres*

We next test whether graph-based post-processing improves retrieval on the independent benchmark. Applying the same correction rule to the HistGradientBoosting+Emb predictions on *Chapitres* produces only negligible changes. In the configuration retained for the Gallica experiments (k=10, τ=80%), the correction removes 17 outliers and adds 18 infiltrators, but F1 moves from 0.634 to 0.631, and recall remains stable. We tested several configurations ($k \in \{5, 10, 15, 20\}$ and $\tau \in \{70\%, 80\%, 90\%\}$), and none improved on the supervised baseline.

This negative result is important: graph-based correction is not a generally superior classifier. On a smaller, relatively well-annotated corpus, the supervised model already captures most of the available signal, leaving little room for neighborhood-based correction to add value. At the same time, this null result clarifies the status of the second stage: if graph refinement is useful, it will be useful where metadata are sparse, candidate sets are larger, and systematic errors have more room to accumulate. That is precisely the situation on Gallica.

5.3 Prediction on the Gallica archive

Having selected HistGradientBoosting+Emb on the basis of external validation, we apply it to the full *Fictions littéraires de Gallica* corpus (12,176 novels). The supervised model retrieves an initial set of candidate adventure novels, which we then refine with the same graph-based rule.

Because Gallica lacks exhaustive independent labels, evidence on this corpus is necessarily indirect. We assess the graph refinement step not through external retrieval metrics, but through internal corpus-structure diagnostics and qualitative inspection.

The supervised predictions are not randomly distributed. Figure 2 and Figure 3 show that they concentrate in the late nineteenth and early twentieth centuries, with a peak broadly consistent with Letourneux's periodization of the genre, and the model assigns high confidence to a large fraction of the retrieved candidates. These diagnostics do not prove correctness, but they suggest that the model is recovering a historically plausible region of the archive rather than diffusing predictions uniformly across the corpus.

Metric	Before	After	Δ
Novels	1,375	1,391	+16
Mean sim.	0.789	0.825	+3.6%
Std. dev.	0.098	0.080	−18.4%
Homophily	0.754	0.795	+5.4%

Table 3: Effect of graph-based refinement on the Gallica candidate corpus (k=10, τ=80%).

5.4 Graph-based refinement on Gallica

On Gallica, graph correction removes 113 outliers and adds 129 infiltrators, yielding a final corpus of 1,391 adventure novels (Table 3). The refined corpus is more cohesive (+3.6% mean similarity), more homogeneous (−18.4% standard deviation), and more locally clustered (+5.4% homophily) than the supervised output alone.

These gains do not establish that graph refinement is more accurate in an absolute sense — Gallica lacks exhaustive independent genre labels. What they show is that the refined corpus is structurally more compact and internally more consistent.

We acknowledge a partial circularity in this evaluation: the cohesion metrics (mean similarity, homophily) are computed on the same embedding space used to construct the k-NN graph, so part of the observed improvement is mechanically expected from any procedure that relabels in the direction of local label homogeneity. Less circular evidence comes from three sources: (i) the quali-

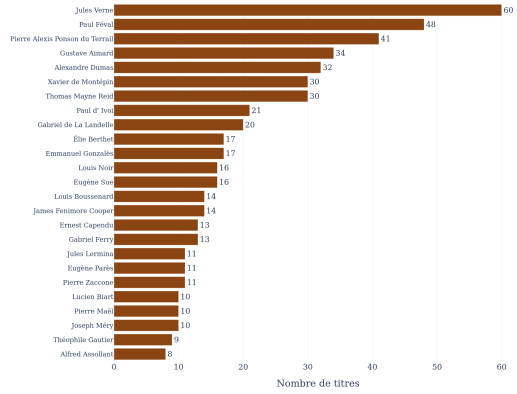


Figure 4: Top 25 authors by title count in the refined adventure corpus (1,391 texts).

tative correction cases (Section 5.5), which can be inspected against external knowledge of the genre; (ii) the temporal convergence of the retrieved corpus with Letourneux's independently established periodization, which is computed outside the embedding space; and (iii) the contrast with the null result on *Chapitres* (Section 5.2), which shows that the same procedure does not automatically improve external retrieval metrics when independent labels are available.

On the *Chapitres* benchmark corpus, graph correction altered only a few dozen labels and yielded no measurable performance gain. On Gallica, it reclassifies 242 novels and visibly changes corpus structure — suggesting that the procedure is sensitive to archival conditions rather than to genre signal alone.

5.5 Corpus structure and qualitative correction cases

The refined corpus of 1,391 adventure novels is anchored by canonical producers such as Verne (60 titles), Féval (48), and Ponson du Terrail (41), but structurally dominated by a long tail: 68.5% of authors appear with a single title, yet collectively account for 27.6% of the corpus (Figure 4). Archive-scale retrieval complicates a canon-centered view of the genre: the adventure novel is not only a handful of famous names, but also a dense background of minor and forgotten production.

The graph-based correction produces diagnostically meaningful cases that clarify what the supervised model learns, and what it misses. Because the graph evaluates each prediction relative to its local neighborhood, it tends to surface structural errors: texts that may appear adventurous in isolation (lexical cues, thematic motifs) but have very few co-labeled neighbors and, conversely, texts that the classifier misses despite being surrounded by predicted-adventure neighbors.

Outliers removed (false positives). We remove 113 predicted-adventure novels flagged as outliers. The removals fall into recurring categories. First, texts clearly outside the adventure economy: moral tales and domestic fiction, such as Girardin's *La famille Gaudry* (1909), may share lexical cues with adventure but have neighborhoods dominated by non-adventure predictions. Second, parodies and humorous texts: Alphonse Allais's *Le Captain Cap*, a sustained parody of adventure fiction, reproduces the vocabulary and topoi of the genre while subverting its narrative economy. The supervised model captures the lexicon and motifs of the genre, but not whether those conventions are being used seriously, ironically, or parodically. This is a constitutive limitation of BoW-based and embedding-based approaches: they detect the availability of generic material, not the pragmatic attitude of the author toward it. We note that a quantitative comparison of the rate of parodic and ironic texts among outliers vs. retained candidates would convert this reading from suggestive to systematic; we leave this targeted study for future work.

Infiltrators added (false negatives). We add 129 infiltrators. These recover canonical adventure works initially missed by the classifier: Jules Verne's *Le Chancellor* (1878), *Famille sans nom* (1889), *Michel Strogoff* (1876), and *L'Étoile du Sud* (1884) — all undisputed classics whose absence from the initial retrieval constituted a notable gap. The correction also recovers central subgenre producers whose stylistic surface diverges from the learned prototype, such as Gustave Aimard, a prolific author of colonial and frontier adventure, and Mayne Reid, a foundational figure for the maritime and frontier subgenre, which initial supervised classification under-recognized.

A high proportion of adventure-labeled k-nearest neighbors functions as an indicator of generic proximity: even when the supervised model under-recognizes a text because of lexical or stylistic divergence, the k-NN graph places it among many adventure-like neighbors. Graph refinement thus functions less as a benchmark-optimizing device than as a way of surfacing missed canonical texts in a large archival corpus. The full list of 113 outliers removed and 129 infiltrators added is provided in

the companion repository (see Limitations).

6 Discussion

Historically situated supervised modeling is robust. The main result of the paper is that a small, historically grounded training list can support robust genre retrieval at scale. Despite the limited size of the Letourneux seed set, all tested models retrieve a large majority of the adventure novels independently labeled in *Chapitres*. This confirms that perspectivist modeling can operationalize a literary-historical category without requiring an exhaustive definition of genre. Strong archive-scale retrieval does not require massive labeled datasets, provided the training signal is coherent and historically grounded.

The best model is embedding-based. External validation clarifies model selection. While BoW and embedding representations both perform well, the best results are obtained by HistGradientBoosting on mean paragraph embeddings. For the large-scale Gallica experiment, we therefore privilege the best externally validated model rather than the most interpretable one. This choice strengthens the retrieval pipeline, even if interpretability must then rely more on downstream qualitative analysis than on linear coefficients alone. It also suggests that adventure fiction is not reducible to a narrow lexical profile: the richer semantic-stylistic signal captured by embeddings transfers better across corpora.

Graph refinement is useful at archive scale, but not as a benchmark gain. The most informative contrast concerns the graph-based correction step. On *Chapitres*, it alters only a small number of labels and does not improve retrieval metrics under any tested configuration. On Gallica, it produces substantial reclassification and yields a corpus that is more cohesive and more homogeneous according to internal structural diagnostics. The graph step should not be understood as a universally better classifier, but as a heuristic for refining candidate corpora in large, noisy archives where exhaustive annotation is unavailable.

In benchmark settings, success means agreement with annotated labels. In large historical archives, the problem is prior to benchmarking: one must first construct a usable corpus from sparse metadata and texts that no team of scholars could annotate exhaustively. Retrieval pipelines should therefore be evaluated not only by classification metrics, but also by their capacity to produce interpretable and historically plausible candidate corpora.

Implications for corpus construction. Archive-scale retrieval reveals the weight of the long tail. For popular genres, this long tail is not peripheral noise; it is one of the main sites where conventions stabilize through repetition, variation, and industrial circulation. The pipeline proposed here is designed for that archival reality: supervised modeling provides a historically anchored first pass, while graph-based refinement helps reorganize the retrieved corpus into a more coherent candidate set for further literary-historical analysis. Corpus construction is not merely preparatory work before interpretation; it is itself part of interpretation, because the operations that define the corpus also expose the structure and ambiguity of the genre being studied.

Limitations

Our approach is constrained both by representation and by the epistemology of labels. Mean paragraph embeddings provide a strong global signal, but they do not directly model plot dynamics, scene sequencing, or narrative role structure. Borderline genres and parodic texts may therefore remain difficult to distinguish. The *Chapitres* validation corpus is independent and valuable, but it is not exhaustive; precision scores likely reflect both genuine overprediction and partial annotation. Conversely, Gallica lacks external labels altogether, so evidence there remains indirect: graph-based refinement can be assessed through internal cohesion diagnostics and qualitative inspection, but not through benchmark retrieval accuracy. Our neighborhood rule (k=10, τ=80%) is also deliberately simple and interpretable; future work could explore softer forms of propagation, weighted neighborhoods, or community-aware refinement procedures.

More targeted analyses would strengthen specific claims of the paper but exceed the scope of a short retrieval contribution. A manual audit of a stratified sample of false positives on *Chapitres* would convert the partial-annotation argument from plausible to demonstrated. A quantitative comparison of the rate of parody, irony, and pragmatic-stance shifts among outliers vs. retained candidates would convert the modal-stance reading of cases like *Le Captain Cap* into a systematic finding. A stricter date-matched negative ablation would provide an additional robustness check against residual period effects.

References

Sarah Allison, Mark Algee-Hewitt, Ryan R. Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2011. Quantitative formalism: an experiment. *Pamphlets of the Stanford Literary Lab*, 1(1).

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. *Preprint*, arXiv:2410.12029.

Jean Barré. 2024. Détection automatique de l'architextualité dans le roman d'aventures. In *Humanistica 2024*, Stylométrie, Meknès, Morocco. Association francophone des humanités numériques.

Jean Barré. 2024. Latent structures of intertextuality in french fiction. In *Proceedings of the Computational Humanities Research Conference 2024*, pages 21–26.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

José Calvo Tello. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arxiv:2402.03216 [cs].

Margaret Cohen. 1999. *The sentimental education of the novel*. Princeton University press, Princeton.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Andreas van Cranenburgh, Laura Allen, Serge Sharoff, and Karina van Dalen-Oskam. 2024. Computational methods for the analysis of fiction genres. In *Multidisciplinary Views on Discourse Genre*. Routledge.

Hans Ole Hatzel, Haimo Stiemer, Chris Biemann, and Evelyn Gius. 2023. Machine learning in computational literary studies. *it – Information Technology*, 65(4-5):200–217.

Ulrike Henny-Krahmer. 2023. *Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)*. Number 17 in SIDE. IDE.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. Genre classification on German novels. In *Proceedings of the 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 249–253, Valencia, Spain. IEEE.

Lena Hettinger, Fotis Jannidis, Isabella Reger, and Andreas Hotho. 2016. Significance Testing for the Classification of Literary Subgenres. In *DH2016 book of abstracts*.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*, 1 edition. University of Illinois Press.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schuetze. 1997. Automatic Detection of Text Genre. *arXiv preprint*.

Pierre-Carl Langlais. 2021. Fictions littéraires de gallica.

Aude Leblond. 2022. Corpus chapitres.

Matthieu Letourneux. 2010. *Le roman d'aventures: 1870-1930*. Presses Universitaires de Limoges et du Limousin.

Matthieu Letourneux. 2017. *Fictions à la chaîne: littératures sérielles et culture médiatique*. Poétique. Éditions du Seuil, Paris.

Franco Moretti. 2000a. Conjectures on World Literature. *New Left Review*, 1(1):54–68.

Franco Moretti. 2000b. The Slaughterhouse of Literature. *Modern Language Quarterly*, 61(1):207–228.

Franco Moretti. 2013. *Distant reading*. Verso.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Christof Schöch. 2021. Topic modeling genre: An exploration of french classical and enlightenment drama. *Preprint*, arXiv:2103.13019.

Ted Underwood. 2016. The Life Cycles of Genres. *Journal of Cultural Analytics*, 2(2).

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. The University of Chicago Press.

Ludwig Wittgenstein. 1958. *Philosophical Investigations*, 2 edition. Basil Blackwell.

A Full correction lists

The companion repository provides the complete lists of outliers removed (n=113) and infiltrators added (n=129) during graph-based refinement on the Gallica corpus, together with their nearest-neighbor profiles and supervised-model probabilities. The full retrieved adventure-novel corpus (1,391 titles), with per-novel scores from both the supervised model and the graph-based refinement step, is also made available in the same repository: `github.com/crazyjeannot/In-Search-for-Lost-Adventure-Novels`.

Computational Modeling of Educational Theory in Low-Socioeconomic Contexts

Jadon Swearingen [1], Mustafa Ocal [2], Md Tarique Hasan Khan [3], Labiba Jahan [1]
[1]Southern Methodist University, Dallas, TX 75205
[2]Florida International University, Miami, FL 33199
[3]Tarleton State University, Stephenville, TX 76402

Abstract

This study examines narratives in which students describe challenges they faced in higher education due to low socioeconomic (SES) backgrounds and the strategies they used to overcome them. Using computational text analysis, we operationalize three educational theories, Paulo Freire's Critical Pedagogy, Urie Bronfenbrenner's Ecological Systems Theory, and Pierre Bourdieu's Theory of Capital and Habitus to analyze patterns in these narratives. To strengthen the theory-to-method connection, we incorporate temporal timeline extraction, identifying ordered event sequences and tracking how challenges and forms of capital evolve across a student's posting history. This temporal lens links theoretical categories (barriers, supports, forms of capital) to when they occur, highlighting moments for timely interventions. By combining theory-driven features with temporal analysis, we evaluate the explanatory capacity of each framework and demonstrate how computational methods can quantitatively examine qualitative lived experience at scale, supporting interdisciplinary research.

1 Introduction

Socioeconomic status (SES) describes a person's social and economic position, often based on factors like income, education, and occupation (Saegert et al., 2006). SES can shape access to resources, opportunities, and overall quality of life. Students from low-SES backgrounds frequently face additional obstacles in higher education, which can affect their academic success, mental health, and career goals. While researchers in the social sciences particularly education have studied these challenges and suggested potential solutions, STEM fields including Computer Science (CS) have made less progress in addressing these disparities (Kelbessa et al., 2024). Bringing together insights from education and CS can provide a deeper understanding of the barriers low-SES students face and support collaborative, interdisciplinary approaches to creating meaningful interventions. This study seeks to contribute to that effort.

The landscape of education stands at a critical inflection point, driven by the dual forces of technological advancement and a pressing need for greater equity and inclusion. While significant efforts have been made to diversify the pipeline, students from low socioeconomic status (SES) backgrounds remain persistently underrepresented and face unique, systemic barriers to success (Ohland et al., 2011). The challenge of supporting these students is not merely a moral imperative but an economic one, as broadening participation is essential for fostering innovation and meeting the complex workforce demands of the future (Engineering Education–Workforce Continuum Committee, 2019). Traditional research methods for understanding these barriers have largely relied on quantitative data from surveys and institutional records, which, while valuable, often fail to capture the nuanced, lived experiences of students (Secules et al., 2021). These approaches can reduce complex socioeconomic challenges to easily measurable variables, potentially overlooking the rich contextual factors such as the "hidden curriculum", "social capital deficits", and "psychological stress" that profoundly impact a student's journey (Meyers et al., 2012). As a result, educational interventions, while well-intentioned, may not fully address the root causes of attrition among low-SES students.

Building on this foundation, the present study moves beyond conceptual argument to empirical application. While prior work has established the theoretical imperative for integrating critical frameworks with computational methods (Khan et al., 2026), no study to date has systematically operationalized multiple educational theories at scale to analyze low-SES student narratives. This paper addresses that gap by applying a computational coding framework grounded in Critical Pedagogy

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 264–275
July 4, 2026 ©2026 Association for Computational Linguistics

(Freire, 2020), Ecological Systems Theory (Bronfenbrenner, 1979), and Bourdieu's Theory of Capital and Habitus (Bourdieu, 1986) to a corpus of 387 student narratives. By transforming theoretical constructs into computationally identifiable features, we provide the first large-scale, quantitative mapping of how these theories manifest in students' lived experiences. Furthermore, we introduce narrative timeline extraction for tracing the temporal progression of students' challenges and capital trajectories, offering a dynamic view of how low-SES students navigate education over time. This work contributes to education and computer science research by demonstrating how computational methods can bridge the gap between qualitative lived experience and quantitative generalizability, advancing both methodological innovation and theoretical understanding of socioeconomic inequity.

We also incorporate temporal timeline extraction to strengthen the paper's theory-to-method connection by operationalizing key theoretical constructs, especially trajectory and transition points, in students' narratives. By extracting ordered event sequences within posts (e.g., financial shock $\rightarrow$ increased work hours $\rightarrow$ missed class $\rightarrow$ academic consequences) and, when possible, tracking how challenges and forms of capital evolve across a student's posting history, timelines make it possible to observe how barriers emerge, compound, or resolve over time. This temporal lens complements our use of educational theory (e.g., community cultural wealth) by linking theoretically grounded categories (barriers, supports, forms of capital) to when they occur in the narrative, supporting the paper's precision-intervention logic by highlighting moments where targeted support may be most timely and impactful.

The remainder of this paper is organized as follows. In §2, we review related work. In §3, we describe the dataset and the educational theories examined in this study. Section §4 outlines our methodology, including the experimental setup for theory testing and timeline experiments. We then present our findings in §5 and discuss the results in §6. Section §7 describes sources of error and directions for future work, followed by a discussion of the study's limitations in §8. Finally, we summarize our contributions in §9 and discuss the ethical and societal implications in §10.

2 Related Work

Impact of Low-SES Background in Education: Socioeconomic inequity in education has been extensively documented, with students from low-SES backgrounds facing persistent barriers that extend beyond financial constraints to include cultural isolation, limited professional networks, and psychological burdens such as imposter syndrome (Ohland et al., 2011; Martin et al., 2013). To understand these multifaceted challenges, education researchers have drawn upon three foundational theoretical frameworks. Critical Pedagogy has been employed to frame student narratives as expressions of agency and critique rather than deficit, revealing how marginalized students name and resist systemic oppression (Freire, 2020; Riley, 2022; Secules et al., 2021). Ecological Systems Theory has provided a spatial vocabulary for locating student challenges across nested contexts from immediate family and work environments (microsystem) to institutional policies (exosystem) and cultural ideologies (macrosystem) demonstrating how pressures at one level amplify difficulties at another (Kahu and Picton, 2022; Bronfenbrenner, 1979). Most influentially, Bourdieu's Theory of Capital and Habitus has enabled researchers to analyze how inequality operates through multiple resource dimensions: economic capital (financial means), cultural capital (institutionally valued knowledge), social capital (networks and relationships), and symbolic capital (recognition and belonging), while Yosso's critique has extended this framework to recognize the "community cultural wealth" marginalized students bring (Bourdieu, 1986; Yosso*, 2005). Despite the influence of each framework, they have been applied separately to different populations using different methods, making it impossible to assess their relative explanatory power. The present study addresses this gap by operationalizing all three theories on a common corpus, enabling the first systematic comparison of how each illuminates the lived experiences of low-SES students.

Computer Science Approaches to Research on Low-SES Students: Contributions from Computer Science researchers in this area remain limited (Kelbessa et al., 2024). Earlier work by Lee et al. (2008) examined the educational development of low-SES students using data from the National Educational Longitudinal Study and applied categorical regression to analyze patterns in the data. Robinson (2020) investigated STEM intentions

among low-SES students at a high school in California, combining quantitative analysis via logistic regression with qualitative insights from in-depth interviews. Kelbessa et al. (2022) explored higher education contexts, identifying patterns in low-SES student experiences through analysis of social media data. Recently, Abdelgaber et al. (2025) developed a semi-automated dataset of narratives written by low-SES students collected from Reddit, which serves as the foundation for the current study.

TimeML Annotation and Timeline Extraction: TimeML is a widely used specification for annotating temporal information in text, designed to support machine-readable representations of when events happen and how they relate in time (Sauri et al., 2006). It represents events, time expressions (e.g., dates or durations), and temporal relations (e.g., BEFORE/AFTER/INCLUDES), enabling document-level temporal graphs for tasks like narrative understanding, temporal question answering, and timeline construction (Liu et al., 2012; Saquete et al., 2004; Navarro-Colorado and Saquete, 2016). Several automated TimeML pipelines (e.g., TARSQI, CAEVO, ClearTK, CATENA) extract events/times and predict temporal links (Verhagen et al., 2005; Chambers et al., 2014; Bethard, 2013; Mirza and Tonelli, 2016), but their outputs can be structurally flawed. Ocal et al. (2022) report that about one-third of automatically generated TimeML graphs are temporally inconsistent, and they contain roughly 20% more temporal indeterminacy than gold annotations, hurting downstream timeline construction. More recently, large language models have improved temporal extraction, with evidence that recent GPT models can better capture implicit temporal cues than traditional systems (Eirew et al., 2025; Wang and Weiss, 2025). Given TimeML annotations, a natural next step is timeline extraction: converting a set of temporally constrained events into an ordered event sequence that better reflects narrative progression. This conversion is non-trivial because TimeML graphs may be incomplete, underspecified, or inconsistent, and therefore require principled inference and constraint handling. The TLEX algorithm addresses this problem by first converting TimeML annotations into a temporal graph, and then transforming that representation into a temporal constraint graph. Timeline construction is performed by solving the resulting constraints using a constraint satisfaction problem (CSP) solver (Ocal et al., 2026).

3 Data and Theories

Data: We used the corpus introduced by Abdelgaber et al. (2025), consisting of Reddit narratives written by students from low-SES backgrounds. The dataset captures students' lived experiences in higher education and was created using a hybrid approach combining machine learning and human validation, expanding a small seed set into a larger, high-quality corpus.

The dataset consisted of English Reddit posts collected from U.S.-centric subreddits related to low-SES and student experiences, including communities such as r/college, r/ApplyingToCollege, r/financialaid, and r/fafsa. In the data collection process, strict inclusion criteria ensured relevance: narratives had to describe low-SES experiences, challenges faced, and responses to those challenges. General commentary or advice without personal experience was excluded. Since Reddit users are anonymous, no personally identifiable metadata about the authors was available. Low-SES status was identified based on self-reported narratives containing phrases such as "I grew up poor", "My parents had no money", and "I received a Pell Grant". The corpus is publicly available at `https://github.com/NahedAbdelgaber/NLP-for-Social-Science/tree/main/CoNLL%202025`.

The semi-automated data collection pipeline began with 128 manually labeled posts and iteratively expanded through classifier-based candidate selection, dimensionality reduction and clustering to remove outliers, and human validation to retain accurate narratives. This process produced a final dataset of 387 narratives, which were further processed with LLaMA to extract structured information on background, challenges, and strategies. Sentiment analysis revealed an increase in neutral and negative posts in the expanded dataset, providing broader representation of student experiences.

Theories: We wanted to use a strong theoretical framework for this corpus to move beyond descriptive observations and generate analytically meaningful insights into the socioeconomic gap in higher education. To achieve this, our study is grounded in a synthesis of three complementary theoretical perspectives drawn from the social sciences, following the theoretical synthesis proposed by Khan et al. (2026). These frameworks provide conceptual structure for interpreting students' narratives and for systematically mapping lived experiences

onto established theoretical constructs. The Table 1 presents a concise overview of each theory and its role in guiding our analytical approach.

4 Methodology

Computational Coding Framework: To operationalize the three theoretical frameworks at scale, we developed a prompt-based coding system using GPT-4.1 (OpenAI, 2025a). The prompt defined all three frameworks with specific coding categories, provided concrete linguistic examples for each code, required JSON-formatted outputs with evidence citations, and instructed the model to code constructs as absent when they were not explicitly discussed rather than inferring deficits.

The prompting keywords used for computational coding were derived directly from the theoretical constructs operationalized in this study rather than selected post hoc to improve model output. For Critical Pedagogy, the prompt emphasized terms related to structural barriers, oppression, agency, resistance, voice, and critical consciousness, reflecting Freire's framing of education as a process through which marginalized learners identify and challenge oppressive conditions (Freire, 1970; Riley, 2008; Secules et al., 2021). For Ecological Systems Theory, keywords such as microsystem, mesosystem, exosystem, macrosystem, family, work, institutional policy, financial aid, and cultural expectation were used to locate student experiences across nested social and institutional contexts, consistent with Bronfenbrenner's ecological model (Bronfenbrenner, 1979; Kahu & Picton, 2022). For Bourdieu's Theory of Capital and Habitus, the prompt included economic capital, cultural capital, social capital, symbolic capital, belonging, recognition, hidden curriculum, professional networks, and institutional familiarity to identify how socioeconomic disadvantage operates through unequal access to material, cultural, relational, and symbolic resources (Bourdieu, 1986; Yosso, 2005; Martin et al., 2013). Thus, the keywords functioned as theory-grounded coding anchors that translated established educational concepts into computationally identifiable narrative features. They were not designed to force clustering, but to ensure that the language model applied each framework consistently across the corpus.

Because the prompt required the model to assign narratives to predefined theoretical categories rather than generate categories inductively, visualizations based on these labels should be interpreted as prompt-guided separations of coded constructs rather than as fully unsupervised discoveries of naturally occurring clusters. GPT-4.1 was selected for the theory-based coding task because the task required consistent, structured classification across the full corpus rather than open-ended temporal reasoning. In preliminary testing, GPT-4.1 produced reliable JSON-formatted outputs for the predefined theoretical categories while allowing the full set of 387 narratives to be processed efficiently.

For Critical Pedagogy, we coded student agency (high/moderate/low), structural barriers (present/absent), and consciousness of barriers (high/moderate/low). For Ecological Systems Theory, we identified the dominant system level (microsystem, mesosystem, exosystem, macrosystem, chronosystem) and secondary influences. For Bourdieu's Theory, we classified four forms of capital, economic, social, cultural, and symbolic as deficit, asset, or absent. All 387 narratives were successfully processed with 0 errors. Results were exported as structured JSON and parsed into CSV format with full narrative text preserved for validation purposes. The methodology produced a dataset of 387 multi dimensional theoretical profiles, enabling comparative analysis across frameworks.

Validation and Data Quality: The improved prompt design explicitly distinguishes between narratives that discuss a construct as a deficit versus narratives that do not discuss it at all (absent). This refinement, validated on a sample of 51 narratives before full run execution, improved the model's ability to reflect actual narrative content without over interpreting missing information. To support reproducibility, the full analysis pipeline, prompts, and data processing scripts used in this study are publicly available at `https://github.com/Jadonsw123/low-ses-narrative-analysis`.

Applying TLEX: We applied TLEX to construct explicit temporal structure from students' posts by combining LLM-based TimeML annotation with constraint-based timeline extraction. Specifically, we used few-shot prompting with GPT-5.2 Thinking (OpenAI, 2025b) to produce TimeML-style annotations for each post. GPT-5.2 Thinking was used for the TimeML annotation task because temporal extraction requires more complex reasoning over implicit event ordering, temporal relations, and narrative sequence. For the few-shot exemplars, we selected fully TimeML-annotated documents from the AQUAINT corpus (Consortium

Theory	Core Idea	Function in This Work	Relevance to Low-SES Student Narratives
Critical Pedagogy (Freire)	Education is a political act that can reproduce or challenge inequality.	Frames the study as an act of empowerment amplifying student voices rather than diagnosing deficits.	Positions low-SES students as agents whose narratives expose structural barriers in education.
Ecological Systems Theory (Bronfenbrenner)	Student experiences are shaped by nested systems (micro → macro).	Provides a multi-layered lens to interpret narrative data across personal, institutional, and societal contexts.	Helps map each narrative theme (family, work, policy) to specific systemic layers influencing student outcomes.
Bourdieu's Theory of Capital & Habitus (Bourdieu)	Inequality persists through economic, cultural, and social "capitals" and internalized dispositions (habitus).	Explains how resource deficits cascade into behavioral, academic, and psychological outcomes.	Reveals how habitus mismatch and unrecognized "community cultural wealth" shape belonging and persistence.
Integrative Role		Combines Critical Pedagogy (why voices matter), Ecological Systems Theory (where challenges occur), and Capital Theory (how inequality operates).	Produces a unified interpretive framework connecting personal experience, systemic structure, and institutional culture.

Table 1: Theoretical Frameworks and Their Analytical Functions

et al., 2002), which provide high-quality demonstrations of how events, temporal expressions, and temporal links are encoded in standard TimeML notation. Given these examples, we prompted GPT-5.2 to annotate each post by identifying events, time expressions, and the temporal relations connecting them. In our experiments, GPT-5.2 generated annotations in a well-formed TimeML format, enabling direct downstream processing.

Next, we fed the resulting TimeML annotations into jTLEX (Ocal et al., 2023, the Java library for the TLEX algorithm). jTLEX first converts the annotations into temporal graphs and performs temporal consistency checking to identify violations in the induced temporal constraints. Finally, using the consistent constraint structure, jTLEX extracts within-post timelines by ordering events and time expressions into a coherent temporal sequence, which we use as the basis for analyzing narrative progression and identifying temporal trajectories in students' accounts.

5 Results

Critical Pedagogy: Agency and Consciousness: Two key dimensions emerge from the Critical Pedagogy framework: student agency and consciousness of structural barriers. As shown in Figure 1, while 71.6% of students demonstrate high consciousness of the systemic barriers they face, only 48.1% exhibit high agency in responding to those barriers. This 23.5 percentage point gap suggests a distinction between critical awareness and empowered action, echoing Freire's argument that reflection must be joined with action, or praxis, in struggles against oppression (Freire, 2020).

The clear separation in Figure 1 likely reflects the discrete coding structure imposed by the prompt, which asked the model to classify agency and consciousness into predefined levels. Therefore, the figure should be interpreted as a visualization of model-coded theoretical categories rather than evidence that the narratives naturally separate into distinct groups without theoretical guidance.

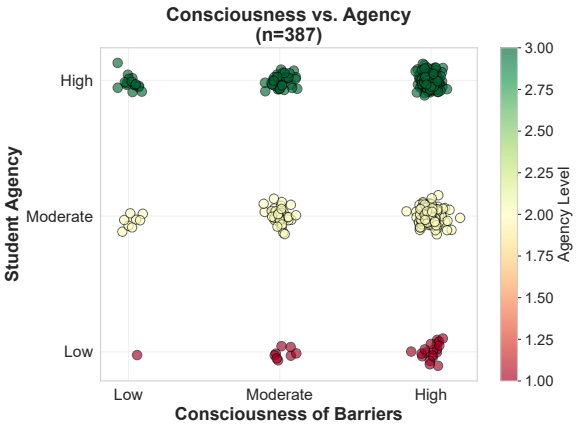


Figure 1: The Awareness Action Gap: Student consciousness of barriers (x-axis) versus student agency (y-axis) reveal that high consciousness is more prevalent than high agency, suggesting that awareness alone is insufficient for consistent empowered action. Each point represents one narrative (n=387).

Additionally, 92.5% of narratives explicitly mention structural barriers such as poverty, institutional policies, discrimination, or family instability. Moderate agency (44.7%) represents students who attempt to navigate barriers repeatedly but encounter constraints, while high agency (48.1%) reflects decisive, sustained action despite obstacles. Low agency (7.2%) appears in narratives expressing resignation or powerlessness. Figure 2 and Figure 3 show the distributions of agency and consciousness respectively.

Ecological Systems: Where Challenges Occur: Bronfenbrenner's framework reveals that stu-

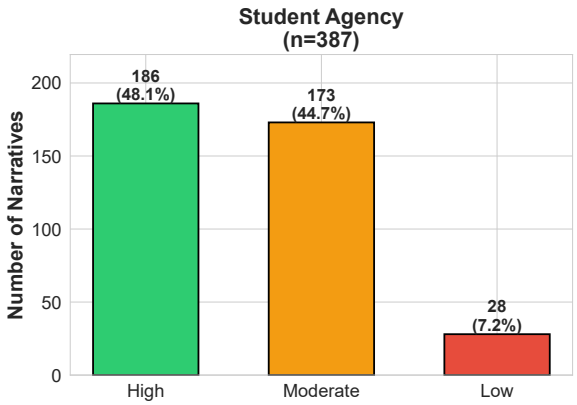


Figure 2: Agency Distribution: Nearly half of students (48.1%) demonstrate high agency, with 44.7% moderate and only 7.2% low, indicating that most low-SES students actively work to overcome barriers despite structural constraints.

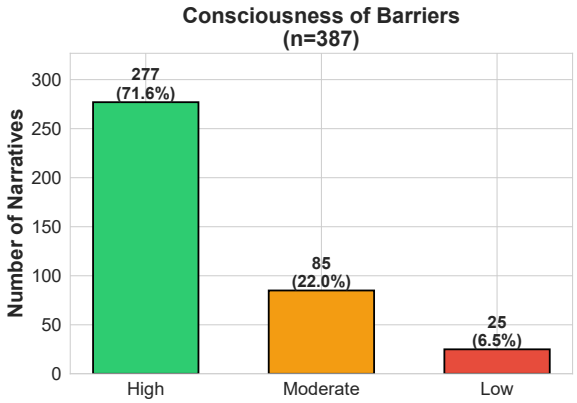


Figure 3: Consciousness of Barriers: The majority of students (71.6%) show high consciousness of systemic barriers affecting their education, aligned with Freire's concept of 'conscientization,' or the awakening to structural injustice.

dents' experiences are dominated by their immediate environments. As shown in Figure 4, the microsystem (family, school, peer relationships, direct work environment) is the dominant influence for 62.8% of narratives, followed by the exosystem (institutional policies, parental employment, financial aid systems) at 24.0%, and macrosystem (societal ideologies, laws, class structures) at 12.9%.

Bourdieu's Capital: Resources and Inequality: The most striking finding emerges from Bourdieu's theory: the distribution of capital types reveals not just deficits, but systematic absence of discussion around certain forms of capital. Figure 6 shows the distribution of students across different combinations of the following capitals.

Economic Capital: 85.8% deficit (as expected for low-SES students), 8.3% asset, 5.9% absent. Few narratives omit money concerns.

Social Capital: Highly polarized with 42.1% absent, 32.6% asset, 25.3% deficit. Many students either describe robust social support or do not mention relationships at all.

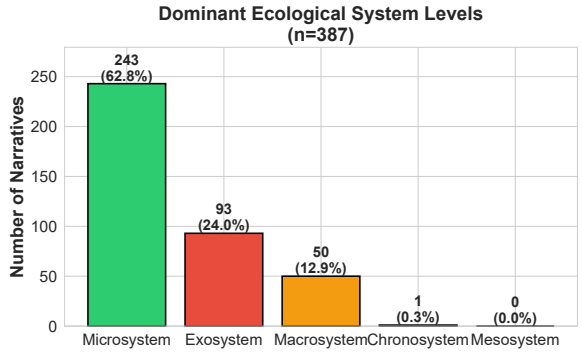


Figure 4: Ecological System Levels: Microsystems dominate (62.8%), indicating that intimate relationships and immediate contexts matter most. However, exosystem factors (policies, indirect institutional influences) affect nearly a quarter of narratives, demonstrating that structural barriers beyond the student's direct control are significant.

Cultural Capital: Nearly split with 46.0% asset (students discuss education, credentials, knowledge), 42.1% absent, 11.9% deficit. This suggests that while many low-SES students build educational capital, others' narratives remain silent on this dimension.

Symbolic Capital: The most absent 48.6%, indicating that students rarely narrate their status, recognition, or sense of belonging. When discussed, 32.3% frame it as deficit (shame, imposter syndrome) and only 19.1% as asset (pride, recognition).

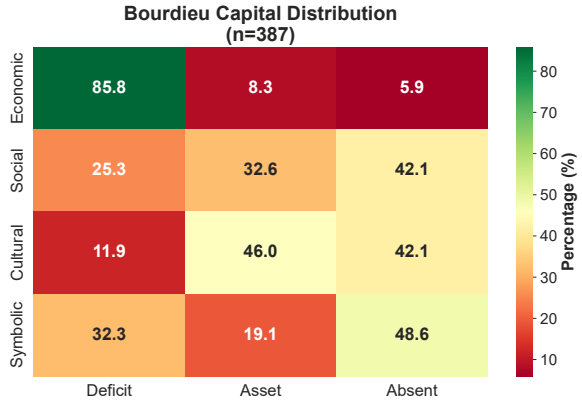


Figure 5: Bourdieu's Capital Distribution (n=387): Economic capital shows near universal deficit (85.8%), confirming low-SES background. However, social capital, cultural capital, and symbolic capital are frequently absent from narratives (42.1%, 42.1%, and 48.6% respectively), indicating students either lack these resources or do not articulate them. Cultural capital is split nearly evenly between asset (46.0%) and absent (42.1%), suggesting education/credentials are either leveraged strategically or invisible in their narratives.

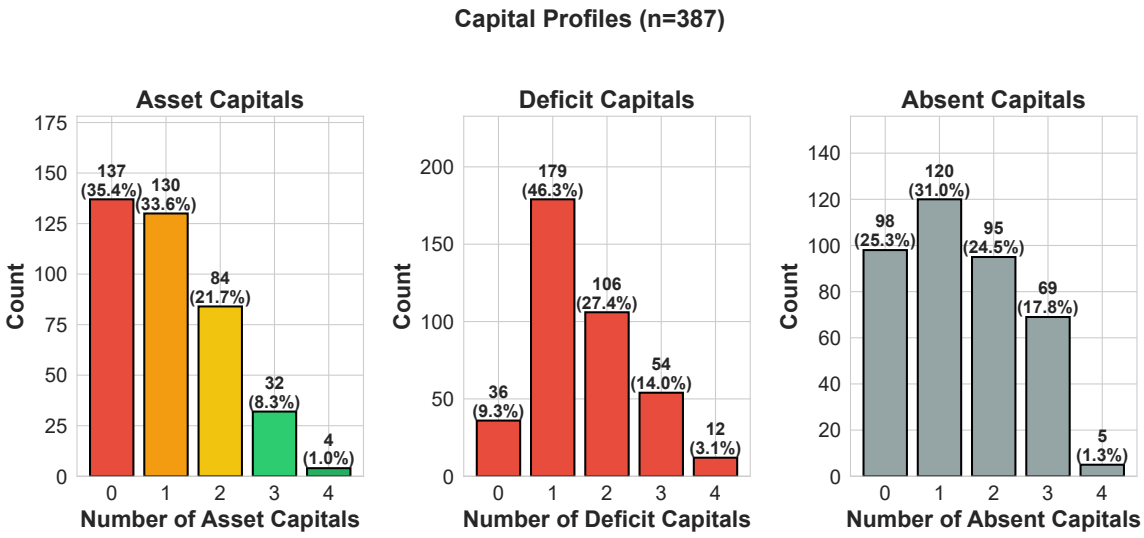


Figure 6: Capital Profile Diversity (n=387): The left panel shows asset distribution, the middle panel shows deficit distribution, and the right panel shows absence distribution across the four capital categories.

Narrative Timelines: In addition to producing narrative timelines, we observe a clear skew in how often writers narrate outcomes beyond schooling. Out of 387 extracted narrative timelines, only 21 include explicit discussion of the author's post-education life, suggesting that most posts remain focused on in-school challenges rather than longer-term resolution. Among those 21 post-education timelines, 11 describe a positive outcome. We labeled a timeline as positive when the writer explicitly reports outcomes such as obtaining a good job, advancing in their career, achieving financial stability (e.g., earning more than their parents, paying off loans), expressing happiness with where they ended up, or explicitly framing their struggles as "worth it." Several of the remaining timelines were mixed rather than clearly positive. For instance, authors describe doing better overall while still reporting ongoing debt, stress, or lingering hardship.

6 Discussion

The three theory framework reveals complementary insights: Critical Pedagogy exposes a consciousness-action gap where students are aware but constrained; Ecological Systems Theory pinpoints the microsystem as the primary site of intervention but acknowledges exosystem barriers; and Bourdieu's Theory reveals that symbolic capital (sense of belonging, recognition, and self worth) is the most invisible and potentially most damaging absence in low SES students' narratives.

The prevalence of absent capital categories suggests that students either genuinely lack those resources, or do not articulate or frame their experiences using those constructs. Each interpretation has implications: it points to resource scarcity and a potential mismatch between students' self narratives and institutional recognition frameworks.

This study used three influential educational theories on a common corpus of 387 low SES student narratives, enabling the first systematic comparison of their explanatory power. Our findings reveal that Bourdieu's Theory of Capital and Habitus provides the most comprehensive framework for understanding socioeconomic inequity in education. While Critical Pedagogy illuminates students' critical consciousness (71.6% high awareness of structural barriers) and Ecological Systems Theory maps the locations of struggle (62.8% microsystem dominance), Bourdieu's framework uniquely explains how disadvantage operates through multiple resource dimensions. The near universal economic capital deficit (85.8%) confirms material hardship, but more significantly, Bourdieu reveals the hidden dimensions of inequality: social capital polarized between asset (32.6%) and absence (42.1%), cultural capital nearly split between asset (46.0%) and absence (42.1%), and most strikingly, symbolic capital belonging and recognition either absent (48.6%) or framed as deficit (32.3%). These findings demonstrate that low SES students face not merely financial constraints but a multi dimensional capital deprivation that quantitative metrics alone cannot capture.

However, each theory also exhibits distinctive blind spots that justify a synthetic approach with

Bourdieu as foundation. Bourdieu's framework, with its emphasis on structural reproduction, risks portraying students as passive carriers of capital rather than active critics of unjust systems. Critical Pedagogy corrects this by revealing that students exercise remarkable agency despite constraints (48.1% high agency) and explicitly name structural barriers (92.5% of narratives). Yet Critical Pedagogy alone cannot specify what resources students lack or where interventions should target. Ecological Systems Theory supplies spatial precision, locating 62.8% of challenges in the microsystem (immediate environments) and 24.0% in the exosystem (policies and institutions), but cannot explain how problems operate within those contexts. Bourdieu supplies the missing mechanism: capital deficits cascade through ecological levels, transforming material scarcity into social isolation, cultural confusion, and psychological non-belonging.

The implications for education are clear. First, interventions must address multiple capitals simultaneously: financial support (economic capital), explicit teaching of hidden curricula (cultural capital), intentional network-building structures (social capital), and active affirmation of students' worth (symbolic capital). Second, the consciousness-action gap documented through Critical Pedagogy warns that raising awareness alone is insufficient, and students need tangible resources, not just recognition of injustice. Third, the ecological distribution of narratives reminds us that microsystem supports cannot compensate for exosystem barriers like financial aid policies or rigid scheduling. A synthesized theoretical framework Bourdieu as the diagnostic lens for what students lack, Freire for understanding how they critique their situation, and Bronfenbrenner for identifying where to intervene offers educators the most powerful toolkit for designing equitable, precisely targeted interventions.

Narrative Timelines: Qualitative Trajectories and Transition Points The extracted within-post timelines reveal how students narrate trajectories, not only in terms of events, but also in how experiences accumulate and shift over time. Even when posts contain few explicit time expressions, the GPT-5.2→TLEX pipeline often recovers coherent temporal ordering by leveraging implicit temporal cues (e.g., discourse structure, causal phrasing, and aspectual markers). This is especially valuable for educational-theory-grounded analysis because it helps operationalize transition points, making it easier to identify when barriers emerge, intensify, or begin to resolve. In addition, timelines frequently make mood/affect shifts visible across the narrative sequence: in one illustrative post, the author describes a difficult period during their college years and then contrasts it with their present circumstances, noting that 17 years later they are happy with their life. That turning point appears clearly in the timeline, with earlier events associated with hardship and later events reflecting a more positive state, highlighting narrative change in a way that aligns with the paper's emphasis on trajectories and theoretically meaningful moments for potential intervention.

7 Error Analysis and Future Work

The dataset for this study was drawn from prior work (Abdelgaber et al., 2025). Although the original data collection included human validation, some inconsistencies or low-quality narratives may remain. During exploratory analysis, we identified at least one narrative that only weakly matched the dataset's inclusion criteria. To maintain methodological consistency with the prior study, these data points were retained, though they may introduce minor noise. Future work will aim to expand the dataset with additional manually validated narratives and implement stricter filtering to ensure all posts are accurate.

In the theory experiments, GPT-4.1 enabled scalable analysis across the 387 narratives; however, the use of a large language model introduces potential error. Despite carefully designed prompts specifying theoretical constructs and required evidence, LLMs may still misinterpret ambiguous language or incorrectly map content. A notable example involves distinguishing between absent and deficit categories, which often depend on subtle contextual cues. Future work will incorporate human validation to improve reliability. Another improvement is to add an automated evaluation step for the labeling procedure. For example, repeated runs on the same narratives could be compared to assess label stability, and model-generated labels could be checked against a smaller human-coded validation set using agreement metrics such as Cohen's kappa. This would provide a more explicit measure of consistency in the theoretical coding process.

Our analysis primarily reports relative frequencies of theoretically coded categories. Future work

should move beyond descriptive distributions by examining correlations or associations among coded factors, such as the relationship between agency, consciousness of barriers, ecological system level, and different forms of capital. For example, additional analysis could test whether economic capital deficits co-occur with lower agency, whether symbolic capital deficits are associated with higher consciousness of barriers, or whether certain ecological system levels correspond to specific capital profiles. This would provide a stronger quantitative account of how the theoretical dimensions interact rather than treating each category separately.

A further limitation is that two GPT model versions were used for different subtasks. Although this choice reflected the different complexity and resource demands of the tasks, future work should include a controlled comparison in which both models are evaluated on the same subset of narratives. This would help determine whether model choice affects coding consistency, temporal annotation quality, or downstream findings.

Two primary limitations affect narrative timeline extraction. For cross-post ordering, future improvements include enhanced detection of temporal expressions and the use of these references as alignment points. When explicit anchors are unavailable, each post's platform timestamp can serve as a proxy to reconstruct a higher-level sequence, enabling longitudinal tracking of barriers and affect. To reduce within-post fragmentation, we will introduce a linking step that identifies bridging temporal cues such as connectives ("later", "after that", "by then"), causal indicators, and narrative shifts, generating candidate links between otherwise disconnected event chains. This approach is expected to yield more complete and interpretable timelines.

8 Limitation

First, the dataset is relatively small with 387 narratives and drawn from Reddit, which may bias the sample toward students comfortable sharing experiences publicly. Second, the study relies on large language models for theoretical coding. While effective at identifying patterns, these automated interpretations approximate rather than perfectly reproduce expert qualitative judgment. Third, narrative content is inherently selective. Students may emphasize some experiences and omit others, so the absence of a concept does not imply absence in reality. Finally, temporal timeline analysis has constraints. Cross-post timelines are difficult to construct due to missing or ambiguous dates, and some within-post timelines produce disconnected event chains, limiting the ability to fully capture sequential progression in a single narrative.

9 Contributions

We have four major contributions. First, this study takes an interdisciplinary approach, bridging Computer Science and Education by combining computational methods with established educational theories. In particular, it translates core constructs from Freire's Critical Pedagogy, Bronfenbrenner's Ecological Systems Theory, and Bourdieu's Theory of Capital and Habitus into computationally measurable features, enabling systematic analysis of low-SES student narratives. Second, the paper incorporates temporal timeline extraction, capturing ordered event sequences and showing how challenges and forms of capital evolve over time, linking theoretical categories such as barriers, supports, and resources to specific moments in students' experiences. Third, it provides a comparative assessment of the educational frameworks, evaluating how each theory explains different dimensions of students' lived experiences and responses. Finally, we made our code public for reproducibility.

10 Ethical and Societal Impact

This research has important ethical and societal implications. By analyzing narratives from low-SES students, it highlights systemic barriers and inequities in higher education, helping educators, policymakers, and institutions better understand the challenges faced by marginalized populations. Ethically, the study ensures privacy and anonymity by using only publicly available data, removing identifying information, and aggregating findings to prevent exposure of individual students. Moreover, the work contributes to equity-centered interventions by identifying patterns of barriers, supports, and forms of capital across students' educational experiences. By linking theoretical frameworks to practical, temporally structured insights, this research can inform targeted support strategies, mentorship programs, and policy initiatives aimed at reducing educational disparities. Additionally, the interdisciplinary approach demonstrates how computational methods can be ethically applied to socially relevant educational research.

References

Nahed Abdelgaber, Labiba Jahan, Arham Vinit Doshi, Rishi Suri, Hamza Reza Pavel, and Jia Zhang. 2025. Bridging the socioeconomic gap in education: A hybrid ai and human annotation approach. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 348–364.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second joint conference on lexical and computational semantics (* SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 10–14.

Pierre Bourdieu. 1986. The forms of capital. In John Richardson, editor, *Handbook of Theory and Research for the Sociology of Education*, pages 241–258. Greenwood Press.

Urie Bronfenbrenner. 1979. *The ecology of human development: Experiments by nature and design*. Harvard university press.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Linguistic Data Consortium and 1 others. 2002. The aquaint corpus of english news text.

Alon Eirew, Kfir Bar, and Ido Dagan. 2025. Beyond pairwise: Global zero-shot temporal graph generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31428–31446.

Engineering Education–Workforce Continuum Committee. 2019. *Understanding the educational and career pathways of engineers*. National Academies Press.

Paulo Freire. 2020. Pedagogy of the oppressed. In *Toward a sociology of education*, pages 374–386. Routledge.

Ella R Kahu and Catherine Picton. 2022. Using photo elicitation to understand first-year student experiences: Student metaphors of life, university and learning. *Active Learning in Higher Education*, 23(1):35–47.

Motti Kelbessa, Ilyas Jamil, and Labiba Jahan. 2022. Impacts of low socio-economic status on educational outcomes: A narrative based analysis. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 59–69.

Motti Kelbessa, Estephanos Jebessa, and Labiba Jahan. 2024. Addressing educational inequalities of low ses students: Leveraging natural language processing for impact. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '24, page 388–391, New York, NY, USA. Association for Computing Machinery.

Md. Tarique Hasan Khan, Labiba Jahan, and Mustafa Ocal. 2026. An inflection point in equity research: Leveraging ai and educational theory to decode the socioeconomic gap in engineering education. In *Proceedings of the ASEE Gulf-Southwest Annual Conference*. American Society for Engineering Education (ASEE).

Sang Min Lee, M Harry Daniels, Ana Puig, Rebecca A Newgent, and Suk Kyung Nam. 2008. A data-based model to predict postsecondary educational attainment of low-socioeconomic-status students. *Professional School Counseling*, 11(5):2156759X0801100504.

Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. 2012. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol.*, 3(2):25:1–25:28.

Julie P Martin, Denise R Simmons, and Shirley L Yu. 2013. The role of social capital in the experiences of hispanic women engineering majors. *Journal of Engineering Education*, 102(2):227–243.

Kerry L Meyers, Matthew W Ohland, Alice L Pawley, Stephen E Silliman, and Karl A Smith. 2012. Factors relating to engineering identity. *Global Journal of Engineering Education*, 14(1):119–131.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75.

Borja Navarro-Colorado and Estela Saquete. 2016. Cross-document event ordering through temporal, lexical and distributional knowledge. *Know.-Based Syst.*, 110(C):244–254.

Mustafa Ocal, Adrian Perez, Antonela Radas, and Mark Finlayson. 2022. Holistic evaluation of automatic timeml annotators. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 1444–1453.

Mustafa Ocal, Akul Singh, Jared Hummer, Antonela Radas, and Mark Finlayson. 2023. jtlex: a java library for timeline extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 27–34.

Mustafa Ocal, Ning Xie, and Mark A Finlayson. 2026. Tlex: an efficient method for extracting exact timelines from timeml temporal graphs. *Language Resources and Evaluation*, 60(1):10.

Matthew W Ohland, Catherine E Brawner, Michelle M Camacho, Richard A Layton, Russell A Long, Susan M Lord, and Mara H Wasburn. 2011. Race, gender, and measures of success in engineering education. *Journal of engineering education*, 100(2):225–252.

OpenAI. 2025a. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed 2026-05-08.

OpenAI. 2025b. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/. Accessed 2026-05-08.

Donna Riley. 2022. Engineering and social justice. In *Engineering and social justice*, pages 47–106. Springer.

Laura Robinson. 2020. The stem selfing process: Nondigital and digital determinants of aspirational stem futures. *American Behavioral Scientist*, 64(7):950–972.

Susan C Saegert, Nancy E Adler, Heather E Bullock, Ana Mari Cauce, William Ming Liu, and Karen F Wyche. 2006. Report of the apa task force on socioeconomic status. *Retrieved from the American Psychological Association website*.

E. Saquete, P. Martínez-Barco, R. Muñoz, and J. L. Vicedo. 2004. Splitting complex temporal questions for question answering systems. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Sauri, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines, version 1.2.1. https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf.

Stephen Secules, Cassandra McCall, Joel Alejandro Mejia, Chanel Beebe, Adam S Masters, Matilde L. Sánchez-Peña, and Martina Svyantek. 2021. Positionality practices and dimensions of impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education*, 110(1):19–43.

Marc Verhagen, Inderjeet Mani, Roser Saurí, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, John Phillips, and James Pustejovsky. 2005. Automating temporal annotation with tarsqi. In *ACL*, pages 81–84.

Jing Wang and Jeremy C Weiss. 2025. A large-language model framework for relative timeline extraction from pubmed case reports. *ArXiv*, pages arXiv–2504.

Tara J Yosso*. 2005. Whose culture has capital? a critical race theory discussion of community cultural wealth. *Race ethnicity and education*, 8(1):69–91.

A Appendix

Prompt Used for Theory-Based Coding: The full prompt used for GPT-based theoretical coding is included below to support reproducibility.

```
# GPT-4 Three-Theory Coding Prompt

You are an expert in educational theory and
     qualitative research coding. Your task is to
     classify a student narrative against three
     major theoretical frameworks.

## THEORETICAL FRAMEWORKS

### 1. CRITICAL PEDAGOGY (Freire)
Education is a political act that can reproduce or
     challenge inequality. Focus on how narratives
     reveal:
- **Student Agency**: Evidence of student voice,
     empowerment, or willingness to act
- **Structural Barriers**: Systemic obstacles,
     injustice, or inequitable conditions
- **Conscientization**: Student awareness of how
     social systems affect them
- **Liberation/Constraint**: Movement toward freedom
      or deeper constraint

**Coding Instructions**: Look for language showing
     consciousness of barriers, agentive responses,
     or expressions of empowerment/disempowerment.

---

### 2. ECOLOGICAL SYSTEMS THEORY (Bronfenbrenner)
Student experiences are shaped by nested systems
     influencing each other. Identify which level(s)
      most affect the student:
- **Microsystem**: Immediate environment (family,
     school, peer relationships, workplace)
- **Mesosystem**: Connections between microsystems (
     e.g., family-school relationships)
- **Exosystem**: Indirect influences (parents'
     workplace, educational policies, economic
     conditions)
- **Macrosystem**: Societal ideologies, values,
     culture, laws
- **Chronosystem**: Historical/temporal context (
     life transitions, major events)

**Coding Instructions**: Identify which level(s) the
      student emphasizes most. Most narratives will
     mention multiple levels-identify the 1-2
     DOMINANT levels.

---

### 3. BOURDIEU'S THEORY OF CAPITAL & HABITUS
Students accumulate four forms of capital that
     determine educational and social outcomes:
- **Economic Capital**: Financial resources, income,
      material assets, employment status
- **Social Capital**: Networks, relationships,
     institutional connections, support systems
- **Cultural Capital**: Educational credentials,
     knowledge, language skills, cultural practices
- **Symbolic Capital**: Status, prestige,
     recognition, respect, legitimacy in society

For each capital type, classify as:
- **DEFICIT**: Student explicitly mentions lacking,
     struggling with, or needing this capital
- **ASSET**: Student explicitly mentions having,
     building, or leveraging this capital
- **ABSENT**: Capital is NOT mentioned or discussed
     in the narrative at all

**Coding Instructions** (Use specific language/
     themes as evidence, not speculation):
- **Economic deficit** (explicit): "can't afford," "
     need to work," "financial stress," "money
     problems," "struggling financially"
- **Economic asset** (explicit): "got scholarship,"
     "parents support," "earning income," "financial
      stability"
- **Economic absent**: Narrative never discusses
     finances, money, work, or employment
- **Social deficit** (explicit): "no support," "
     isolated," "family problems," "no mentor," "
     alone," "don't have connections"
- **Social asset** (explicit): "family helped," "
     friends support," "mentor/professor helped," "
     community support"
```

```
- **Social absent**: Narrative never discusses
     relationships, support systems, or social
     networks
- **Cultural deficit** (explicit): "unprepared," "
     don't know," "no credentials," "behind," "
     underprepared"
- **Cultural asset** (explicit): "good student," "
     high GPA," "learning," "degree," "educated," "
     knowledge"
- **Cultural absent**: Narrative never discusses
     education, credentials, knowledge, or academic
     readiness
- **Symbolic deficit** (explicit): "shame," "not
     good enough," "imposter," "expected to fail," "
     invisible," "doubt"
- **Symbolic asset** (explicit): "proud," "overcame
     ," "resilient," "recognized," "valued," "
     confident"
- **Symbolic absent**: Narrative never discusses
     status, recognition, respect, or self-
     perception

---

## STUDENT NARRATIVE

{NARRATIVE_TEXT}

---

## REQUIRED OUTPUT FORMAT

Return ONLY valid JSON (no markdown, no extra text)
     with this exact structure:

```json
{
  "critical_pedagogy": {
    "student_agency": "high" | "moderate" | "low" |
         "absent",
    "structural_barriers_evident": true | false,
    "consciousness_of_barriers": "high" | "moderate"
         | "low",
    "summary": "Brief 1-2 sentence summary of
         student's agency/constraint",
    "quotes": ["quote1", "quote2"]
  },
  "ecological_systems": {
    "dominant_level": "microsystem" | "mesosystem" |
          "exosystem" | "macrosystem" | "
         chronosystem",
    "secondary_level": "microsystem" | "mesosystem"
         | "exosystem" | "macrosystem" | "
         chronosystem" | null,
    "microsystem_elements": ["element1", "element2
         "],
    "summary": "Brief explanation of dominant system
          level(s)"
  },
  "bourdieu_capital": {
    "economic": "deficit" | "asset" | "absent",
    "social": "deficit" | "asset" | "absent",
    "cultural": "deficit" | "asset" | "absent",
    "symbolic": "deficit" | "asset" | "absent",
    "dominant_capital_need": "economic" | "social" |
          "cultural" | "symbolic",
    "dominant_capital_strength": "economic" | "
         social" | "cultural" | "symbolic" | null,
    "summary": "Brief summary of capital profile"
  },
  "overall_summary": "2-3 sentence synthesis: How do
       all three theories help us understand this
      student's experience?"
}
```

INSTRUCTIONS FOR CODING

1. **Read the entire narrative carefully** before
     making classifications.
2. **Use evidence from the text** - Do NOT make
     assumptions or inferences beyond what is
     explicitly stated.
3. **For Bourdieu**: IMPORTANT - If a capital type
     is never mentioned or discussed, code it as "
     absent" rather than inferring deficit or asset.
      A narrative can have BOTH deficits and assets
     in the same capital type, OR it can omit the
     capital entirely.
4. **For Ecological Systems**: Most narratives will
     show multiple levels-pick the 1-2 that most
     dominate student's experience.
5. **For Critical Pedagogy**: Look for language of
     consciousness, agency, empowerment, or
     constraint.
6. **CRITICAL RULE**: Never force a classification
     when sufficient evidence is absent. Use "absent
     ," "low," or "null" rather than guessing.
7. **Be specific in summaries**: Don't just repeat
     definitions-show what THIS student's narrative
     reveals using direct language from the text.

VALIDITY CHECK

Before returning JSON:
- [ ] Does the JSON parse and validate?
- [ ] Have I cited evidence from the narrative?
- [ ] Am I using the student's own language/voice?
- [ ] Does the summary synthesize across theories?
```

Lost in Translation?
Exploring the Shift in Grammatical Gender from Latin to Occitan

Ahan Chatterjee[1,2], Matthias Schöffel[1,2],
Matthias Aßenmacher[2,3], Marinus Wiedner[4], Esteban Garces Arias[2,3]

[1]Bavarian Academy of Sciences (BAdW), Munich [2]LMU Munich
[3]Munich Center for Machine Learning (MCML) [4]University of Freiburg
Correspondence: ahan.chatterjee@badw.de

Abstract

The diachronic evolution from Latin to the Romance languages involved a restructuring of the grammatical gender system from a tripartite configuration (masculine, feminine, neuter) to a bipartite one (masculine, feminine) in most Romance languages. In this work, we introduce an interpretable deep learning framework to investigate this phenomenon at both lexical and contextual levels. First, we show that conventional tokenization strategies are insufficiently robust for this low-resource historical setting, and that our proposed tokenizer improves performance over these baselines. At the lexical level, we evaluate the contribution of morphological features to gender prediction. At the contextual level, we quantify the contributions of different part-of-speech categories to grammatical gender prediction. Together, these analyses characterize the distribution of gender information between the lemma and its sentential context. We make our codebase, datasets, and results publicly available at https://github.com/ahan-2000/Lost-in-Translation-.

1 Introduction

Despite substantial advances in natural language processing (NLP), contemporary research remains concentrated on fewer than two dozen of the nearly 7,000 languages spoken worldwide. The vast majority of historical and regional languages are categorized as low-resource languages, defined by data scarcity, minimal digital presence, and a lack of standardized resources (Singh, 2008). Medieval Occitan, a Romance language historically spoken in southern France, the Val d'Aran, and parts of Piedmont (cf. Figure 1), played an important role in medieval cultural and economic life all over Europe. Despite this, UNESCO currently classifies it as an endangered language (Mothe, 2024). Medieval Occitan presents many of the challenges typical of low-resource languages. In addition to severe data scarcity and the lack of annotated gold-standard

Figure 1: Historical spread of the Occitan language (Poujade et al., 2024).

resources, the language displays substantial instability: it shows extensive orthographic variation, with lexical items attested in multiple spellings both across and within texts (Garces Arias et al., 2023; Schöffel et al., 2025), as well as dialectal fragmentation stemming from the absence of a standardized norm (Zampieri et al., 2020). As a result, existing work consistently describes Occitan as a neglected low-resource Romance language with severe resource limitations (Woller et al., 2021).

As a direct descendant of Vulgar Latin (Pasquini and Serva, 2021), Occitan, too, underwent a transition from a tripartite to a bipartite gender system, as did most Romance languages. As Vulgar Latin evolved, morpho-phonological changes weakened the stable neuter category of Classical Latin, ultimately leading to the collapse and absorption of the neuters from the second declension class predominantly into the masculine gender (Szlovicsák, 2023). However, the specific factors that governed this reassignment in Occitan, whether semantic, phonological, or morphological, remain insufficiently studied, especially for nouns inherited from the third declension class (Marzo and Wiedner, 2025; Polinsky and Van Everbroeck, 2003). This work addresses this gap through a computational study of Medieval Occitan, examining how gram-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 276–296
July 4, 2026 ©2026 Association for Computational Linguistics

matical gender information is distributed between morphological features and morpho-syntactic context, and how these two sources contribute to model predictions for nouns descended from the Latin neuters. Methodologically, we propose a general framework for disentangling lemma-internal and contextual signals in grammatical gender prediction for low-resource historical languages. Our study is based on annotated corpora spanning law (*Lo Codi*), medicine (*Albuc*), and poetry (*Croisade*). Using these heterogeneous but sparse resources, we examine how morphological features and morpho-syntactic context jointly contribute to gender prediction for nouns descended from the Latin neuter class, through the following research questions:

RQ1 (Lexical-Level Analysis): To what extent can the grammatical gender of Occitan nouns derived from the Latin neuter class be predicted by word-level features, including their phonological and morphological characteristics?

RQ2 (Contextual Analysis): How is grammatical gender information distributed between morphological features and morpho-syntactic context in Occitan, and how do these sources contribute to model predictions?

2 Related Work

During the medieval period, the collapse of Latin neuter nouns often led to their absorption into the masculine gender in Romance languages, i.e., for neuters from the second declension class ending in *-um* (Klingebiel, 2019; Loporcaro, 2018). Although morpho-phonological cues provide strong signals (e.g., nouns in *-a* are typically feminine, while many others default to masculine), there are important exceptions, such as consonant-final gender-ambiguous nouns such as *mar* ('sea'), as well as Grecisms in *-a*, e.g., *propheta* ('prophet'). These irregularities suggest that accurate gender assignment may require additional information, including stress patterns, Latin etyma, and, of course, sentence context, given that gender is not a morphological but a morpho-syntactic category and that Old Occitan lacks an overt gender system.

One core research question is how effectively grammatical gender can be assigned to a noun solely on the basis of its form, including its lexical, phonological, and morphological characteristics. Early work by Brugmann (1897) emphasized the critical role of both phonological and semantic cues in gender assignment. However, these approaches are largely rule-based and language-specific, limiting their generalizability across diverse languages or linguistic families. Classic typological research, such as Corbett (1991), highlights that noun gender assignment typically involves a combination of morpho-phonological cues and semantic principles (e.g., natural gender for animates; but see the Greek loanwords as mentioned before). In Occitan, purely semantic gender applies in certain contexts, but for inanimate nouns, formal phonological and morphological cues predominate in gender determination and sometimes even supersede semantic criteria, e.g. the Greek loanwords.

Computational studies have attempted to quantify and predict gender from lexical features. Rule-based approaches to gender assignment have been extensively developed for languages such as French, producing long lists of endings and their most probable genders (Lyster, 2006). Nastase and Popescu (2009) analyze the prediction of grammatical gender using orthographic features and report that using a word's orthographic form in a statistical classifier improves gender prediction beyond baseline. These studies confirm that morpho-phonological cues have strong predictive power for gender. However, purely form-based prediction is not enough. Recent work by Williams et al. (2019) took an information-theoretic approach to languages such as German and Czech, measuring how much of gender assignment can be explained by a noun's form, meaning, or inflection class. They found that a combination of features provides the best predictions, highlighting that no single feature (orthography, phonology, semantics) accounts for everything, which is supported by recent experimental evidence (Basirat et al., 2021). Chronologically, the literature progressed from early descriptive grammars and implicit rules to manual rule compilations, then to data-driven classification, and now to neural and interpretable models. For Occitan, however, published computational work is still sparse. While some nouns have inherent grammatical gender, sentence context helps to identify gender assignment. In Occitan, as in related Romance languages like French and Spanish, determiners and adjectives agree in gender with nouns, participles, or pronouns. For example, the presence of the feminine article *la* before a noun signals that the noun is feminine in that context. Thus, the noun *torista* ('visitor') may be ambiguous in isolation, but in the phrase *la torista*, the

article disambiguates it as feminine in this context. While nouns predominantly have fixed grammatical gender, a few remnants, primarily from Latin neuter, exhibit atypical behaviour. Early computational work by Cucerzan and Yarowsky (2003) demonstrated that combining morphological analysis with contextual information significantly improves grammatical gender identification. Using a small annotated lexicon together with contextual cues such as co-occurrence with gendered articles and adjectives, their approach infers the gender of previously unseen words with high accuracy.

3 Data Description

The primary dataset for this study is drawn from three key Medieval Occitan sources. The first, *Lo Codi*, was annotated by Tobias Schmid as part of the ALMA Project (Heidelberg Academy of Sciences and Humanities, Bayerische Akademie der Wissenschaften, Academy of Sciences and Literature Mainz, 2025; Prifti et al., 2023). The second, the *Chanson de la Croisade Albigeoise*, was prepared and revised by Marinus Wiedner. The third source is the DOM Dictionary Project (Bayerische Akademie der Wissenschaften, 2025). The resulting annotated dataset comprises Latin–Occitan pairs, including Latin words, their corresponding Occitan lemmata, and the grammatical gender of each form, with a data distribution of 40.85% of unique lemmas from the *DOM* data source, 46.39% from *Lo Codi*, and 12.76% from *Croisade*. In addition, we use raw Occitan texts to analyze contextual cues (cf. Appendix A.2).

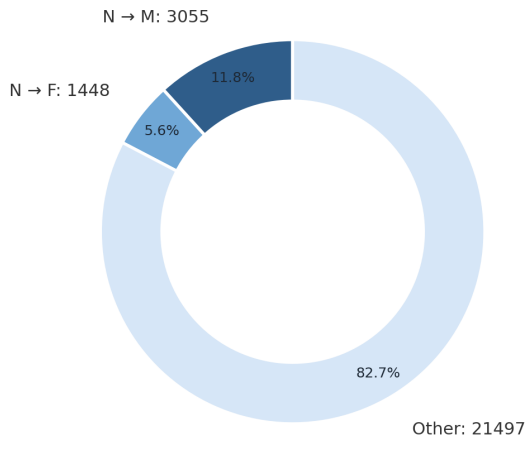


Figure 2: Gender Shift Frequencies across all three investigated corpora.

Our initial analysis confirms the complete absorption of the Latin neuter class into masculine and feminine genders in Occitan. As Figure 2 shows, the dominant shift is from neuter to masculine (3,055 cases), while a smaller but still substantial number of nouns shift to feminine (1,448 cases). A closer look at the orthographic features driving this divergence (Figure 7, Appendix A.1) reveals that specific endings are highly predictive of the outcome. The role of endings in this process is more nuanced than raw frequency counts suggest. While the ending *-um* is overwhelmingly associated with a masculine outcome, it is also, paradoxically, the single most common ending for nouns that become feminine. This is explained by the fact that the overall shift to masculine was far more prevalent, meaning any frequent neuter ending would appear dominant in that category. This finding underscores the importance of moving beyond simple frequency counts to understand the underlying mechanisms. By contrast, other endings, such as *-ia* and *-la*, provide a clearer signal and correlate strongly with feminine outcomes, further supporting the importance of morphological cues in gender (re)assignment.

4 Preliminary Analysis: Model and Tokenization Selection

We run a targeted set of probes to select (i) the embedding family that best captures Medieval Occitan variation in a Latin–Occitan setting and (ii) a tokenization policy that is robust to heavy orthographic noise. Concretely, we evaluate embedding models under three complementary criteria: (P_1) a frozen-encoder linear probe for Occitan gender prediction, (P_2) retrieval of Occitan orthographic variants given a Latin lemma, and (P_3) unsupervised structure in the embedding space via clustering.

4.1 Embedding Model Selection

We conduct a preliminary backbone selection study comparing FastText, mBERT, and ByT5 on three complementary probes: frozen gender prediction, Latin→Occitan variant retrieval, and clustering of Occitan forms. mBERT performs best across all three probes, suggesting that it provides the most reliable representation of lexical and cross-lingual structure for Medieval Occitan. We therefore adopt it as the embedding backbone in all downstream experiments (cf. Appendix B.1 for detailed results).

4.2 Tokenization Policy Selection

Medieval Occitan exhibits frequent spelling variation and sparse type coverage, making segmentation a primary bottleneck. We therefore evaluate tokenization policies via (i) OOV rate and (ii) masked token recovery accuracy on an Occitan masked language modeling (MLM)-style objective.

Example 1	**Example 2**
`de lay del primpcipat`	`En sa cambra secretament`
Hybrid: `de, la, y, del, pri, mp, ci, pat`	*Hybrid:* `En, sa, cambra, s, ec, ret, amen, t`

Figure 3: Examples of hybrid tokenization capturing orthographic and morphological variation in Medieval Occitan. In *primpcipat*, the subword `mp` isolates consonant-cluster variation, helping the model remain robust to spelling differences such as `nc`/`mp`. In *secretament*, the final `t` is segmented separately, reflecting a common Old Occitan alternation where the adverbial `-t` may be elided (e.g., *secretamen* vs. *secretament*). Such fine-grained segmentation supports better generalization across predictable historical variants.

Subword vs. Hybrid Segmentation. We evaluate tokenization policies using (i) **OOV rate**, defined as the proportion of tokens mapped to `[UNK]`, and (ii) **masked token recovery**, defined as top-1 accuracy at masked *subword* positions. In this experiment, a hybrid policy (Occitan-adapted BPE with a word-level fallback) preserves full coverage (zero `[UNK]`) and yields the best masked recovery (25.23%), indicating that explicit fallback coverage is crucial while still benefiting from corpus-adapted subword units (cf. Appendix B.2). Qualitatively, the hybrid tokenizer also produces more interpretable subword boundaries than generic WordPiece segmentation (Fig. 3).

4.3 Domain-Adaptive MLM Fine-Tuning

Given the consistent advantages of mBERT and hybrid tokenization, we apply domain-adaptive MLM fine-tuning for 10 epochs with identical hyperparameters across runs and evaluate on a held-out validation split. Fine-tuning substantially improves fit to the Occitan corpus: standard MLM adaptation reduces validation perplexity from 942.85 to 10.44, while the hybrid-vocabulary variant attains the best validation perplexity (9.52). Since perplexity is tokenization-dependent, we interpret these values as within-configuration diagnostics; taken together with the probing and tokenization results, they motivate our final setup: mBERT with a hybrid tokenizer and domain-adaptive MLM fine-tuning.

Based on the preliminary analyses, we adopt mBERT with hybrid tokenization and MLM adaptation as the backbone for all subsequent experiments. We now address our first research question by investigating grammatical gender prediction from lexical features alone, setting up a contrast with the contextual models introduced next.

5 Methodology

5.1 Lexical Grammatical Gender Prediction

Grammatical gender is a nominal classification system; in Occitan, it is bipartite (masculine and feminine) and is typically realized through noun morphology and agreement. In this section, we examine gender assignment based solely on lexical information, without relying on sentential context.

5.1.1 Feature Representation and Engineering

Data Normalization. We lowercase lemmas and apply Unicode NFKD normalization, stripping combining diacritics for character-level features; original forms are retained for the embeddings.

Task and Imbalance Handling. We predict *Occitan grammatical gender* as a bipartite label $y \in \{\mathrm{M}, \mathrm{F}\}$. Since outcomes are highly skewed (the Latin neuter most frequently maps to Occitan masculine), we use class-weighted training and focal loss; we also perform ablations to quantify the contribution of each feature group.

Morphological and Phonotactic Features. From both Latin and Occitan lemmas, we extract initial word substrings and suffix character n-grams ($1 \leq n \leq 4$), emphasizing word-final cues consistent with Romance gender marking (Table 1). We further encode syllabic shape using (i) vowel-run syllable count $S(w)$ and (ii) VC templates $P(w)$ (Table 2), and include length features $|w_{\text{lat}}|$, $|w_{\text{occ}}|$, their difference, and ratio.

Lang	**Lemma**	**Initial Substrings**	**Suffix**
Latin	*domus*	`do`	`us`
Occitan	*dom*	`do`	`om`

Table 1: Example of initial word substrings/suffix bigram extraction ($n = 2$) for aligned Latin–Occitan lemmas.

Stress as a Coarse Proxy. We include a lightweight stress-position proxy (ultimate/penultimate/antepenultimate), derived from a syllable-weight heuristic: monosyllables

Word (w)	**Language**	**Syllable Count** $S(w)$
festum	Latin	2
festa	Occitan	2
tempus	Latin	2
temps	Occitan	1
Word (w)	**Language**	**VC Pattern** $P(w)$
festum	Latin	CVCCVC
festa	Occitan	CVCCV
tempus	Latin	CVCCVC
temps	Occitan	CVCCC

Table 2: Syllabic structure features: vowel-run syllable count $S(w)$ and VC template $P(w)$.

are stressed on the only syllable; disyllables on the penult; for polysyllables, we stress the penult if it is heavy (long vowel or closed syllable), otherwise the antepenult. We treat this feature as an approximate cue rather than as a definitive phonological annotation.

Embedding Features We use frozen pretrained representations as *feature extractors* and compare them as alternative embedding feature sets rather than concatenating them: FastText (subword n-gram composition), mBERT, and ByT5. For mBERT/ByT5, each lemma is embedded in isolation and represented by mean pooling over subword/byte final-layer states; FastText uses standard word-type vectors. These embeddings are then used directly as input features to the downstream classifier.

5.1.2 Experimental Setup

We evaluate feature sets using lemma-grouped 10-fold cross-validation to prevent leakage across orthographic variants. Let $\mathcal{D} = \{(x_i, y_i, \ell_i)\}_{i=1}^{N}$, where x_i are features, $y_i \in \{\mathrm{M}, \mathrm{F}\}$ is the label, and ℓ_i is a lemma ID; folds are formed over lemmas and scores are averaged across folds. We evaluate a diverse set of classifiers to cover complementary inductive biases, ranging from transparent linear models (Logistic Regression), to non-linear tree ensembles (Random Forest, XGBoost), to sequence-aware neural architectures (FFN, BiLSTM, and attention-based variants). This design allows us to test whether grammatical gender is primarily recoverable from simple lexical cues or whether stronger performance requires models that capture higher-order or sequential interactions in the feature space. Hyperparameters are tuned with Optuna (Bayesian optimization), maximizing validation Macro-F1 within the cross-validation protocol.

Although lexical features are highly informative, they do not fully determine grammatical gender in all cases. For nouns such as *psalmista*, the intended gender may only become clear from sentence-level agreement cues, especially the article (*lo/la*). We therefore turn to our second research question, examining whether contextual information improves prediction beyond lemma-internal evidence alone.

5.2 Context-based Grammatical Gender Prediction

In the previous section, we examined the contribution of lexical features to gender prediction in isolation. Here, we study the contribution of *sentence-level context* as a second source. In Occitan, gender is jointly encoded by the noun and its agreeing dependents (articles, adjectives, and other modifiers); we exploit this distributed encoding as a prediction signal when lemma-internal cues are weak.

5.2.1 Dataset & Data Preparation

We use ∼130k tokens of unannotated Occitan texts spanning multiple genres (law, poetry, and medicine). We normalize the corpus by lowercasing, stripping diacritics, and standardizing punctuation. Because parallel Latin sentences are unavailable, we rely on an existing Occitan–Latin lemma lexicon and link each Occitan lemma (cf. Algorithm 1) occurrence to its containing sentence, yielding contextual instances for downstream analysis.

5.2.2 Proposed Methodology

We quantify the contribution of sentential context to Occitan gender prediction using three input settings. Each instance is (X, i, L, G_L, y), where $X = (x_1, \ldots, x_T)$ is an Occitan sentence, i indexes the target noun token $w = x_i$, L is its Latin lemma with Latin gender $G_L \in \{\mathrm{M}, \mathrm{F}, \mathrm{N}\}$, and $y \in \{\mathrm{M}, \mathrm{F}\}$ is the gold Occitan label. A pretrained encoder produces contextual states:

$$H = \mathrm{BERT}_\theta(X) = (h_1, \ldots, h_T), \qquad h_t \in \mathbb{R}^d. \tag{1}$$

All configurations share the same MLP head f_ϕ, with

$$p(y \mid r) = \mathrm{softmax}(f_\phi(r)). \tag{2}$$

(i) Word-only. We form a lexical representation from isolated embeddings and Latin metadata:

$$r_{\text{word}} = [\, e(w);\ e(L);\ \mathrm{onehot}(G_L) \,]. \tag{3}$$

Algorithm 1 Construction of Occitan–Latin Lemma–Gender Dataset

Require: Raw Occitan corpus D
Require: Occitan–Latin lemma lexicon $\mathcal{L}$
Require: Similarity function $\textsc{Sim}(\cdot,\cdot)$ (cf. C.1)
Ensure: Table T of aligned lemmas, contexts and genders
$D_{\text{pos}} \leftarrow \textsc{PoSTag}(D)$ ▷ tag every token in the corpus (cf. E)
$N \leftarrow \{(w, s, \ell_{\text{oc}}) \in D_{\text{pos}} : \text{PoS}(w) = \text{NOUN}\}$ ▷ collect noun tokens with sentence s and lemma ℓ_{oc}
$T \leftarrow \emptyset$
for all $(w, s, \ell_{\text{oc}}) \in N$ **do** ▷ iterate over all noun instances
  if $\exists(\ell_{\text{oc}}, \ell_{\text{la}}, g_{\text{oc}}, g_{\text{la}}) \in \mathcal{L}$ **then**
    $(\hat{\ell}_{\text{oc}}, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}}) \leftarrow (\ell_{\text{oc}}, \ell_{\text{la}}, g_{\text{oc}}, g_{\text{la}})$ ▷ exact lemma match
  else
    Find $(\ell', \ell'_{\text{la}}, g'_{\text{oc}}, g'_{\text{la}}) \in \mathcal{L}$ s.t. $\ell' = \arg\max_{\tilde{\ell} \in \mathcal{L}} \textsc{Sim}(\ell_{\text{oc}}, \tilde{\ell})$ and $\textsc{Sim}(\ell_{\text{oc}}, \ell') \geq \tau$ (τ=0.85).
    if a candidate exists **then**
      $(\hat{\ell}_{\text{oc}}, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}}) \leftarrow (\ell', \ell'_{\text{la}}, g'_{\text{oc}}, g'_{\text{la}})$ ▷ fuzzy lemma match
    else
      continue ▷ skip if no reliable match is found
    end if
  end if
  Append row $(\hat{\ell}_{\text{oc}}, s, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}})$ to T ▷ store Occitan lemma, context, Latin lemma, and both genders
end for
return T

(ii) Context-focused. To target the noun within its sentence, we use noun-conditioned attention over H (cf. Architecture in Figure 4):

$$r_{\text{ctx}} = [\,\text{Attn}(h_i, H, H);\ e(L);\ \text{onehot}(G_L)\,]. \tag{4}$$

(iii) Masked-context. To isolate contextual cues, we mask the noun $x_i \leftarrow [\text{MASK}]$, re-encode $H^{\text{mask}} = \text{BERT}_\theta(X^{\text{mask}})$, and use the state at position i:

$$r_{\text{mask}} = [\,h_i^{\text{mask}};\ e(L);\ \text{onehot}(G_L)\,]. \tag{5}$$

The masked-context setting evaluates how much of a noun's gender can be recovered from the surrounding sentence alone. Because Occitan articles, adjectives, and other dependents inflect to agree with the noun, the surrounding sentence jointly encodes the noun's gender; we therefore read masked-context performance as a measure of this distributed encoding rather than as an independent contextual signal. Comparing the word-only, context-focused, and masked-context configurations lets us bound how much predictive signal each source contributes (cf. Algorithm 2).

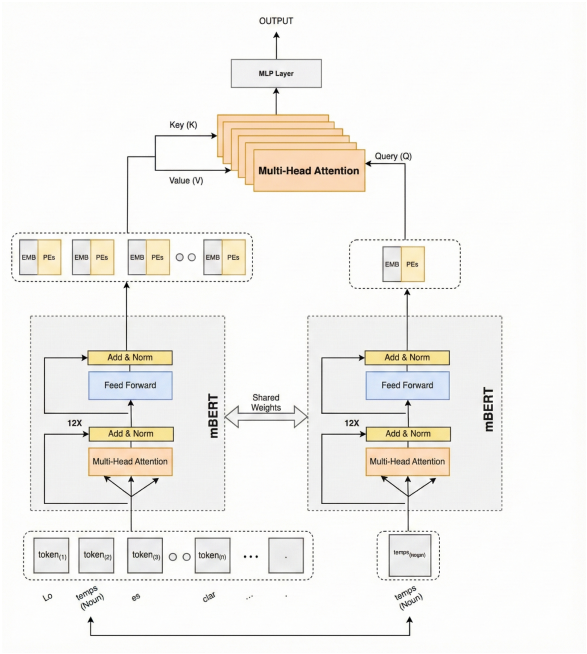


Figure 4: Proposed architecture to assess the impact of contextual cues on nouns' grammatical gender prediction.

All experiments use 3-group k-fold cross-validation, preserving the class distribution of Occitan gender labels across splits. To prevent label leakage, splits are constructed at the lemma level so that orthographic variants of the same lemma do not appear in both training and validation folds. A fixed random seed (13) is used throughout for reproducibility. We further analyze which contextual categories drive predictions by aggregating token-level contributions by PoS tag, e.g., determiners, adjectives, and verbs (cf. Appendix E), yielding tag-wise estimates of their influence on gender prediction.

6 Results and Discussion

6.1 Lemma-level gender prediction

Table 3 reports mean Accuracy and Macro-F1 (10-fold CV) for lemma-level gender prediction across model families and embedding feature sets. Overall, neural sequence models outperform shallow baselines, and attention generally yields further gains. The best Macro-F1 is achieved with a 2× BiLSTM + multi-head self-attention (MHSA) model trained with imbalance-aware objectives (focal loss/class weighting), yielding the strongest results with multilingual encoders (mBERT/ByT5), with pretrained representations providing more informative lexical cues than static embeddings.

FastText (baseline = 0.7734)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.7606	0.0128	1.66%
Meta-features	0.7640	0.0094	1.22%
Occitan n-grams	0.7667	0.0067	0.87%
Syllable counts	0.7667	0.0067	0.87%
VC patterns	0.7710	0.0024	0.31%
Stress patterns	0.7746	-0.0012	-0.16%
mBERT (baseline = 0.8224)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.8092	0.0132	1.61%
Meta-features	0.8168	0.0056	0.68%
Occitan n-grams	0.8169	0.0055	0.67%
Syllable counts	0.8194	0.0030	0.37%
VC patterns	0.8220	0.0004	0.05%
Stress patterns	0.8239	-0.0015	-0.18%
ByT5 (baseline = 0.8106)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.7958	0.0148	1.83%
Meta-features	0.8006	0.0100	1.23%
Occitan n-grams	0.8035	0.0071	0.88%
Syllable counts	0.8087	0.0019	0.23%
VC patterns	0.8091	0.0015	0.19%
Stress patterns	0.8123	-0.0017	-0.21%

Table 4: Feature ablation comparison across FastText, mBERT, and ByT5.

6.2 Feature ablation

To quantify the contribution of each feature group, Table 4 removes one block at a time from the best configuration (per embedding set) and reports the resulting Macro-F1 drop. Latin and Occitan character n-grams, especially suffix cues, are the most influential, producing the largest decreases (1.6–1.8 Macro-F1 points). Length/meta-features are the next strongest contributors (0.7–1.3 points), while VC templates and stress proxies have comparatively smaller effects.

6.3 Feature attributions (SHAP)

SHAP attributions (cf. Figure 5) are broadly consistent with the ablation results: suffix features and length-related meta-features dominate the decision signal across embedding modalities. Stress-related cues occasionally receive non-trivial attribution; however, since stress is derived from a heuristic proxy, we interpret these effects cautiously.

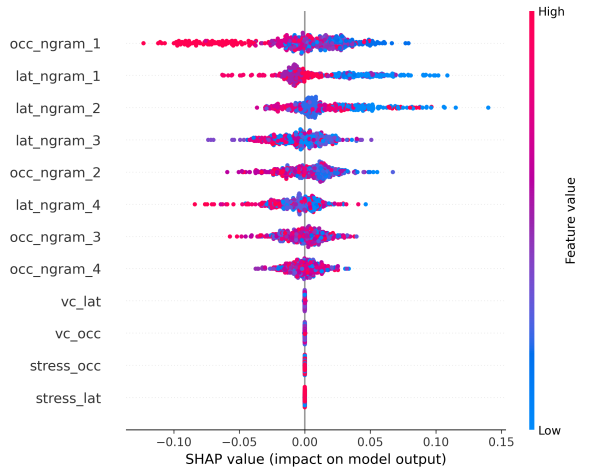


Figure 5: SHAP summary plot for the best-performing lemma-level model.

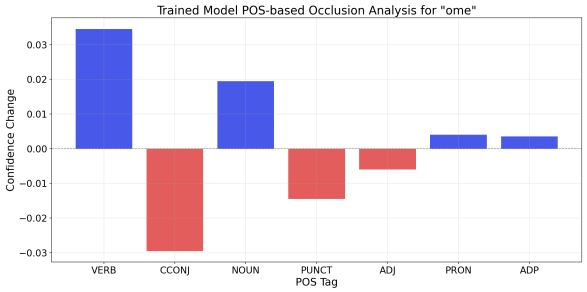


Figure 6: Example in which the lemma-only model misclassifies *ome* as feminine: *"aquill ome qui tenunt uera fe e pois tornunt en heresia deuent auer atrestal pena cum li altre e tant maior quant maior peccat ill fant."* With sentence-level context, the prediction shifts to masculine, with attribution distributed across agreement-bearing tokens such as *aquill*, illustrating how gender information distributed between the lemma and its local context lets the contextual model recover the correct label when the lemma representation alone is insufficient.

6.4 Impact of Contextual Cues

We evaluate contextual induction following Algorithm 2. Table 5 compares three mBERT-based configurations. Adding sentence context yields a substantial gain over the word-only baseline (Macro-F1: 0.665 → 0.929). Masking the noun remains substantially better than word-only (Macro-F1 0.902), but underperforms the unmasked setting, consistent with the noun form carrying most gender signal while context provides additional disambiguation. In cases where the lemma representation alone is insufficient, the contextual model produces a different prediction than the lemma-only model, as illustrated in Figure 6.

Model	Accuracy	Macro F1
ByT5		
Logistic Regression	0.6418 ± 0.0372	0.6139 ± 0.0440
Random Forest	0.7333 ± 0.0265	0.6974 ± 0.0333
XGBoost	0.7141 ± 0.0364	0.7354 ± 0.0492
BiLSTM	0.7550 ± 0.0410	0.7427 ± 0.0397
2×BiLSTM+Attn. (CE, 100 ep)	0.7902 ± 0.0389	0.7867 ± 0.0354
2×BiLSTM+MHSA (CE, LS=0.1, 100 ep)	0.8380 ± 0.0249	0.8106 ± 0.0237
FastText		
Logistic Regression	0.6578 ± 0.0384	0.6319 ± 0.0459
Random Forest	0.7167 ± 0.0392	0.6917 ± 0.0397
XGBoost	0.7152 ± 0.0524	0.7068 ± 0.0528
Feedforward NN (FFN)	0.7266 ± 0.0231	0.7113 ± 0.0283
BiLSTM	0.7763 ± 0.0323	0.7561 ± 0.0293
BiLSTM + Attn. (CE, 100 ep)	0.7480 ± 0.0492	0.7224 ± 0.0536
2×BiLSTM+Attn. (CE+ES, 100 ep)	0.7904 ± 0.0349	0.7512 ± 0.0354
2×BiLSTM+MHSA (CE+LS=0.1, 50 ep)	0.8134 ± 0.0416	0.7734 ± 0.0344
mBERT		
Logistic Regression	0.6749 ± 0.0254	0.6572 ± 0.0273
Random Forest	0.7287 ± 0.0397	0.7097 ± 0.0441
XGBoost	0.7352 ± 0.0402	0.7168 ± 0.0425
BiLSTM	0.7525 ± 0.0355	0.7252 ± 0.0384
BiLSTM+Attn. (CE, 100 ep)	0.7846 ± 0.0272	0.7419 ± 0.0309
2×BiLSTM+Attn. (FL + Class Wts, 50 ep)	0.7840 ± 0.0369	0.7460 ± 0.0341
2×**BiLSTM + MHSA (CE, LS=0.1, 100 ep)**	$\mathbf{0.8327 \pm 0.0365}$	$\mathbf{0.8224 \pm 0.0385}$

Table 3: Mean ± SD over 10-fold lemma-grouped cross-validation for lemma-level grammatical gender prediction. Best per-embedding rows are bolded. Under the shared 2×BiLSTM+MHSA head, the mBERT advantage over ByT5 is significant at the instance level on pooled out-of-fold predictions (paired bootstrap, Δ Macro-F1 $= +0.0395$, 95% CI $[+0.0250, +0.0543]$, $p < 10^{-6}$; full procedure in Appendix F.1).

Experiment (mBERT)	Acc.	Macro F_1
Word-only Model	0.808 ± 0.154	0.665 ± 0.108
Context model (explicit noun attention)	$\mathbf{0.979 \pm 0.012}$	$\mathbf{0.929 \pm 0.034}$
Context model (noun masked)	0.977 ± 0.008	0.902 ± 0.097

Table 5: Classification performance across the three experimental settings. The word-only baseline is lower than in the lemma-level experiments because, for comparability with the contextual models, it uses only the Latin lemma, Occitan lemma, and Latin gender, without the richer lemma-level feature set introduced earlier.

Δ Statistic	Mean	95% CI
Δ_1^{prob}	0.283	[0.281, 0.285]
Δ_2^{prob}	0.279	[0.277, 0.281]
$\Delta_1^{\log p}$	0.294	[0.285, 0.303]
$\Delta_2^{\log p}$	0.340	[0.334, 0.346]

Table 6: Mean values and 95% confidence intervals for the Δ statistics.

To examine whether context increases confidence in the correct label, we report mean probability and log-probability deltas for the gold class (Table 6). Both Δ_1 (context vs. word-only) and Δ_2 (masked-context vs. word-only) are positive, indicating that contextual cues systematically raise the model's confidence in the ground-truth class.

6.5 Model Explainability

For the context model, we use 8-head attention with the target noun state as query and sentence states as keys/values. Figure 8 in Appendix G illustrates that attention concentrates on the noun token, with the associated article typically receiving the next-highest mass, matching Occitan morpho-syntax where articles (e.g., *lo*/*la*) are strong gender cues. Across heads, attention mass is broadly distributed, with no single head consistently specializing in a particular Part-of-Speech (PoS) category.

To quantify which contextual categories contribute most, we run PoS-conditioned occlusion (cf. Algorithm 3 in Appendix C.3) and aggregate token-level deltas by tag. Table 7 shows that nouns contribute the largest positive delta, followed by determiners and adjectives, consistent with gender information being distributed across the noun and its agreeing dependents.

PoS tag	Mean Δ	Count (n)	Sign-flip p
NOUN	$+0.0026$	39,521	$< 10^{-4}$
DET	$+0.0010$	24,042	$< 10^{-4}$
ADJ	$+0.0003$	29,920	$< 10^{-4}$
CCONJ	-0.0010	27,492	$< 10^{-4}$
ADP	-0.0007	26,577	$< 10^{-4}$
VERB	-0.0003	29,336	$< 10^{-4}$
PUNCT	-0.0001	30,408	0.096
PRON	-0.0002	23,893	0.997

Table 7: PoS-wise mean occlusion deltas with sign-flip permutation test (10,000 permutations, two-sided). NOUN, DET, and ADJ contribute reliably positive contextual evidence; CCONJ, ADP, and VERB contribute reliably negative effects; PUNCT and PRON are not significant. Effect magnitudes are small in absolute terms; we read them as stable but modest contextual cues. Full procedure in Appendix F.2.

7 Conclusion

Gender information in Medieval Occitan is distributed across two sources: lemma-internal morphology and sentence-level context. Suffix morphology carries the strongest single signal; articles, adjectives, and other agreeing dependents provide additional morpho-syntactic cues that may inform gender assignment in context-sensitive interpretation, and when the lemma alone is ambiguous, they can shift a model's prediction. Taken together, these findings support a two-layer view of gender in Medieval Occitan: lexical morphology provides the primary structural encoding, while agreement and contextual patterns, that is, morpho-syntactic cues, reflect its realization in usage. Methodologically, the work highlights that historical orthographic instability makes standard tokenization brittle; hybrid tokenization with domain-adaptive MLM enables models to exploit meaningful subword regularities while remaining robust to spelling variation. More broadly, the proposed lexical-versus-contextual comparisons and attribution analyses offer a useful framework for studying grammatical change in noisy historical corpora, though future work with richer gold annotation and improved morpho-syntactic resources would allow for more fine-grained analyses.

Limitations

It is important to acknowledge several limitations. First, while our corpus is genre-diverse, it remains relatively small and label-imbalanced (approximately 2:1 masculine-to-feminine), which may limit minority-class generalization despite mitigation via focal loss and class-weighted training. Second, key components of the preprocessing pipeline were set heuristically, most notably the fuzzy-matching threshold ($\tau = 0.85$) and the stress-position proxy, and our ablations suggest that the stress feature can introduce mild noise. Third, our PoS-conditioned analyses (cf. Appendix E) rely on automatic tagging, and our evaluation shows $\sim$71% tagging accuracy, implying that PoS-based attribution results may be biased by tagging errors. Finally, the contextual model is less reliable in sentences where the target noun occurs at sentence boundaries or where agreement-bearing cues (cf. Appendix I) are sparse, which motivates future work on boundary-aware modeling and richer syntactic supervision. More broadly, our conclusions pertain to the Latin-to-Occitan neuter collapse and should be tested across additional Medieval Romance varieties. Our experiments quantify how gender information is *distributed* between lexical and contextual sources for synchronic prediction; they do not directly test the diachronic question of what drove the historical reassignment of former Latin neuters, which requires parallel diachronic data and a different experimental design.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work uses publicly available datasets and involves no human subjects or personally identifiable information. All data and code, including preprocessing, modeling choices, and evaluation protocols, are released to enable reproducible research and further investigation. Our work is intended exclusively for research purposes, and we encourage careful interpretation of results, particularly in low-resource and historical language settings where annotation uncertainty and data scarcity are common.

Acknowledgments

Esteban Garces Arias sincerely thanks the Mentoring Program of the Faculty of Mathematics, Statistics, and Informatics at LMU Munich and the Munich Center for Machine Learning (MCML) for their ongoing support. Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581 - BERD@NFDI.

References

Ali Basirat, Marc Allassonnière-Tang, and Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*, 7(1):20200048.

Bayerische Akademie der Wissenschaften. 2025. Dictionnaire de l'occitan médiéval (dom). https://dom.badw.de/fr/le-projet.html. Accessed: 25 November 2025.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Karl Brugmann. 1897. *The nature and origin of the noun genders in the Indo-European languages: A lecture delivered on the occasion of the sesquicentennial celebration of Princeton University*. C. Scribner's sons.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old Occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.

Heidelberg Academy of Sciences and Humanities, Bayerische Akademie der Wissenschaften, Academy of Sciences and Literature Mainz. 2025. Alma: Knowledge networks of medieval romance-speaking europe.

Kathryn Klingebiel. 2019. Occitan studies: Language and linguistics. *The Year's Work in Modern Language Studies*, 79(1):181–197.

Michele Loporcaro. 2018. *Gender from Latin to Romance: History, geography, typology*, volume 27. Oxford University Press.

Roy Lyster. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies*, 16(1):69–92.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniela Marzo and Marinus Wiedner. 2025. Remarks on grammatical gender in romance. In *Parla, e sie breve e arguto. Festschrift für Maria Selig / Studies in Honor of Maria Selig*, ScriptOralia 147, pages 201–207, Tübingen. Narr.

Josiane Mothe. 2024. Shaping the future of endangered and low-resource languages—our role in the age of llms: A keynote at ecir 2024. In *ACM SIGIR Forum*, volume 58, pages 1–13. ACM New York, NY, USA.

Vivi Nastase and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, Singapore. Association for Computational Linguistics.

Michele Pasquini and Maurizio Serva. 2021. Stability of meanings versus rate of replacement of words: an experimental test. *Journal of Quantitative Linguistics*, 28(2):95–116.

Maria Polinsky and Ezra Van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from latin to french. *Language*, 79(2):356–390.

Clamenca Poujade, Myriam Bras, and Assaf Urieli. 2024. CorpusArièja: Building an annotated corpus with variation in Occitan. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 66–71, Torino, Italia. ELRA and ICCL.

Elton Prifti, Wolfgang Schweickard, Maria Selig, and Sabine Tittel. 2023. Sprachdatenbasierte modellierung von wissensnetzen in der mittelalterlichen romania (alma): Projektskizze. *Zeitschrift für romanische Philologie*, 139(2):301–332.

Matthias Schöffel, Esteban Garces Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025. Unveiling factors for enhanced pos tagging: A study of low-resource medieval romance languages. *arXiv preprint arXiv:2506.17715*.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Béla Szlovicsák. 2023. Preliminary examination of the latin neuter on inscriptions. *Acta Antiqua Academiae Scientiarum Hungaricae*, 62(4):419–434.

Marinus Wiedner. 2025. Cometa: Corpus de l'occitan médiéval comparatif et annoté: Provence et languedoc. Zenodo.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739, Hong Kong, China. Association for Computational Linguistics.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do not neglect related languages: The case of low-resource occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

A Data Description

A.1 Gender Shift by Lemma Ending

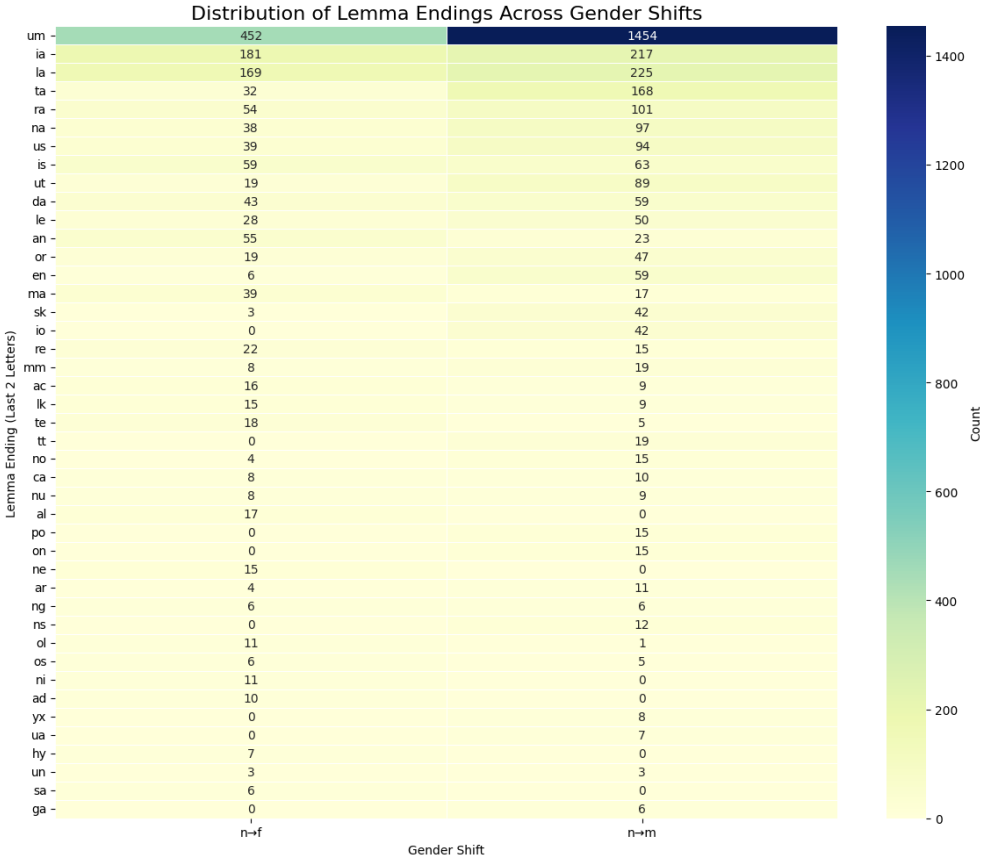


Figure 7: Gender shift frequencies for different lemma endings.

A.2 Lexical Diversity in Raw Occitan Texts

File	Tokens	Types	TTR	MATTR@50	MATTR@100	MATTR@500
Harley_3041.txt	443	227	0.512	0.852	0.776	0.512
Latin_901.txt	1105	492	0.445	0.830	0.739	0.537
Nouvelle_acquisition_française_11180.txt	4312	1324	0.307	0.824	0.735	0.520
Arsenal_6355.txt	11277	2491	0.221	0.849	0.762	0.545
Français_13504.txt	35802	6163	0.172	0.808	0.724	0.536
Roman_de_Flamenca.txt	46724	7852	0.168	0.861	0.787	0.585
Nouvelle_acquisition_française_11151.txt	20268	3151	0.155	0.766	0.661	0.429
Croisade_Albigeoise.txt	81305	12370	0.152	0.822	0.747	0.562
Add__10323.txt	53564	7609	0.142	0.850	0.764	0.552
Add__21218_ohne1-6.txt	31713	4485	0.141	0.805	0.717	0.507
Français_13503.txt	35352	4773	0.135	0.820	0.722	0.493
Français_13509.txt	62489	7867	0.126	0.849	0.763	0.548
Français_2232.txt	34217	4194	0.123	0.799	0.705	0.489
Lays_d-amors.txt	125475	13515	0.108	0.791	0.698	0.493
Français_1049.txt	130220	10425	0.080	0.778	0.680	0.463

Table 8: Lexical diversity metrics for the 15 raw Occitan corpora, including the size-dependent TTR and the more robust MATTR metric, computed with varying window sizes (Data Source: Wiedner (2025)).

B Preliminary Analysis

B.1 Embedding Model Backbone

(1) Frozen-encoder probing: We compare FastText (Bojanowski et al., 2016), mBERT (Devlin et al., 2019), and ByT5 (Xue et al., 2022) as frozen feature extractors with an identical linear classifier for Occitan *grammatical gender* prediction. Each instance is a bilingual pair $(w_{\text{lat}}, w_{\text{occ}})$ with Latin gender g_{lat}; we embed words in isolation, mean-pool subword/byte states, and classify $[e(w_{\text{lat}}); e(w_{\text{occ}}); \text{onehot}(g_{\text{lat}})]$. mBERT performs best on this probe (**Macro F1 = 72.04**), outperforming FastText and slightly exceeding ByT5; we therefore adopt mBERT as our default backbone.

Embedding Model	Macro F1	Accuracy %
FastText	57.51	58.30
mBERT	**72.04**	**73.37**
ByT5	71.24	72.56

Table 9: Comparison of frozen embedding models on the Occitan gender prediction task.

(2) Variant retrieval: To test whether embeddings place one-to-many Latin → Occitan realizations near each other, we cast variant identification as retrieval: given a Latin lemma w_{lat}, rank all candidate Occitan forms w_{occ} in the corpus by cosine similarity between isolated word embeddings (mean-pooled over subword/byte units). We report Recall@3 (and nDCG@3) since each query can have up to three attested variants in this experiment. Again, mBERT performs best on this probe (**Recall@3 = 0.59**), outperforming ByT5 and FastText, indicating that multilingual contextual encoders better cluster orthographic variants than static monolingual embeddings.

Embedding Model	Recall@k	MRR	nDCG
FastText	0.41	0.31	0.33
mBERT	**0.59**	**0.47**	**0.49**
ByT5	0.51	0.40	0.42

Table 10: Retrieval performance for identifying Occitan orthographic variants from a Latin lemma.

(3) Unsupervised structure: We probe intrinsic geometry by applying K-Means to isolated Occitan form embeddings and evaluating cluster agreement with canonical lemma labels. mBERT yields the best cluster separation on this probe (**Silhouette = 0.049**), outperforming ByT5 and FastText; taken together, these findings motivate our use of mBERT in the downstream pipeline.

Embedding Model	Silhouette Score
FastText	0.026
mBERT	**0.049**
ByT5	0.042

Table 11: Clustering performance of different embedding models. Higher scores indicate better-defined and more pure clusters with respect to canonical lemmas.

B.2 Tokenization Policy on BPE

We compare the standard mBERT WordPiece tokenizer against corpus-trained BPE tokenizers (vocabulary sizes 600 and 800) and a hybrid tokenizer that combines Occitan-adapted BPE with a word-level fallback. Tokenizers are evaluated using two criteria: **OOV rate**, defined as the proportion of tokens mapped to `[UNK]`, and **masked token recovery**, defined as top-1 accuracy at masked subword positions.

BPE formulation. BPE iteratively builds a subword vocabulary by merging the most frequent adjacent pair of symbols. At step t,

$$V_{t+1} = V_t \cup \{ab\}, \qquad (a, b) = \arg\max_{(x,y)} f(x, y), \tag{6}$$

where $f(x, y)$ is the corpus frequency of the pair (x, y). Repeating this process for a fixed number of merges yields a vocabulary of reusable subword units.

Summary of findings. Table 12 shows that corpus-trained BPE alone incurs non-zero OOV and very low masked recovery, while the standard mBERT tokenizer preserves full coverage but yields only moderate recovery. The hybrid tokenizer achieves the strongest overall trade-off, retaining zero OOV while substantially improving masked token recovery, which motivates its use in the downstream pipeline.

Tokenization Policy	OOV Rate (%)	Masked Token Recovery (%)
mBERT Tokenizer	0.0	15.78
BPE (vocab=600)	2.63	3.43
BPE (vocab=800)	2.86	4.76
Hybrid (BPE+word-level)	**0.0**	**25.23**

Table 12: OOV rate and masked token recovery for tokenization policies on the Occitan corpus.

C Algorithms

C.1 Algorithm 1: Construction of Occitan–Latin Lemma–Gender Dataset

We define

$$\textsc{Sim}(x, y) = \alpha\, \textsc{CosSim}(x, y) + (1 - \alpha)\, \textsc{LevSim}(x, y),$$

with

$$\textsc{LevSim}(x, y) = 1 - \frac{d_{\text{Lev}}(x, y)}{\max(|x|, |y|)}.$$

Since both $\textsc{CosSim}(x, y)$ and $\textsc{LevSim}(x, y)$ are normalized to $[0, 1]$, $\textsc{Sim}(x, y) \in [0, 1]$. We set $\alpha = 0.3$, i.e.,

$$\textsc{Sim}(x, y) = 0.3\, \textsc{CosSim}(x, y) + 0.7\, \textsc{LevSim}(x, y),$$

and accept a candidate iff

$$\textsc{Sim}(x, y) \geq 0.85.$$

The threshold and the value for α were chosen through qualitative assessment across samples and threshold settings with an Occitan linguistic expert.

C.2 Algorithm 2: Evaluation of Contextual Induction in Grammatical Gender Prediction

Algorithm 2 Evaluation of Contextual Induction in Grammatical Gender Prediction

Require: Dataset D of input instances
Require: Models M_{word} (word-only), M_{ctx} (context), M_{mask} (context with noun masked)
Ensure: Mean delta-probabilities and log-likelihood deltas for contextual induction; classification metrics

for all sample (X, i, W, L, G_L, Y) in D **do**
  $p_{\text{word}} \leftarrow M_{\text{word}}(X, i, W, L, G_L)$
  $p_{\text{ctx}} \leftarrow M_{\text{ctx}}(X, i, W, L, G_L)$
  $p_{\text{mask}} \leftarrow M_{\text{mask}}(X, i, W, L, G_L)$
  ▷ Ground-truth probability under word-only, context, and masked-context settings
  $\Delta_{p1} \leftarrow p_{\text{ctx}} - p_{\text{word}}$ ▷ prob. deltas
  $\Delta_{p2} \leftarrow p_{\text{mask}} - p_{\text{word}}$
  $\Delta_{p1}^{\log} \leftarrow \log p_{\text{ctx}} - \log p_{\text{word}}$ ▷ log-deltas
  $\Delta_{p2}^{\log} \leftarrow \log p_{\text{mask}} - \log p_{\text{word}}$
  ▷ Δ_{p1}: context vs word-only; Δ_{p2}: masked-context vs word-only
  Record deltas and predicted labels for summary
end for
Report mean(Δ_{p1}), mean(Δ_{p2}) ▷ context induction (prob.)
Report mean($\Delta_{p1}^{\log}$), mean($\Delta_{p2}^{\log}$) ▷ context induction (log)
Report accuracy and macro F1 for M_{word}, M_{ctx}, and M_{mask} classification ▷

C.3 Algorithm 3: Estimating the Impact of PoS Tags on Grammatical Gender Prediction

Algorithm 3 Estimating the Impact of PoS Tags on Grammatical Gender Prediction

Require: Sentences S with PoS tags (see Algorithm 1)
Ensure: Influence of PoS Tags

for all sentence $s \in S$ **do**
  Retrieve PoS tags $P = (p_1, p_2, \ldots, p_T)$ for s
  ▷ e.g., attention-, gradient-, or perturbation-based scores
  Construct mapping between tokens and PoS tags
  for $t = 1$ to T **do**
    Mask token x_t and recompute model confidence ▷ occlusion
    Record confidence change Δc_t ▷ per token
  end for
  Aggregate token scores (a_t) and/or confidence changes (Δc_t) by PoS tag for sentence s
end for
Aggregate tag-wise statistics across all sentences
return PoS-tag contributions to gender prediction

D Model Architecture and Hyperparameters for the Experiments

D.1 Lemma Experiment

Embedding	Model	Best hyperparameters
FastText	Logistic Regression	$C = 0.0027$; solver = `liblinear`; class_weight = `None`
FastText	Random Forest	n_estimators = 400; max_depth = 23; min_samples_split = 6; min_samples_leaf = 5; class_weight = `balanced`
FastText	XGBoost	n_estimators = 100; max_depth = 7; learning_rate = 0.0210; subsample = 0.7344; colsample_bytree = 0.8147; gamma = 0.6742; min_child_weight = 5
FastText	Feedforward NN (FFN)	hidden_dim = 64; dropout = 0.3934; lr = 0.0024; batch_size = 16; optimizer = `Adam`
FastText	BiLSTM	hidden_dim = 64; num_layers = 1; dropout = 0.3552; lr = 0.00969; batch_size = 16; optimizer = `Adam`
mBERT	Logistic Regression	$C = 0.0005$; solver = `liblinear`; class_weight = `None`
mBERT	Random Forest	n_estimators = 200; max_depth = 41; min_samples_split = 8; min_samples_leaf = 5; class_weight = `balanced`
mBERT	XGBoost	n_estimators = 150; max_depth = 9; learning_rate = 0.0164; subsample = 0.5431; colsample_bytree = 0.6104; gamma = 4.3585; min_child_weight = 5
mBERT	BiLSTM	hidden_dim = 256; num_layers = 3; dropout = 0.1298; lr = 0.00025; batch_size = 32; optimizer = `AdamW`
ByT5	Logistic Regression	$C = 0.0733$; solver = `lbfgs`; class_weight = `None`
ByT5	Random Forest	n_estimators = 400; max_depth = 37; min_samples_split = 9; min_samples_leaf = 4; class_weight = `None`
ByT5	XGBoost	n_estimators = 200; max_depth = 3; learning_rate = 0.0266; subsample = 0.8787; colsample_bytree = 0.9173; gamma = 2.8813; min_child_weight = 2
ByT5	BiLSTM	hidden_dim = 128; num_layers = 1; dropout = 0.4417; lr = 0.00053; batch_size = 32; optimizer = `AdamW`

Table 13: Best hyperparameter settings for all models and embedding families.

D.2 Context-Level Experiment

Hyperparameter	Exp1	Exp2	Exp3
Model Architecture			
Model Type	FieldsOnlyModel	ContextReaderModel	ContextSimpleModel
Base Encoder	BERT-base-multilingual-cased (768 dim)		
Encoder Frozen	Yes	No	No
Input Features	Word-level only	Context + Word	Masked Context + Word
Attention Mechanism	None	Multi-head (8 heads)	None
Attention Dimensions	N/A	$d_k = 128$, $d_v = 128$	N/A
Relative Position Bias	N/A	Yes (window=64)	N/A
No Self-Peek	N/A	Yes	N/A
Feature Dimensions	768	2560	2304
MLP Hidden Size	512	512	512
Training Hyperparameters			
Epochs	20	20	20
Batch Size	128	128	128
Learning Rate	2×10^{-5}	1×10^{-5}	2×10^{-5}
Weight Decay	0.01	0.01	0.01
Warmup Ratio	0.06	0.06	0.06
Optimizer	AdamW		
LR Schedule	Linear warmup + linear decay		
Gradient Clipping	0.5 (max norm)		
Dropout	0.1	0.1	0.1
Early Stopping Patience	3 epochs		
Early Stopping Metric	Validation loss		
Class Weights	Balanced (sklearn)		
Data Configuration			
Max Sentence Length	128 tokens		
Max Fields Length	32 tokens		
Cross-Validation	Group K-Fold (3 folds)		
Random Seed	13		
Experiment-Specific Details			
Context Usage	None	Full sentence	Masked sentence
Masking Strategy	N/A	N/A	Noun replaced with `NOUNTOKEN`
Fields Format	`[OCC]{word} [LAT]{lemma} [LG]{gender}`		

Table 14: Hyperparameters and training configuration for experiments 1, 2, and 3.

E PoS Tagger in the Study

The PoS tagger used in this study is (Manjavacas et al., 2019). A key limitation is that PoS tags are automatically predicted for the full Occitan corpus, and downstream analyses (including our occlusion-based PoS importance estimates) inherit tagging errors. To quantify tagger quality, we manually annotated a 60,000-token subset and evaluated the tagger against this gold data, obtaining 71.31% overall accuracy. Performance varies by tag: ADJ shows the lowest accuracy, while our primary tag of interest, NOUN, achieves 70.32%. We therefore interpret PoS-conditioned results as informative but potentially biased by tagging noise.

F Statistical Tests of the Experiments

F.1 Statistical Significance for Lemma Experiment

To complement the fold-averaged results in Table 3, we directly compare mBERT and ByT5 under a matched downstream architecture using paired bootstrap resampling over *out-of-fold* (OOF) predictions from the same 10 CV splits. Out-of-fold (OOF) predictions are obtained from the same lemma-grouped CV splits; therefore, each OOF prediction is made on a held-out lemma and is free of lemma-level leakage. This analysis is slightly different from Table 3: there, Macro-F1 is reported as the mean across folds, whereas here we compute a single OOF Macro-F1 over all 4,444 held-out predictions to obtain a more robust paired comparison at the item level. Under the shared 2×BiLSTM + MHSA classifier, mBERT

yields a higher OOF Macro-F1 than ByT5 (0.7608 vs. 0.7213; $\Delta = +0.0395$). The paired bootstrap confirms that this advantage is reliable in the present setup: the 95% confidence interval remains strictly above zero, and no bootstrap resample yields $\Delta \leq 0$.

System	OOF Macro-F1	Notes
mBERT	0.7608	
ByT5	0.7213	
Δ (mBERT $-$ ByT5)	+0.0395	95% CI [+0.0250, +0.0543]
p-value		$p < 10^{-6}$

Table 15: Paired bootstrap comparison between mBERT and ByT5 under the same 2×BiLSTM + MHSA architecture and identical 10-fold CV splits. Unlike Table 3, which reports mean Macro-F1 across folds, this table reports Macro-F1 computed over out-of-fold predictions for a paired item-level comparison. Because Table 3 reports fold means, we additionally perform a paired bootstrap on pooled out-of-fold predictions under the same architecture to test whether the mBERT–ByT5 difference is reliable at the instance level.

F.2 Statistical Significance for PoS Tags

To assess whether the observed PoS-wise occlusion effects are reliably different from zero, we perform a *sign-flip permutation test* for each PoS tag independently. For a given tag p, let $\delta_{i,p}$ denote the per-sample occlusion score, i.e., the change in confidence for the gold label when tokens with tag p are masked. Under the null hypothesis that the mean effect of p is zero, the sign of each $\delta_{i,p}$ is exchangeable; we therefore generate a null distribution by randomly flipping the sign of the per-sample scores over 10,000 permutations and recomputing the mean. We report the observed mean effect together with the resulting p-value from a two-sided test of whether the mean effect differs from zero, which asks whether a PoS tag provides a reliably positive or negative contextual contribution. The full numerical results are reported in the main text in Table 7.

G Explainability through Attention Heads on Contextual Cues Experiments

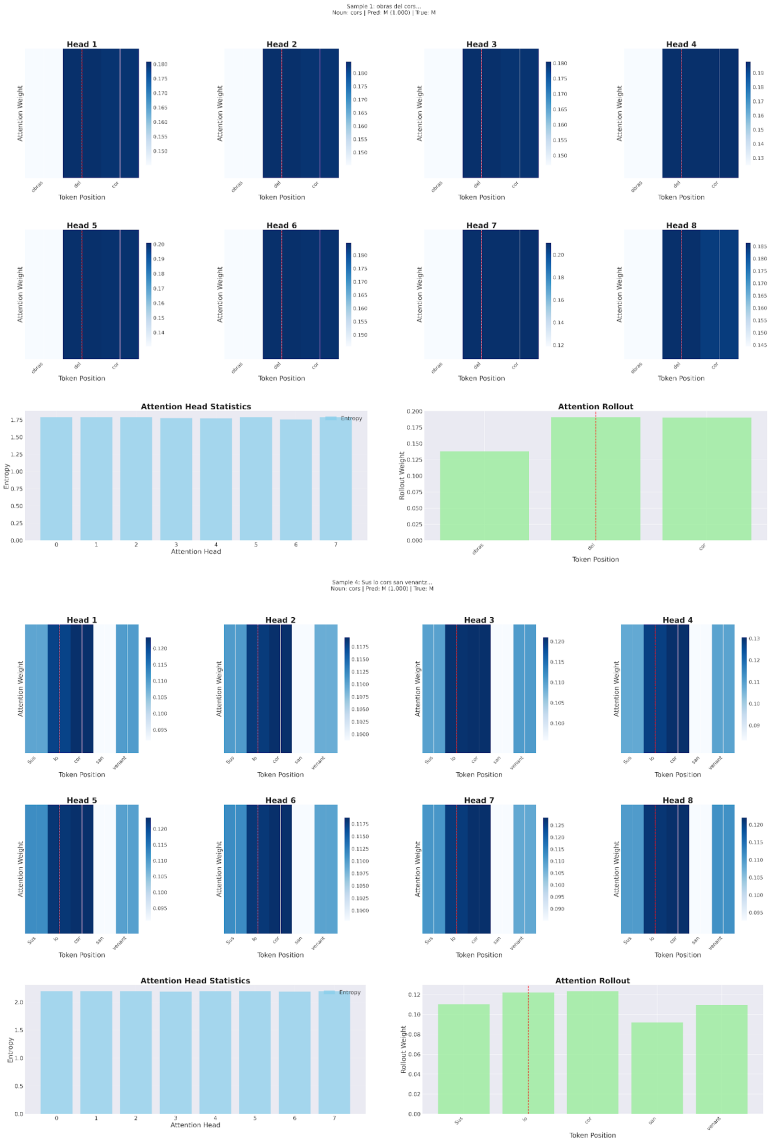


Figure 8: Attention-based contextual evidence for grammatical gender prediction shown for two representative Occitan sentences (top and bottom panels; target noun: *cors*). For each example, we visualize the 8 MHSA heads when using the noun representation as the query and the full sentence as keys/values; the dashed red line marks the target noun position. Across heads, attention concentrates on the noun and nearby agreement-bearing tokens (often including determiners/articles), consistent with morpho-syntactic cues for gender assignment. The per-head entropy plot (left) indicates broadly distributed head behavior, and attention rollout (right) summarizes aggregate token attribution across the sentence.

H Ablation Study on Contextual Cues Experiment

Ablation: removing Latin features. To quantify the contribution of Latin etymological information in the contextual setting, we ablate the Latin lemma and Latin gender from the input and retrain/evaluate the same context model under 3-fold lemma-grouped cross-validation. Table 16 reports mean performance (± std across folds). While the context model remains effective without Latin features, the contextual Δ gains in confidence are substantially reduced: with Latin lemma and gender, context increases the gold-class probability by $\sim$0.28 relative to the word-only baseline, whereas without Latin this increase drops to $\sim$0.09–0.11 (about 3$\times$ smaller). This indicates that Latin features provide critical complementary signal that amplifies the benefit of context.

Setting	**Accuracy**	**Macro-F1**
Context + full noun (no Latin)	0.961 ± 0.018	0.879 ± 0.036

Table 16: Contextual ablation without Latin lemma/gender. Results are mean $\pm$ std over 3-fold lemma-grouped cross-validation.

I Error Analysis

We analyse the 294 misclassifications made by the BiLSTM+Attention model across all folds using a SHAP-based surrogate approach. We train an XGBoost error predictor on 57 interpretable features capturing morphology (Latin/Occitan suffix cues, length, vowel ratio), frequency, sentence properties (length, noun position), and local syntax (PoS fractions and neighbouring tags) and explain its decisions with TreeSHAP (5-fold CV; ROC-AUC = 0.62). The strongest error drivers fall into three groups: (i) *context sparsity*—sentences with fewer agreement-bearing categories, especially adjectives in the immediate right context, yield more errors; (ii) *morphological ambiguity*—errors are more common when nouns occur at either sentence boundary; and (iii) *length/frequency effects*—errors are more common in shorter contexts and for mid-frequency items, consistent with a regime where very frequent forms are memorised and rare regular forms generalise more easily. Overall, masculine items exhibit a higher error rate than feminine ones, though the difference is very low.

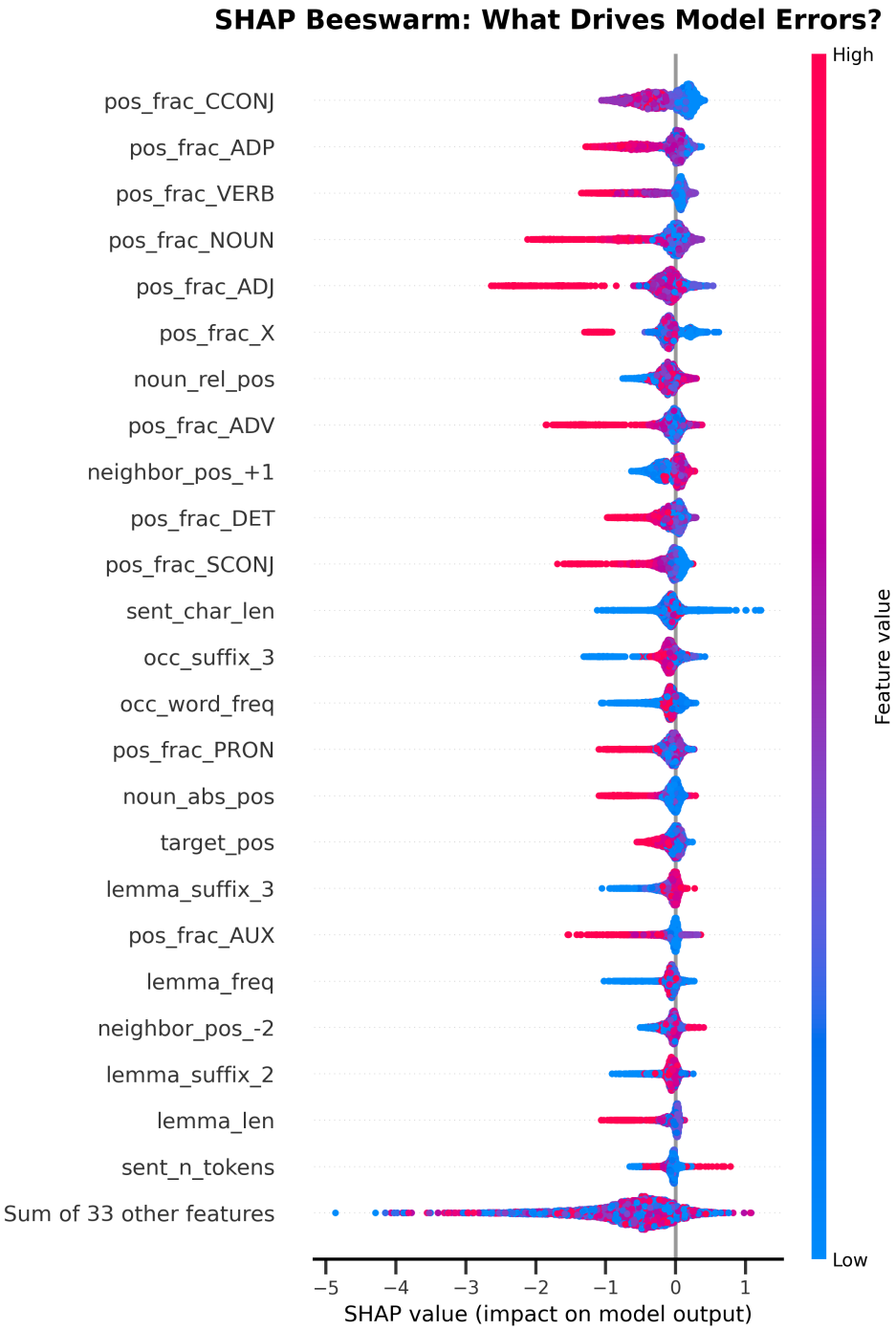


Figure 9: SHAP beeswarm plot showing feature contributions to model error prediction. Each dot represents a sample; the x-axis indicates the SHAP value (positive = pushes toward error, negative = toward correct), and colour encodes feature value (red = high, blue = low). The top five drivers are all POS composition features — fraction of coordinating conjunctions, adpositions, verbs, nouns, and adjectives in the sentence — indicating that syntactically sparse contexts lacking agreement-bearing words are the primary source of errors. Morphological features (Occitan suffix, Latin lemma suffix) and contextual factors (noun position, sentence length, word frequency) contribute secondarily. The immediate right-neighbour POS tag (rank 9) confirms that local agreement context, particularly an adjacent adjective, is a key disambiguating signal.

From Traditional Taggers to LLMs: A Comparative Study of POS Tagging for Medieval Romance Languages

Matthias Schöffel[1] **Esteban Garces Arias**[2,3]
[1]Bavarian Academy of Sciences (BAdW), Munich, Germany
[2]Department of Statistics, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML), LMU Munich, Germany
matthias.schoeffel@badw.de
esteban.garcesarias@stat.uni-muenchen.de

Abstract

Part-of-speech (POS) tagging for Medieval Romance languages remains challenging due to orthographic variation, morphological complexity, and limited annotated resources. This paper presents a systematic empirical evaluation of large language models (LLMs) for POS tagging across three medieval varieties: Medieval Occitan, Medieval Catalan, and Medieval French. We compare traditional rule-based and statistical taggers with modern open-source LLMs under zero-shot prompting, few-shot prompting, monolingual fine-tuning, and cross-lingual transfer learning settings.

Experiments on historically grounded datasets show that LLM-based approaches consistently outperform traditional taggers, with fine-tuning and multilingual training yielding the largest improvements. In particular, cross-lingual transfer learning substantially benefits under-resourced varieties, while targeted bilingual training can outperform broader multilingual configurations for specific target languages. The results highlight the importance of linguistic proximity and dataset characteristics when designing transfer strategies for historical NLP.

These findings provide empirical insights into the applicability of modern neural methods to medieval text processing and provide practical guidance for deploying LLM-based POS tagging pipelines in digital humanities research. All code, models, and processed datasets are released for reproducibility[1].

1 Introduction

The computational analysis of historical texts plays an increasingly important role in digital humanities, enabling large-scale investigation of linguistic change, cultural transmission, and textual variation. A fundamental prerequisite for such analyses is reliable linguistic annotation, with part-of-speech (POS) tagging serving as a key preprocessing step for downstream tasks including syntactic parsing, semantic analysis, and diachronic linguistic modeling (Piotrowski, 2012; Ehrmann et al., 2020).

POS tagging for Medieval Romance languages presents particular challenges. These varieties exhibit substantial orthographic instability, dialectal diversity, and complex morphological systems, while annotated corpora remain comparatively small and domain-specific (Schöffel et al., 2025a). An overview of the geographic distribution as well as the spelling characteristics in all languages is shown in Figure 1. As a result, computational tools originally developed for modern standardized languages often show limited robustness when applied to medieval textual material.

Traditional approaches to historical POS tagging have relied on rule-based systems and statistical models adapted from modern resources. While tools such as COLaF[2] and UDPipe[3] provide useful baselines, their performance can degrade in the presence of spelling variation, lexical sparsity, and genre-specific linguistic patterns. Recent advances in large language models (LLMs) offer new opportunities to address these challenges by leveraging contextual representations learned from large-scale multilingual corpora. However, systematic evaluation of LLM-based approaches for medieval Romance POS tagging remains limited, particularly with respect to prompting strategies, fine-tuning regimes, and multilingual transfer effects.

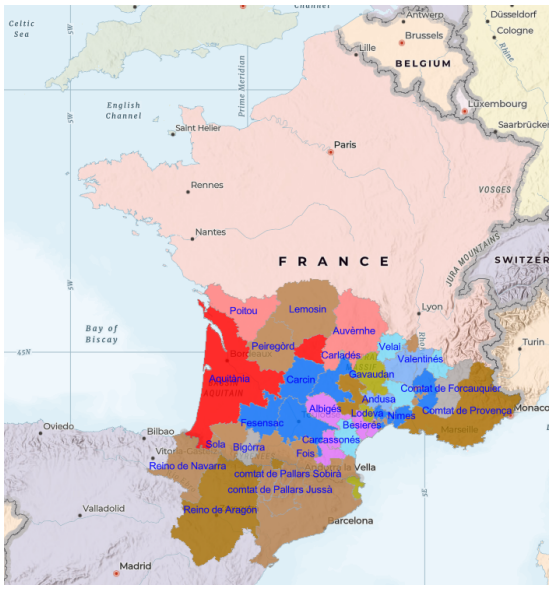
This study provides a comparative evaluation of traditional and neural methods across three medieval Romance datasets representing different genealogical branches and textual domains: Medieval Occitan, Medieval Catalan, and Medieval French. We investigate the following research questions:

[1]https://github.com/msch38/medieval-romance-pos/tree/main

[2]https://colaf.huma-num.fr/deucalion/occ-cont

[3]https://lindat.mff.cuni.cz/services/udpipe/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 297–313
July 4, 2026 ©2026 Association for Computational Linguistics



(a) Map of Medieval Occitan and Medieval Catalan variations, 13th century (Cabré, 2014).

Spelling variants	Modern/Known spelling
Medieval French	
deffendre	défendre (engl. 'to defend')
joinncture	jointure (engl. 'knuckle')
sun	son (engl. 'his')
pruz	preux (engl. 'brave')
Medieval Catalan	
ssaber	saber (engl. 'to know')
Ffrança	França (engl. 'France')
hòmens	homes (engl. 'men')
jóvens	joves (engl. 'youth')
Medieval Occitan	
deceplina	disciplina, disiplina, desiplina (engl. 'discipline')
falssa	fals (engl. 'false')
liech/lech	lloc (engl. 'place')
fuoc/foc	foc (engl. 'fire')

(b) Spelling characteristics across medieval language variations.

Figure 1: Geographic distribution and spelling characteristics of medieval Romance languages (13th century). Left: regional variation of Medieval Occitan and Medieval Catalan. Right: spelling variants across Medieval French, Medieval Catalan, and Medieval Occitan.

1. How do LLM-based approaches compare to established POS tagging tools for medieval Romance varieties?

2. To what extent do prompting strategies and decoding configurations influence tagging performance?

3. Can cross-lingual transfer learning improve performance across related historical languages?

Our contributions are threefold. First, we present a controlled empirical comparison of traditional taggers and modern LLMs across multiple training and inference regimes. Second, we analyze the effectiveness of bilingual and trilingual transfer configurations, highlighting the role of linguistic proximity in multilingual historical NLP. Third, we provide practical recommendations for selecting modeling strategies under different resource constraints. Together, these contributions aim to support the development of more robust annotation pipelines for historical text collections.

2 Related Work

Computational processing of historical language has evolved from early rule-based approaches toward statistical and neural methods that better accommodate linguistic variation and limited supervision. Foundational surveys (Piotrowski, 2012) highlight key challenges in historical NLP, including orthographic instability, domain specificity, and data scarcity. Subsequent work has explored normalization and annotation strategies for early language stages, demonstrating the importance of adapting preprocessing and modeling techniques to non-standard linguistic input (Scheible et al., 2011).

For Romance historical varieties, research has focused on developing specialized corpora and task-specific annotation tools. Studies on Classical and Early Modern French show that domain-adapted models can substantially improve lemmatization and POS tagging accuracy compared to general-purpose systems (Camps et al., 2021). Neural approaches have further expanded these capabilities. Work on historical text normalization and lemmatization demonstrates that sequence-to-sequence and contextualized models can effectively capture variation-rich morphological patterns (Bollmann et al., 2019; Manjavacas et al., 2019; Kestemont et al., 2016). Related efforts in OCR post-correction and handwritten text recognition highlight the broader applicability of neural pipelines in historical document processing (Springmann and Lüdeling, 2016; Koch et al., 2023; Garces Arias et al., 2023; Wiedner, 2023; Pavlopoulos et al., 2024; Sarawgi et al., 2026).

Recent studies have begun to investigate large language models in historical linguistic contexts. Prompt-based experiments suggest that pretrained multilingual models can transfer useful linguistic knowledge to low-resource historical varieties, although systematic comparisons across training regimes remain scarce (Schöffel et al., 2025a; Schöffel et al., 2025b). In parallel, research on cross-lingual transfer learning indicates that multilingual training can improve performance on under-resourced languages when genealogical or typological proximity is present (Karthikeyan et al., 2020; Müller et al., 2021). Building on this line of work, the present study provides a unified evaluation of traditional taggers, prompting-based LLM inference, fine-tuning, and multilingual transfer learning for medieval Romance POS tagging. By jointly analyzing multiple datasets and transfer configurations, we aim to clarify the relative strengths of these approaches in realistic historical NLP scenarios.

3 Methodology

We design a comparative experimental framework to evaluate traditional and neural approaches to part-of-speech (POS) tagging for medieval Romance languages. The study considers five experimental settings: direct application of traditional taggers, prompting-based large language model (LLM) inference, monolingual fine-tuning, bilingual cross-lingual transfer learning, and trilingual multilingual training. This setup enables systematic analysis of how varying levels of supervision and multilingual exposure affect tagging performance. In particular, the bilingual and trilingual configurations allow us to examine the role of linguistic relatedness and dataset composition in cross-lingual transfer. An overview of the experimental configuration is provided in Table 3.

3.1 Datasets

We employ three historically grounded datasets representing distinct medieval Romance varieties and textual domains, summarized in Table 1. Further dataset references are available in the accompanying GitHub repository. The three resources differ substantially in size, genre, and century of composition, which we revisit when interpreting cross-lingual transfer results in Section 4 and as a limitation in the closing discussion.

Dataset	Language	Genre	Tokens
NAF	Med. Occitan (14th c.)	Literary	45,457
CAT	Med. Catalan (13th c.)	Chronicle	59,359
Chauliac	Med. French (15th c.)	Medical	2,443

Table 1: Summary of the three medieval Romance datasets used in this study.

Medieval Occitan: the Nouvelle Acquisition Française 6195 (NAF6195), also known as manuscript M of the *Vida de Sant Honorat*, dating from the 14th century. This literary manuscript reflects Provençal tradition and contains approximately 45,457 manually annotated tokens (Wiedner, 2025). **Medieval Catalan**: the *Llibre dels Fets*, a historical chronicle describing the reign of James I of Aragon, composed in the 13th century. This chronicle comprises approximately 59,359 tokens with consistent morphological annotation (Pujol i Campeny and Meelen, 2021). **Medieval French**: the *Anathomie* section from Gui de Chauliac's *Grande Chirurgie*, a 15th-century medical treatise containing specialized technical vocabulary and approximately 2,443 annotated tokens (Tittel, 2004)[4].

3.2 Models

We include traditional POS taggers as baselines. **UDPipe** (Straka et al., 2016) is a neural pipeline trained on Universal Dependencies treebanks, offering tokenization, tagging, lemmatization, and parsing for many modern languages. **COLaF** is a neural tagger from the COLaF project covering French, the languages of France, and modern Occitan. Neither is natively trained on Medieval Catalan or Medieval Occitan, providing a realistic baseline for off-the-shelf generalization to historical Romance varieties. We additionally evaluate two open-source LLMs, Gemma3-12B (Gemma-Team et al., 2024) and Phi4-14B (Abdin et al., 2024), under prompting, monolingual fine-tuning, and multilingual fine-tuning. Except for COLaF and UDPipe on Medieval French, none of the evaluated models were explicitly trained on Medieval Occitan, Catalan, or French. Representative language coverage per model is summarized in Table 6 (Appendix A).

[4]All datasets were preprocessed through tokenization and sentence segmentation, followed by manual verification of annotations. Tagsets were harmonized using Universal Dependencies conventions to ensure cross-dataset comparability.

3.3 Prompting Strategies

As summarized in Table 2, we investigate two prompting configurations: zero-shot and few-shot. In the **zero-shot** setting, models receive task instructions specifying the Universal Dependencies POS tagset and output format without exposure to target-domain examples. In the **few-shot** setting, prompts additionally include a small fixed set of annotated tokens reflecting morphological variability and orthographic diversity across the three target varieties. The example block is held constant across all evaluations to ensure prompt-level comparability and to avoid variance introduced by random example sampling. We deliberately mix examples from all three medieval varieties in a single block to expose the model to the orthographic heterogeneity it must cope with at inference time, rather than to optimize per-language prompts; we revisit this design choice as a limitation. This design allows assessment of how limited in-context supervision influences tagging performance.

3.4 Decoding Strategies

To assess the robustness of prompting-based inference, we evaluate several commonly used decoding approaches (Wiher et al., 2022; Garces Arias et al., 2025). Specifically, we consider beam search, temperature sampling (Ackley et al., 1985), top-k sampling (Fan et al., 2018), and nucleus (top-p) sampling (Holtzman et al., 2019). Hyperparameter ranges for each decoding configuration are summarized in Table 3.

3.5 Fine-tuning Experiments

We investigate three fine-tuning scenarios, all using a single fixed 80%/20% train/test split per dataset. The same 20% test partition for each dataset is reused across the monolingual, bilingual, and trilingual settings, so that test tokens are never seen during training in any configuration and accuracy numbers are directly comparable across regimes.

First, in **monolingual fine-tuning**, each LLM is trained on the 80% partition of one dataset and evaluated on the corresponding 20% test partition.

Second, in **bilingual cross-lingual transfer learning** (CLTF), models are trained on the union of the 80% partitions of two datasets (CAT+OCC, CAT+FR, FR+OCC, where OCC = NAF and FR = Chauliac) and evaluated on the held-out 20% partition of each constituent language.

This setup allows analysis of how genealogical proximity and corpus characteristics influence transfer effectiveness. Third, in **trilingual CLTF**, models are trained on the combined 80% partitions of all three datasets and evaluated separately on each language's 20% test partition. This configuration tests whether broader multilingual exposure improves performance, particularly for lower-resource varieties. A true held-out cross-lingual evaluation (e.g., CAT+FR→NAF, with the target language entirely absent from training) is left for future work. Fine-tuning hyperparameters are provided in Table 8.

3.6 Evaluation Metrics

Model performance is assessed using standard classification metrics (see Appendix F). **Accuracy** measures the proportion of correctly predicted POS tags across all tokens, providing an overall performance indicator. **Macro-averaged F1** captures balanced performance across frequent and infrequent POS categories by averaging class-level F1-scores.

3.7 Experimental Setup

A comprehensive description of the models, datasets, experiments and tasks is summarized in Table 3.

4 Results

4.1 Overall Performance Comparison

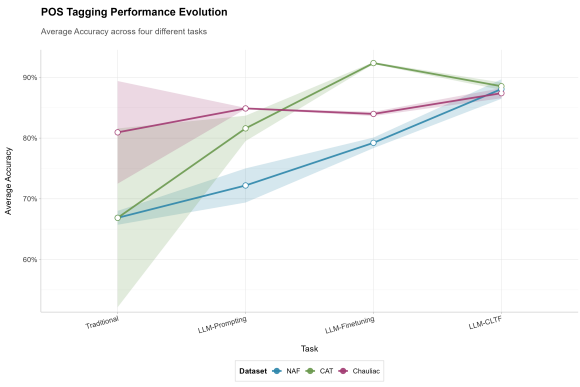


Figure 2: Illustration of performance evolution, in terms of Average Accuracy, per dataset (NAF in blue, CAT in green, Chauliac in purple), from traditional POS taggers (UDPipe, COLaF) through LLM prompting and monolingual fine-tuning to trilingual cross-lingual transfer learning (CLTF). Shaded areas represent variability across model and decoding configurations within each method family.

Prompting Strategy	Prompt
Zero-shot	*You are a linguistic expert in Medieval Romance languages.* *Analyze the given text and assign Universal Dependencies Part-of-Speech tags (UPOS) to each token.* *Available tags: "ADJ", "ADP", "ADV", "AUX", "CCONJ", "DET", "INTJ", "NOUN", "NUM", "PART", "PRON", "PROPN", "PUNCT", "SCONJ", "VERB", "X", "SYM".* *Return a JSON array of objects, each with only "word" and "UPOS" keys.* *Output only the JSON array, properly formatted and closed, with no extra text or explanation.*
Few-shot	***Zero-shot prompt* +** *Consider syntactic and semantic relationships, including agreement, word order, and morphology. Medieval Romance languages often exhibit significant spelling variation; for example, Old Occitan: 'ansy', 'eynsi', or 'anes'; Old Catalan: 'fiyl', or 'conseyl'; Middle French: 'norryr' or 'norrir'.* *Example format:* `[{"word": "bo", "UPOS": "ADJ"}, {"word": "volch", "UPOS": "VERB"}, {"word": "seyor", "UPOS": "NOUN"}, {"word": "homps", "UPOS": "NOUN"}, {"word": "sant", "UPOS": "ADJ"}, {"word": "iorn", "UPOS": "NOUN"}, {"word": "ilz", "UPOS": "PRON"}, {"word": "addicions", "UPOS": "NOUN"}, {"word": "deffendre", "UPOS": "VERB"}]`

Table 2: Comparison of different prompting strategies for UD POS tagging.

Figure 2 summarizes tagging accuracy across experimental configurations. Traditional POS taggers achieve competitive results on selected datasets but exhibit substantial performance variability across language varieties. Prompting-based LLM inference generally yields higher and more stable accuracy, while supervised fine-tuning further improves performance across all datasets. The largest gains are observed under multilingual training regimes, particularly for the lower-resource Medieval Occitan dataset.

In aggregate terms, traditional approaches reach an average accuracy of 71.56%, with pronounced dispersion across cells (52.15%–89.40%). Prompting-based LLM inference improves average accuracy to 77.75% (range 62.53%–84.98% across cells), while also reducing per-dataset variability relative to the traditional baselines. Dataset-specific patterns are evident: the Medieval Catalan dataset exhibits the largest performance gap between traditional systems and fine-tuned LLMs (10.93 percentage points), whereas improvements for NAF and Chauliac are more moderate (approximately 5 percentage points on average).

Monolingual LLM fine-tuning yields a further increase in mean accuracy to 85.19% (range 78.36%–92.52% across cells). The strongest results are obtained on the CAT dataset, where both Gemma3 and Phi4 exceed 92% accuracy. At the same time, the broader performance range observed under fine-tuning suggests sensitivity to dataset size, domain, and linguistic characteristics.

Bilingual cross-lingual transfer learning (CLTF) produces heterogeneous effects depending on language pairing. The CAT+OCC configuration achieves 89.25% accuracy on NAF, approaching trilingual performance levels, while CAT+FR reaches 93.14% on Chauliac, representing the highest accuracy observed for Medieval French across all experimental conditions. This bilingual configuration outperforms both monolingual fine-tuning (+9.50 percentage points) and trilingual CLTF (+4.91 percentage points) for this dataset. In contrast, the FR+OCC pairing results in only marginal improvement on NAF (80.31% vs. 80.09% monolingual), indicating that transfer effectiveness depends strongly on the specific language combination.

Trilingual CLTF achieves an average accuracy of 88.01% with the narrowest spread among all method families (range 86.48%–89.68% across cells), suggesting comparatively stable behavior across datasets. Relative to the UDPipe baseline, trilingual training leads to substantial gains on NAF (+21.67 percentage points) and CAT (+7.57 percentage points), while performance on Chauliac decreases slightly (-1.17 percentage points). Notably, the bilingual CAT+FR configuration remains superior for Medieval French, illustrating that broader multilingual exposure does not neces-

Models & Datasets	
Traditional	COLaF, UDPipe
LLMs	Gemma3-12B (Gemma-Team et al., 2024), Phi4-14B (Abdin et al., 2024) *Language support in Table 6, Appendix A*
Datasets	NAF (Medieval Occitan, 14th c.), CAT (Medieval Catalan, 13th c.), Chauliac (Medieval French, 15th c.)
Experimental Tasks	
Task 1: Traditional	Direct evaluation using COLaF and UDPipe on all datasets
Task 2: LLM Prompting	Zero-shot & few-shot prompting (Table 2) Decoding: beam search ($w \in \{1, 15\}$), temperature ($\tau \in \{0.6, 0.8, 0.9\}$), top-$k$ ($k \in \{5, 20, 50\}$), top-p ($p \in \{0.75, 0.85, 0.95\}$)
Task 3: LLM Fine-tuning	80/20 train/test split per dataset Each model fine-tuned and tested on same dataset (1-to-1 mapping)
Task 4: Bilingual CLTF	80% of two datasets for training, 20% per dataset for testing Pairwise: CAT+OCC, CAT+FR, FR+OCC (2-to-1 transfer)
Task 5: Trilingual CLTF	80% of all datasets for training, 20% per dataset for testing Multilingual training → monolingual testing (N-to-1 transfer)

Table 3: Experimental setup for POS tagging of medieval Romance languages. Evaluation focused on accuracy with precision, recall, and F1-measures available (Appendix F). All experiments used NVIDIA H100-96GB GPU. Hyperparameters detailed in Appendices B and C.

sarily yield optimal results for all target languages. A detailed comparison is provided in Table 4[5].

Overall, the progression from traditional methods (71.56%) to prompting (77.75%), fine-tuning (85.19%), and multilingual transfer (up to 88.01%) reflects systematic improvements associated with increased supervision and multilingual training signals. However, the results also suggest that optimal transfer configurations are language-dependent: trilingual training appears beneficial for Medieval Occitan, whereas targeted bilingual pairing with Catalan is more effective for Medieval French.

4.2 Task-Specific Analysis

4.2.1 Traditional vs. LLM-based Approaches

Across datasets, LLM-based methods consistently outperform traditional POS taggers. UDPipe achieves relatively strong results on the Chauliac dataset (89.40% accuracy), but performance declines substantially on NAF (68.01%), highlighting the challenges posed by orthographic variability and the limited training data available in Medieval Occitan.

[5]**Caveat.** Token counts differ by more than an order of magnitude across datasets (Table 1), so the strong CAT+FR→Chauliac result partly reflects the additional training material contributed by Catalan rather than genealogical proximity alone. Disentangling size, genre, and relatedness through controlled subsampling is left for future work.

4.2.2 Prompting Strategy Effectiveness

Few-shot prompting yields systematic improvements over zero-shot configurations across all models and datasets (see Figure 4 in Appendix E). Absolute gains range from 2.94 percentage points on Chauliac with Gemma3 to 10.24 percentage points on NAF with Phi4. Among the evaluated models, Phi4 demonstrates stronger prompting performance overall. In addition, deterministic decoding strategies tend to outperform sampling-based approaches, with beam search (beam width = 15) consistently achieving the highest accuracy. Detailed decoding-level results and variability analyses are provided in Tables 9 and 10 (Appendix D.1).

4.2.3 Monolingual Fine-tuning vs. Multilingual CLTF

Monolingual fine-tuning achieves the highest performance on Medieval Catalan (92.52% with Gemma3). In contrast, trilingual CLTF yields the largest improvement for Medieval Occitan, increasing accuracy from 80.09% under monolingual fine-tuning to 89.68%. This pattern suggests that cross-lingual supervision is particularly beneficial for lower-resource datasets. The relative impact of multilingual training across datasets and models is illustrated in Figure 3.

Task	Model/Strategy	NAF		CAT		Chauliac	
		Acc.	F1	Acc.	F1	Acc.	F1
Traditional	UDPipe	**68.01**	**67.29**	**81.59**	**81.19**	**89.40**	**89.53**
	COLaF	65.73	65.47	52.15	51.50	72.50	67.43
Prompting	Gemma3 Zero-shot	62.53	61.81	72.54	74.03	82.49	82.58
	Gemma3 Few-shot	69.39	69.22	79.48	80.49	84.80	85.20
	Phi4 Zero-shot	72.78	71.94	80.84	81.01	84.45	84.61
	Phi4 Few-shot	**75.01**	**74.31**	**83.69**	**83.75**	**84.98**	**85.19**
Fine-tuning	Gemma3	**80.09**	**79.99**	**92.52**	**92.50**	83.64	83.74
	Phi4	78.36	78.35	92.20	92.13	**84.33**	**84.10**
Bilingual CLTF	Gemma3 (CAT+OCC)	**89.25**	**89.18**	91.62	91.54	–	–
	Gemma3 (CAT+FR)	–	–	91.28	91.19	**93.14**	**93.02**
	Gemma3 (FR+OCC)	80.31	80.22	–	–	85.74	85.61
Trilingual CLTF	Gemma3	**89.68**	**89.66**	**89.16**	**89.11**	**88.23**	**88.09**
	Phi4	86.48	86.39	87.94	87.74	86.57	86.48
$\Delta_{\text{best, UDPipe}}$	Best CLTF vs UDPipe	+21.67	+22.37	+7.57	+7.92	+3.74	+3.49

Table 4: Overall performance comparison across methods and datasets. Best result per method is in **bold**; best overall result per column is highlighted in **green**. For each bilingual CLTF row, results are reported on the held-out partition of each constituent language; cells marked "–" indicate languages outside the training pair.

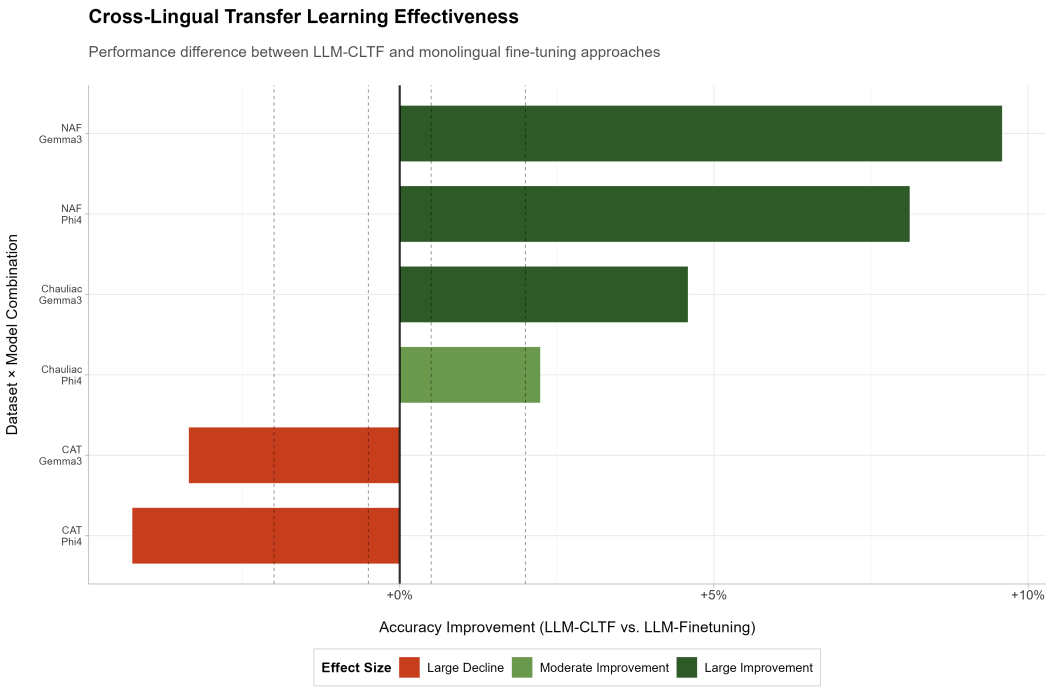


Figure 3: Effect of trilingual CLTF with respect to single-dataset LLM fine-tuning.

4.2.4 Bilingual vs. Trilingual Transfer

Bilingual CLTF results indicate that targeted language pairing can match or exceed trilingual training for specific datasets (see Table 4). Three observations emerge. First, **Medieval Catalan functions as an effective bridge language**. Pairing Catalan with either Occitan (CAT+OCC → NAF: 89.25%) or French (CAT+FR → Chauliac: 93.14%) leads to substantial gains over monolingual fine-tuning. This pattern is consistent with Catalan's intermediate genealogical position between Gallo-Romance and Occitano-Romance varieties. Second, **transfer effectiveness varies across language combinations**.

The configuration FR+OCC yields only minimal improvement for NAF (+0.22 percentage points), suggesting that linguistic proximity alone does not guarantee successful transfer. Corpus size imbalance and domain differences may also influence performance.

Third, **increasing the number of training languages does not uniformly improve results**. For Medieval French, bilingual CAT+FR training (93.14%) clearly outperforms trilingual CLTF (88.23%). This likely reflects the small size of the Chauliac dataset, which benefits from focused exposure to closely related Catalan material; all three corpora are domain-specific, so corpus size appears to be the more salient factor here. Conversely, Medieval Occitan shows the strongest performance under trilingual training, suggesting complementary transfer signals from both Catalan and French.

5 Error Analysis

5.1 Part-of-Speech Class Performance

Analysis of POS-specific F1-scores reveals systematic differences across methods. Table 5 (NAF) is shown here; CAT and Chauliac results are in Appendix D.2. Adjectives and adverbs remain particularly challenging for traditional approaches, with UDPipe yielding comparatively low F1-scores for these categories. LLM-based approaches, especially fine-tuned models, show marked improvements, which suggests that contextual representations may be helpful for disambiguating semantically richer POS classes. Pronouns also show consistent gains across LLM-based methods, potentially reflecting improved handling of context-dependent dependencies.

By contrast, function words such as adpositions and coordinating conjunctions maintain high performance across most methods. This pattern suggests that these categories may be more reliably identified from local syntactic cues. Verb classification also remains comparatively stable across approaches, indicating that morphological marking provides relatively strong signals even in the presence of orthographic variation.

5.2 Cross-Lingual Transfer Effects

Cross-lingual transfer results indicate that content-word categories (e.g., nouns, adjectives, verbs) tend to benefit more from multilingual training than function-word categories. The NAF dataset shows the largest gains under trilingual CLTF, with accuracy increasing from 80.09% under monolingual fine-tuning to 89.68%. These improvements are particularly pronounced for lower-frequency POS classes, suggesting that multilingual exposure can help mitigate data sparsity in historical language processing. The bilingual transfer results refine this picture further. The strong performance of the CAT+OCC configuration on NAF suggests that Catalan provides useful lexical and morphological information for Occitan. Similarly, the gains observed for CAT+FR on Chauliac indicate that focused bilingual training may be advantageous for small, domain-specific datasets.

6 Discussion: Implications for Historical NLP

The results of this study provide several insights for the development of computational pipelines in historical language processing. First, the observed performance differences across training regimes suggest that model adaptation strategies should be carefully aligned with resource availability and linguistic relatedness. In particular, multilingual training appears to offer robust benefits for lower-resource historical varieties, while targeted bilingual configurations may yield stronger performance for small, domain-specific datasets.

Second, the analysis highlights the importance of contextual modeling for linguistically variable input. Improvements observed for semantically richer POS classes such as adjectives, adverbs, and pronouns indicate that contextualized neural representations can mitigate challenges associated with orthographic instability and morphological variation. This finding supports recent trends toward integrating large pretrained models into digital humanities workflows, where annotation robustness is often constrained by heterogeneous textual sources.

Third, the heterogeneous transfer effects observed across language pairings underline the need for historically informed model design. Genealogical proximity alone does not fully determine transfer success; corpus size, genre, and annotation consistency also appear to play important roles. These factors suggest that future historical NLP research should move beyond purely language-centric transfer assumptions and adopt more data-driven criteria for multilingual training configuration. These findings indicate that modern neural tagging approaches can meaningfully complement traditional linguistic annotation methods.

POS	UDPipe	Phi4 FS	Gemma3 FT	Gemma3 CLTF	Δ
PROPN	25.85	72.34	78.31	**92.47**	+66.62
NUM	28.92	61.39	**91.89**	86.01	+62.97
AUX	38.39	45.08	53.58	**61.04**	+22.65
PRON	45.80	52.77	76.38	**81.51**	+35.71
ADV	50.61	54.19	66.92	**74.38**	+23.77
SCONJ	52.94	54.73	57.97	**94.62**	+41.68
ADJ	65.29	72.05	71.17	**73.58**	+8.29
VERB	67.77	79.91	75.79	**89.00**	+21.23
DET	73.81	73.48	**89.97**	87.99	+16.16
NOUN	76.45	83.65	82.81	**89.44**	+12.99
CCONJ	83.06	81.72	86.67	**96.34**	+13.28
ADP	85.51	89.93	88.59	**92.78**	+7.27

Table 5: Per-class F1 on **NAF**: low-performing (top) and high-performing (bottom) classes. FS = few-shot, FT = fine-tuned, Δ = improvement of best LLM-based method over UDPipe. Best per row in **green**.

7 Conclusion

This paper presents a systematic empirical evaluation of large language models for POS tagging across three medieval Romance language datasets. Relative to traditional tagging tools, LLM-based approaches yield consistent performance improvements under prompting, fine-tuning, and multilingual training settings. Cross-lingual transfer learning is particularly beneficial for lower-resource varieties, while targeted bilingual configurations can outperform broader multilingual training for specific target languages. Beyond reporting performance differences, the analysis highlights the importance of linguistic proximity and dataset composition when designing multilingual training strategies for historical NLP. Overall, the findings suggest that contemporary neural models can contribute to more reliable linguistic annotation pipelines for medieval texts and may support broader digital humanities workflows. Future work may include additional historical languages, examine downstream task effects, and explore alternative multilingual adaptation strategies.

Limitations Our study has several limitations. First, dataset sizes range from 2,443 to 59,359 tokens and represent different genres. We did not run controlled subsample or learning-curve experiments, so corpus size, genre, and genealogical proximity cannot be fully disentangled in the cross-lingual transfer results. Second, the few-shot block mixes examples from all three varieties; per-language prompts and example-selection strategies were not systematically explored. Third, the strongest results are obtained with fine-tuned 12–14B LLMs, which are substantially more expensive than UDPipe or COLaF. We did not evaluate smaller LLMs (e.g., 1–3B) or a modern pre-trained encoder fine-tuned for token classification (e.g., XLM-R). Moreover, Gemma3-12B and Phi4-14B were chosen as strong openly available instruction-tuned models that fit on a single H100 under LoRA; the 7–9B range and proprietary models were not explored. Finally, generalization to other historical language families and to downstream tasks (lemmatization, parsing, NER) remains untested.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work uses publicly available datasets and involves no human subjects or personally identifiable information. All code and data are released to enable reproducible research and further investigation.

Acknowledgments

We would like to express our gratitude to the ALMA (*Wissensnetze in der mittelalterlichen Romania*) project for their support and for granting us access to a partially annotated subset of the Chauliac dataset. We also sincerely thank Marinus Wiedner for his work on the annotation and publication of Medieval Occitan corpora. His efforts provided an important resource that made the broader analyses in the present work possible. Additionally, we thank the Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ) for providing the computational resources essential to this research. Esteban Garces Arias sincerely thanks the Mentoring Program of the Faculty of Mathematics, Statistics, and Informatics at LMU Munich and the Munich Center for Machine Learning (MCML) for their ongoing mentorship and financial support.

References

Marah Abdin, Jyoti Aneja, Ahmed Awadallah, Abhishek Awasthi, Ankur Bapna, Aseem Bhargava, Sarah Benchekroun, Aaron Bingeman, Shuo Chen, Weizhu Cheng, and 1 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2019. Large-scale historical text normalization with neural networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3904–3914. Association for Computational Linguistics.

Miriam Cabré. 2014. Trob-eu - troubadours and the european identity. `https://www.trob-eu.net/`. Accessed: 2025-06-30.

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. 2021. Corpus and models for lemmatisation and pos-tagging of classical french theatre. *Journal of Data Mining and Digital Humanities*.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Named entity recognition and classification in historical documents: a survey. *ACM Computing Surveys*, 53(4):1–47.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old Occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of the 8th International Conference on Learning Representations*.

Mike Kestemont, Enrique Manjavacas, and Folgert Karsdorp. 2016. Lemmatization for variation-rich languages using deep learning. In *Digital Humanities 2016: Conference Abstracts*, pages 193–195.

Philipp Koch, Gilary Vera Nuñez, Esteban Garces Arias, Christian Heumann, Matthias Schöffel, Alexander Häberlin, and Matthias Assenmacher. 2023. A tailored handwritten-text-recognition system for medieval Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 103–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1493–1503. Association for Computational Linguistics.

Lukas Müller, Rico Sennrich, and Martin Volk. 2021. First experiments in neural translation of historical texts. In *Proceedings of the Workshop on Language Technologies for Historical and Ancient Languages*, pages 23–32. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps, and Franz Fischer. 2024. Challenging error correction in recognised byzantine Greek. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 1–12, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Afra Pujol i Campeny and Marieke Meelen. 2021. Annotated corpora of historical catalan (hiscat) - llibre dels fets.

Anjali Sarawgi, Esteban Garces Arias, and Christof Zotter. 2026. Digitizing nepal's written heritage: A comprehensive htr pipeline for old nepali manuscripts. *Preprint*, arXiv:2512.17111.

Silke Scheible, Sarah Schulz, and Michael Wojatzki. 2011. Token-based chunking of historical german. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social*

Sciences, and Humanities, pages 84–89. Association for Computational Linguistics.

Matthias Schöffel, Esteban Garces Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025b. Unveiling factors for enhanced pos tagging: A study of low-resource medieval romance languages. Under review. ArXiv preprint, submitted.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025a. Modern models, medieval texts: A pos tagging study of old occitan. *Preprint*, arXiv:2503.07827.

Uwe Springmann and Anke Lüdeling. 2016. Automatic quality evaluation and (semi-)automatic improvement of mixed models for ocr on historical documents. In *Proceedings of the 12th Conference on Natural Language Processing (KONVENS 2014)*, pages 179–185.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Sabine Tittel. 2004. Die "anathomie" in der "grande chirurgie" des gui de chauliac : wort- und sachgeschichtliche untersuchungen und edition.

Marinus Wiedner. 2023. Old Occitan handwriting. (modell-nr. 52822, CER=3,51%), PyLaia-Modell for handwritten Occitan from the 13th and 14th century.

Marinus Wiedner. 2025. Cometa : Corpus de l'occitan médiéval comparatif et annoté: Provence et languedoc.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012.

Appendix

A Supported languages (pre-training)

Language	COLaF	UDPipe	Phi4-14B	Gemma3-12B
Occitan (modern)	✓		✓	
Medieval Occitan				
Catalan (modern)		✓	✓	
Medieval Catalan				
French (modern)	✓	✓	✓	✓
Medieval French	✓	✓		
Spanish (modern)		✓	✓	✓
Italian (modern)		✓	✓	✓
Portuguese (modern)		✓	✓	✓
Romanian (modern)		✓	✓	✓
Galician (modern)		✓	✓	
Asturian (modern)			✓	
Sardinian (modern)			✓	
Sicilian (modern)			✓	
Ligurian (modern)			✓	
Lombard (modern)			✓	
Venetian (modern)			✓	
Friulian (modern)			✓	
Arabic		✓	✓	✓
English		✓	✓	✓

Table 6: Language support (modern vs. medieval) across traditional POS taggers (COLaF, UDPipe) and LLMs (Phi4-14B and Gemma3-12B).

B Hyperparameters for LLM Prompting

Category	Hyperparameter	Value
Tokenizer	Max Length	8192
	Padding Side	left
	Data Type	torch.bfloat16 (Gemma) torch.float16 (Phi-4)
Model	Max New Tokens	300
	Batch Size	8
Processing	Chunk Size	20
	Window Length	5

Table 7: Hyperparameters used for LLM prompting experiments.

C Hyperparameters for LLM Fine-tuning

Category	Hyperparameter	Value
LoRA	LoRA Rank (r)	16
	LoRA Alpha (α)	32
	LoRA Dropout	0.1
	Target Modules	q_proj, v_proj, k_proj, o_proj
Training	Learning Rate	2×10^{-4}
	Batch Size	4
	Number of Epochs	10
	Optimizer	AdamW
	Weight Decay	0.01

Table 8: Hyperparameters used for LLM fine-tuning experiments with LoRA.

D Performance Analysis

D.1 Effect of Decoding Strategies

Model	Strategy	NAF	CAT	Chauliac	Average	Std Dev
Gemma3	Zero-shot + Beam-15	62.53	72.54	82.36	72.48	10.08
	Few-shot + Beam-1	69.24	79.37	84.27	77.63	7.51
	Few-shot + Beam-15	**69.39**	**79.52**	84.51	77.81	7.50
	Few-shot + Top-k-5	69.22	79.48	**84.80**	**77.83**	7.79
	Few-shot + Top-k-20	69.29	79.33	84.35	77.66	7.51
	Few-shot + Top-k-50	69.17	79.41	84.28	77.62	7.56
	Few-shot + Top-p-0.75	69.33	79.47	84.56	77.79	7.62
	Few-shot + Top-p-0.85	69.34	79.42	84.44	77.73	7.54
	Few-shot + Top-p-0.95	69.27	79.31	83.99	77.52	**7.34**
	Few-shot + Temp-0.6	69.23	79.46	84.49	77.73	7.63
	Few-shot + Temp-0.8	69.30	79.35	84.31	77.65	7.48
	Few-shot + Temp-0.9	69.35	79.43	84.39	77.72	7.52
Phi4	Zero-shot + Beam-15	72.77	80.84	84.09	79.23	6.67
	Few-shot + Beam-1	74.86	83.47	84.60	80.98	5.37
	Few-shot + Beam-15	**75.01**	**83.69**	**84.98**	**81.23**	5.32
	Few-shot + Top-k-5	74.02	82.88	84.31	80.40	5.15
	Few-shot + Top-k-20	73.76	82.85	83.98	80.20	5.55
	Few-shot + Top-k-50	73.80	82.81	84.11	80.24	5.21
	Few-shot + Top-p-0.75	74.50	83.46	84.51	80.82	5.53
	Few-shot + Top-p-0.85	74.49	83.34	84.00	80.61	5.43
	Few-shot + Top-p-0.95	74.19	83.16	84.56	80.64	5.70
	Few-shot + Temp-0.6	74.48	83.23	84.53	80.75	5.53
	Few-shot + Temp-0.8	74.27	82.94	83.78	80.33	**4.84**
	Few-shot + Temp-0.9	74.26	82.97	84.69	80.64	5.95

Table 9: Comprehensive decoding strategy performance analysis. Best results per model are highlighted in **bold**, while best overall results per column are highlighted in **green**.

Strategy Type	Mean Acc.	Std Dev.	CV	Range	Recommendation
Phi4 Few-shot					
Beam Search	81.23	0.12	0.001	0.5	Most reliable
Top-k Sampling	80.28	0.20	0.002	1.1	Good alternative
Top-p Sampling	80.69	0.18	0.002	0.8	Balanced performance
Temperature	80.57	0.21	0.003	0.9	Moderate variance
Gemma3 Few-shot					
Beam Search	77.81	0.14	0.002	0.3	Consistent but lower
Top-k Sampling	77.70	0.11	0.001	0.5	Very consistent
Top-p Sampling	77.68	0.14	0.002	0.4	Stable performance
Temperature	77.70	0.08	0.001	0.2	Most consistent

Table 10: Decoding strategy robustness and variance analysis. CV = coefficient of variation (Std Dev / Mean), Range = max - min across datasets. Highlighted cells indicate the best combination of performance and stability.

D.2 POS Class Performance

POS Class	UDPipe	Phi4 Few-shot	Gemma3 Fine-tuned	Gemma3 CLTF	Improvement
ADJ	54.12	58.11	**79.75**	71.94	+25.63
ADV	51.19	58.79	**77.30**	72.93	+21.74
PRON	62.19	68.66	**84.47**	81.10	+22.28
DET	71.69	74.40	87.32	**89.38**	+17.69
PROPN	79.90	76.26	**98.07**	91.22	+18.17
NOUN	86.14	85.34	**91.84**	88.72	+5.70
VERB	91.31	**93.55**	92.29	87.91	+2.24
ADP	**94.34**	93.09	94.16	92.65	-0.18
CCONJ	95.02	95.86	**98.89**	96.22	+3.87

Table 11: Per-class F1 on **CAT**: low-performing (top) and high-performing (bottom) classes. Best per row in **green**.

POS Class	UDPipe	Phi4 Few-shot	Gemma3 Fine-tuned	Gemma3 CLTF	Improvement
NUM	48.78	60.87	83.33	**88.04**	+39.26
AUX	**56.45**	32.97	40.00	49.30	-7.15
ADJ	68.66	64.52	57.14	**70.27**	+1.61
ADV	75.59	**77.83**	64.15	74.02	+2.24
PROPN	76.19	66.67	75.00	**90.47**	+14.28
VERB	86.32	82.55	67.39	**86.71**	+0.39
PRON	**88.50**	76.66	83.72	79.90	-4.78
DET	**91.79**	82.72	76.47	89.25	-2.54
NOUN	**92.88**	90.87	89.51	88.29	-2.01
ADP	**93.14**	87.70	90.62	92.16	-0.98
CCONJ	93.99	89.42	91.30	**95.65**	+1.66

Table 12: Per-class F1 on **Chauliac**: low-performing (top) and high-performing (bottom) classes. Best per row in **green**.

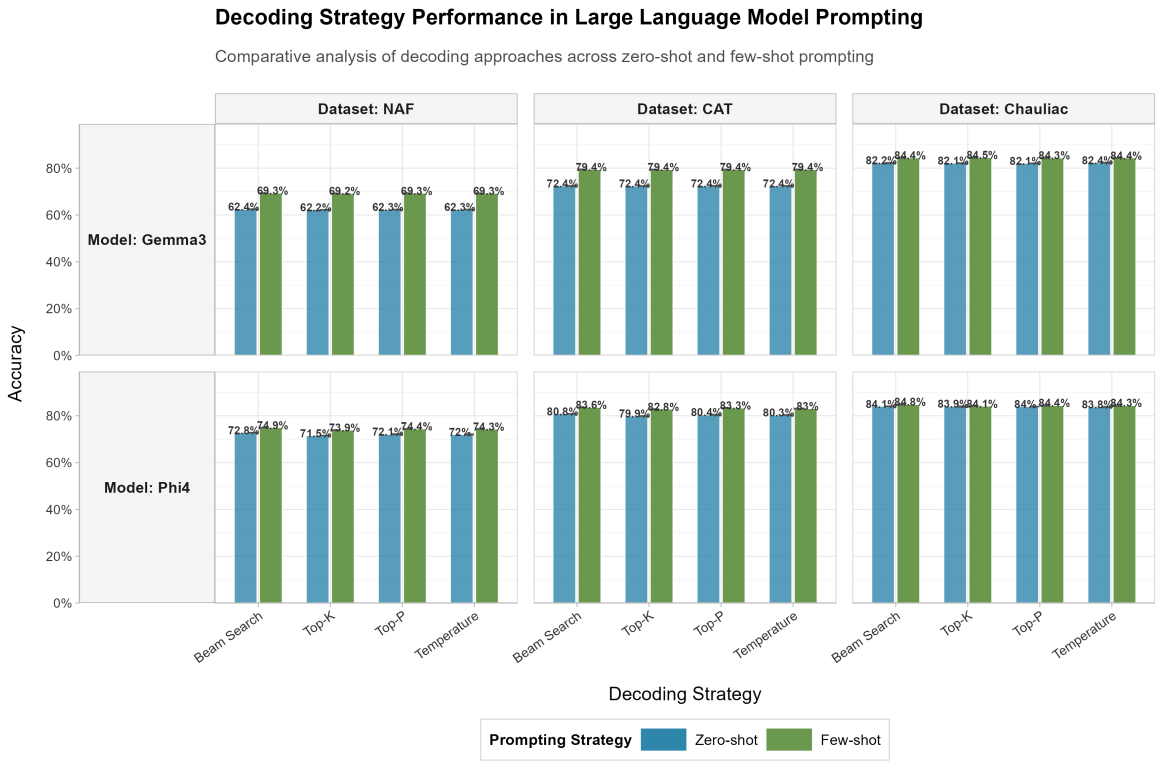


Figure 4: Decoding strategy performance across varying prompts, models and datasets.

E Decoding effects

F Evaluation metrics

We assessed our model using several standard metrics, defined as follows.

Accuracy Accuracy quantifies the proportion of tokens for which the predicted POS tag matches the gold label:

$$\text{Accuracy} = \frac{\#\text{correct predictions}}{\#\text{tokens}}. \tag{1}$$

Precision Precision measures the fraction of correct POS tag predictions among all instances predicted as a given tag:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

Recall Recall determines the proportion of actual POS tag instances that were correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

F1-score The F1-score, representing the harmonic mean of precision and recall, is computed as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

G Practical Recommendations

G.1 Method Selection Framework

Performance analysis reveals distinct optimal strategies depending on computational resources and target language characteristics, as illustrated in Table 13.

For resource-scarce languages like Medieval Occitan, trilingual CLTF provides substantial gains (+21.67 percentage points over traditional methods). Medieval Catalan benefits most from dedicated fine-tuning, while Medieval French achieves peak performance through bilingual pairing with Catalan, surpassing both trilingual CLTF and traditional tools. Under resource constraints, UDPipe remains a competitive option for Medieval French.

G.2 Implementation Guidelines

Prompting Configuration Few-shot prompting consistently outperforms zero-shot across all conditions, with improvements ranging from 2.94 to 10.24 percentage points. Phi4 demonstrates superior prompting capabilities, achieving 81.23% average accuracy compared to Gemma3's 77.81%. For decoding, beam search with width 15 provides optimal results across all datasets and models.

Cross-Lingual Transfer Learning CLTF shows particular effectiveness for under-resourced varieties. Medieval Occitan achieves the largest improvement (+9.59 percentage points over monolingual fine-tuning with trilingual CLTF), while Medieval French benefits most from targeted bilingual pairing with Catalan (+9.50 percentage points over monolingual fine-tuning). We recommend trilingual CLTF when the target language can benefit from broad cross-lingual exposure, and bilingual CLTF when a closely related language with sufficient training data is available—particularly for small target datasets where trilingual training may dilute the transfer signal.

Strategic Language Pairing Our bilingual experiments demonstrate that linguistic similarity should guide language pairing decisions. Catalan, occupying an intermediate position between Gallo-Romance and Occitano-Romance, serves as an effective bridge language for both Occitan and French. In contrast, the French–Occitan pairing yields minimal gains (+0.22 percentage points), suggesting that more distant genealogical relationships or domain mismatches can limit transfer effectiveness. We recommend practitioners prioritize pairings based on genealogical proximity and complementary corpus characteristics rather than defaulting to the largest possible training set.

Performance-Cost Trade-offs The progression from prompting (77.75% average accuracy) to fine-tuning (85.19%) to CLTF (up to 93.14% with optimal pairing) represents varying returns relative to computational investment. For production systems processing single languages, the 7.44 percentage point improvement from prompting to fine-tuning may justify computational costs. Bilingual CLTF offers an attractive middle ground: it requires only one additional language's training data yet can deliver substantial gains, particularly for small target datasets. Trilingual CLTF provides the most consistent performance across all languages but may be suboptimal for individual targets where a strong bilingual pairing exists.

G.3 Quality Assurance Considerations

Error analysis reveals systematic performance patterns across POS classes. Content words (ADJ, ADV, PRON) show the largest improvements with neural methods, with F1-score gains exceeding 25 percentage points for adjectives and adverbs. Function words (ADP, CCONJ) maintain consistently high performance (>90% F1) across all methods, suggesting reliable baseline capabilities. For production deployment, we recommend implementing class-specific validation protocols, particularly for content word categories where traditional methods show substantial limitations (ADJ: 54.12% F1 with UDPipe vs. 79.75% with fine-tuned models).

G.4 Resource Allocation Strategy

Based on performance variance analysis, trilingual CLTF provides the most stable results across cells (sample SD = 1.31 percentage points, CV ≈ 0.015), whereas traditional methods show high variability (sample SD $\approx$ 13.0 percentage points across cells). For multi-language digital humanities projects, trilingual CLTF training followed by language-specific evaluation provides robust performance with predictable resource requirements. However, when the goal is to maximize accuracy for a specific target language, practitioners should evaluate bilingual pairings with genealogically proximate languages before defaulting to trilingual training.

Dataset	High Resources	Limited Resources
NAF (Medieval Occitan)	Trilingual CLTF (89.68% acc.)	Few-shot Prompting (75.01% acc.)
CAT (Medieval Catalan)	Fine-tuning (92.52% acc.)	Few-shot Prompting (83.69% acc.)
Chauliac (Medieval French)	Bilingual CAT+FR (93.14% acc.)	UDPipe (89.40% acc.)

Table 13: Method selection by dataset and computational constraints.

Statistical Structure in Indus Sign Sequences

Tanishk Tiwari
DSEU
t.high1945@gmail.com

Abstract

This paper introduces a computational framework for evaluating structural properties of the undeciphered Indus script.The study uses a corpus of 6,579 inscriptions.The analytical approach combines unsupervised visual clustering of sign morphology, entropy-based sequence analysis, Kullback-Leibler divergence comparison, and neural sequence modeling (BiLSTM). The results indicate directional asymmetry and structured combinatorial patterns in sign sequences. We conclude that the Indus sign sequences exhibit statistical properties consistent with structured symbolic systems and not easily explained by random generation.

1 Introduction

The Indus Valley Civilization (c. 2600–1900 BCE) produced a symbolic system primarily found on steatite seals, copper tablets, and pottery sherds. Figure 1 shows an example inscription from the dataset. The short sequences—averaging four signs in length—and the absence of bilingual artifacts have prevented decipherment.

Figure 1: Example Indus seal artifact containing a short sequence of signs above an animal motif.

The inability to decipher the text has generated distinct hypotheses regarding its nature. Some scholarship posits that the sign system exhibits sequential statistical properties consistent with structured symbol systems (Mahadevan, 1977; Parpola, 1994). Conversely, other researchers argue that the sign sequences represent non-linguistic administrative emblems or heraldic symbols (Farmer et al., 2004).

This work does not assume that the Indus inscriptions represent a fully developed writing system. Instead, it investigates whether their statistical properties resemble those observed in known linguistic symbol systems, non-linguistic structured symbol systems, or random sequences.

The central question of this study is whether the statistical structure of Indus sign sequences more closely resembles that of linguistic writing systems or structured non-linguistic symbol systems. Specifically, the study evaluates three questions:

1. Do sign sequences exhibit directional asymmetry?
2. Do signs form stable visual clusters that may correspond to functional categories?
3. Do sequence models detect ordering constraints within inscriptions?

To address these questions, the study applies a multimodal computational framework combining computer vision, statistical sequence analysis, and neural sequence modeling.

2 Related Work

Early efforts by Hunter (1934) pioneered the cataloging of sign variations. Subsequently, Knorozov (1965) and Knorozov et al. (1979) used early computational methods to conduct positional analysis, identifying recurrent sign combinations suggestive of structured sequential composition. Mahadevan (1977) and Parpola (1994) analyzed positional

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 314–319
July 4, 2026 ©2026 Association for Computational Linguistics

statistics to suggest the presence of terminal sequence markers.

Recent computational interventions have sought to quantify the script's flexibility. Rao et al. (2009) utilized first-order conditional entropy to suggest that the script's sequential flexibility aligns with natural languages. However, Sproat (2014) demonstrated that entropic measures alone are insufficient diagnostics for language, as they fail to differentiate written language from non-linguistic symbol systems with internal rules. Mukhopadhyay (2023) proposed an administrative model concerning taxation and trade, while Yajnadevam (2024) attempted direct phonetic mapping.

Our work acknowledges the critiques raised by Sproat (2014). We combine sequence modeling with computer vision and differential cross-system baselines to identify statistical regularities in sign ordering and positional constraints, rather than using isolated metrics to claim linguistic status.

3 Methodological Framework

To evaluate the structural constraints of the script, we process the data through a four-stage analytical pipeline. First, we curate a standardized corpus from historical catalogs. Second, to avoid modern human bias in sign classification, we extract visual features directly from glyph images and apply unsupervised clustering to group structurally similar signs independently of sequence statistics. Third, we apply information-theoretic metrics to test for syntax directionality and compare these metrics against linguistic and non-linguistic baselines. Finally, we train a neural sequence model. By feeding the model systematically altered synthetic sequences, we observe which structural manipulations the model assigns lower probability to, thereby analyzing hidden statistical structure.

4 Methodology

4.1 Dataset

The digitized corpus used in this study is derived from publicly available Indus sign lists and inscription catalogs (Mahadevan, 1977; Wells, 2011). Inscriptions containing illegible signs, damaged portions, or non-numeric tokens were filtered. The dataset was partitioned into an 80/20 train/test split, stratified based on archaeological provenance metadata (object type). The vocabulary consists of numeric sign identifiers bounded dynamically by padding (<PAD>), start-of-sequence (<SOS>), and end-of-sequence (<EOS>) tokens.

4.2 Visual Feature Extraction

To establish sign families based on physical traits, we processed binarized glyph images. We extracted Hu Moments to achieve rotational invariance, as signs are frequently carved at varying angles. We additionally extracted Histogram of Oriented Gradients (HOG) features (64×64 pixels, 8×8 cells, 9 orientations) to capture local texture.

4.3 Sign Clustering

We applied feature standardization to the combined visual feature space. To reduce dimensionality while preserving non-linear geometric relationships, we applied Uniform Manifold Approximation and Projection (UMAP). Subsequently, we applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) followed by agglomerative clustering to identify morphological sign communities.

Purpose for Humanities Scholars: The goal of this unsupervised clustering is *not* to reduce the number of signs for modeling, but to discover natural visual families that may reflect functional or scribal categories used by ancient Indus scribes. All subsequent statistical analyses (entropy, KL divergence, sequence modeling) were performed on the **original** fine-grained sign inventory (400 signs), not on the clusters. The 8–20 morphological families serve primarily as an independent validation of structure and help interpret which visual types of signs carry positional or combinatorial importance.

4.4 Entropy Analysis

We calculated unigram (H_1), bigram (H_2), and trigram (H_3) conditional entropies for both the original corpus and a mechanically reversed corpus to test for statistical irreversibility.

Humanities Interpretation: In this context, *conditional entropy* measures the predictability of a sequence. Lower entropy indicates tighter syntactical rules (e.g., if Sign A is almost always followed by Sign B, entropy is low). *Markov orders* refer to the window of context: unigram (1 sign), bigram (2 signs), and trigram (3 signs).

We also calculated *Positional Rigidity* (R_i) for every absolute position i in the sequence, defined as $R_i = 1 - (H_i/H_1)$, where H_i is the positional entropy $H(X|\text{position} = i)$. Positional rigidity measures how strictly a sign is tied to a specific

physical location in a text (e.g., a sign that only ever appears at the very end of a line has high positional rigidity).

4.5 Cross-System Validation

To evaluate whether positional rigidity provides meaningful structural information, we apply the metric to three systems: a known writing system (Linear B), a synthetic heraldic sequence system, and the Indus corpus. Linear B was chosen because its linguistic status is firmly established and it exhibits short administrative inscriptions similar in length to Indus texts. The heraldic system serves as our structured non-linguistic baseline; it mimics European Blazonry (coats of arms), possessing strict positional rules (e.g., crests must be terminal, charges central) but mapping to no spoken language.

4.6 KL Divergence Analysis

To quantify the directional dependence of the joint probability distributions, we computed the Kullback-Leibler (KL) Divergence between the original forward transition probability matrix (P_{fwd}) and the mechanically reversed matrix (P_{bwd}):

$$D_{KL}(P_{fwd}||P_{bwd}) = \sum_{x \in \mathcal{X}} P_{fwd}(x) \log_2 \left(\frac{P_{fwd}(x)}{P_{bwd}(x)} \right) \quad (1)$$

Humanities Interpretation: KL Divergence measures *directional syntax*. A high divergence indicates that reading the text left-to-right follows entirely different structural rules than reading it right-to-left, strongly implying a fixed, intended reading direction.

4.7 Representation Learning of Sign Sequences

We modeled the syntax using a Categorical Hidden Markov Model (HMM), evaluating state distributions using the Bayesian Information Criterion (BIC). We then trained a Bidirectional Long Short-Term Memory (BiLSTM) network to predict the next token to evaluate structural constraints via *perplexity* (a measure of how "confused" or uncertain the model is by a sequence; lower perplexity means the sequence appears more natural to the model's learned grammar).

We procedurally generated three datasets of synthetic inscriptions to test against the model:

1. **Concatenation A+B:** Two inscriptions with identical artifact types and similar lengths were joined.

2. **Prefix-Core-Suffix Swap:** The core of Inscription B was injected between the terminal signs of Inscription A.

3. **Suffix Replacement:** The final sign of A was replaced with the final sign of B.

4.8 Implementation Details

Visual features were extracted after binarization using Otsu's thresholding method and base/affix splitting (components occupying $> 30\%$ of the total glyph area were classified as base). All features were standardized with z-score normalization. Dimensionality reduction was performed with UMAP (10 components, 15 neighbors, min dist 0.1). Clustering used agglomerative clustering with complete linkage. The optimal number of clusters ($k = 8$) was selected by maximizing the silhouette score over $k = 2$ to 20, yielding a global silhouette score of 0.414. The BiLSTM model used an embedding dimension of 128 and hidden size of 256, trained for 10 epochs.

5 Results

Results reported in this section are computed on the Indus inscription corpus unless otherwise stated. Synthetic datasets and cross-system corpora are used only for baseline comparison.

5.1 Clustering Results

The application of HDBSCAN on the UMAP-reduced feature space yielded stable partitions. Figure 2 visualizes the resulting clusters. Figure 3 illustrates the network connectivity (cosine similarity) between these morphological families, and Figure 4 displays algorithmically generated cluster representatives. Several clusters show coherent visual themes (e.g., complex composite signs vs. simple linear strokes), suggesting that ancient scribes may have treated them as related categories despite surface variation.

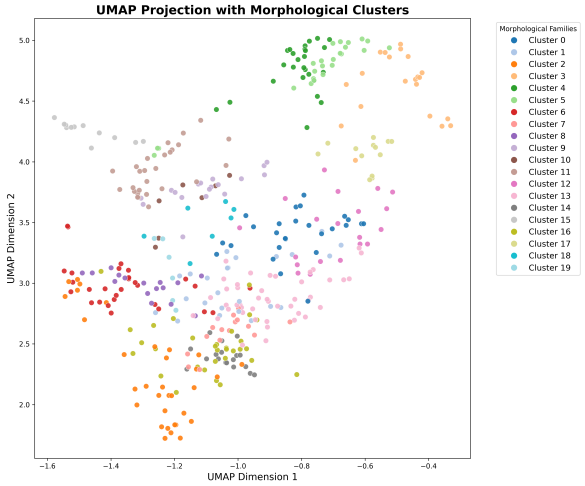


Figure 2: UMAP projection of visual glyph embeddings revealing morphological clusters discovered via HDBSCAN.

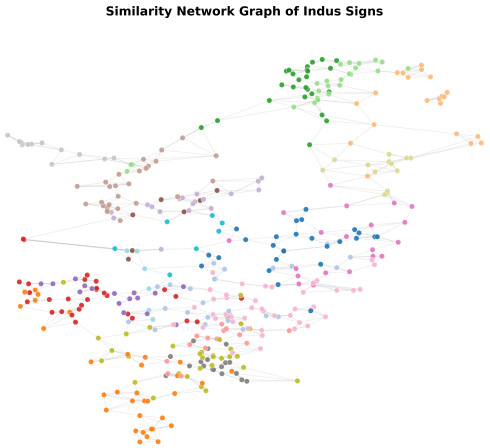


Figure 3: Similarity network graph of Indus script signs based on visual features.Edge thickness represents cosine similarity between morphological families.

5.2 Entropy and Cross-System Validation

The entropy calculations for the original and reversed Indus sequences, alongside comparative baselines, are summarized in Table 1. The original Indus sequence direction yields lower conditional entropy at higher Markov orders compared to its reversed counterpart, indicating a clear directional rule-set.

Metric	Indus (Orig)	Indus (Rev)	Linear B	Heraldic
Unigram (H_1)	6.58	6.58	5.59	3.78
Bigram (H_2)	3.60	4.11	4.43	3.73
Trigram (H_3)	1.42	1.61	2.12	3.45

Table 1: Comparison of conditional entropy decay across systems (measured in bits).

Table 2 presents the cross-system validation of Positional Rigidity (R_i). Linear B shows near-zero rigidity across positions, while the heraldic system exhibits strong rigidity at constrained boundary positions. Indus inscriptions show intermediate behavior. These positional constraints are visualized in Figure 5. While the average Indus sequence is 4-5 signs long, the rightward spike on the x-axis demonstrates that terminal signs remain highly constrained regardless of total sequence length.

System	R_0	R_1	R_2	R_3	R_4	R_5
Linear B	0.023	-0.008	0.009	0.012	0.011	0.014
Heraldic	1.779	-0.126	-0.016	0.071	0.301	1.780
Indus	0.082	0.110	0.050	0.140	0.220	0.740

Table 2: Cross-system comparison of Positional Rigidity (R_i) tracking constraint strengths across specific sequence indices.

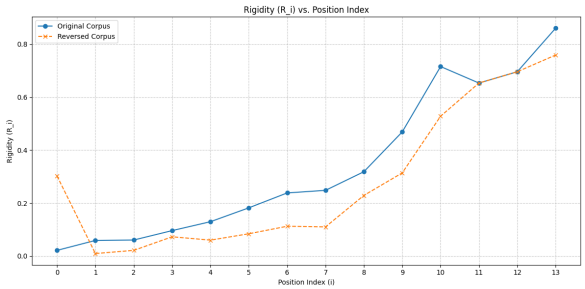


Figure 5: Positional Rigidity (R_i) plotted against absolute position index. The original Indus corpus exhibits terminal boundary constraints that distinguish it from random distribution.

5.3 KL Divergence Results

The KL divergence results comparing the forward and backward joint probability distributions are presented in Table 3.

Metric	Value (bits)
KL Divergence (Bigram)	16.45
KL Divergence (Trigram)	21.43

Table 3: Kullback-Leibler divergence between original and reversed joint distributions measuring statistical irreversibility.

5.4 Sequence Model Predictability

Based on BIC penalty functions, a 4-state HMM was optimal for the dataset. Figure 6 displays the proportional dominance of each hidden state by

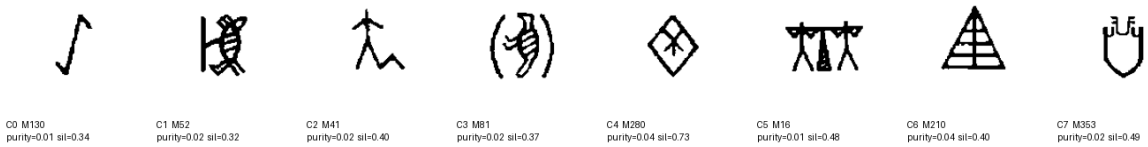


Figure 4: Cluster representatives extracted from the HDBSCAN partition. Labels indicate the standard Mahadevan ID alongside individual silhouette scores for each morphological prototype.

absolute sequence position. These latent states represent hidden functional categories (e.g., "start" and "stop" structural slots) within the script's syntax.

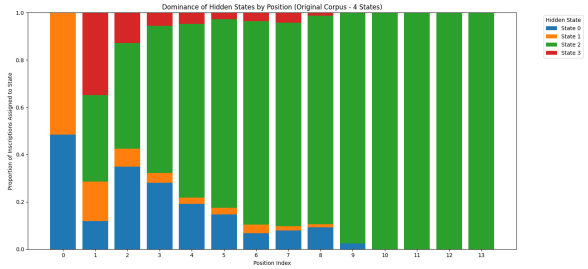


Figure 6: HMM state dominance by sequence position. State 1 dictates sequence beginnings, while State 4 dictates sequence terminals, confirming structured functional transitions.

The BiLSTM model results for synthetic sequence evaluation are presented in Table 4. The ΔPPL metric represents the ratio of synthetic mean perplexity to the real test set mean perplexity (2.60 Abs. PPL). The extremely small p-values ($< 10^{-12}$) indicate that the observed differences between real and synthetic sequence distributions are statistically robust. Figure 7 visualizes these perplexity distributions, and Figure 8 shows the gradient saliency map indicating where the model applies its "attention" during next-token prediction.

Dataset	Abs. PPL	ΔPPL	p-value	Cliff's δ
Real Test Set	2.60	1.00	–	–
Suffix Replace	1.10	0.42	6.53×10^{-13}	0.176
Concat A+B	1.39	0.53	2.65×10^{-177}	-0.698
Prefix-Core Swap	2.42	0.92	2.13×10^{-166}	-0.676

Table 4: Absolute and ratio perplexity scores alongside non-parametric statistical significance evaluations for synthetic sequence modifications.

6 Discussion

The results demonstrate directional structural constraints within the script. The lower conditional entropies for the original sequence orientation (Table 1), combined with the increasing KL divergence

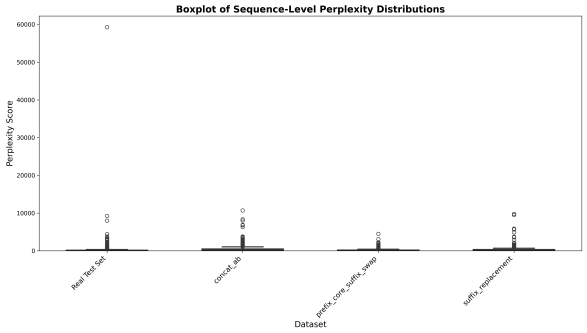


Figure 7: Boxplots of sequence-level perplexity distributions. The BiLSTM smoothly parses suffix replacements (tight low variance) while assigning higher variance to structural concatenations and core swaps.

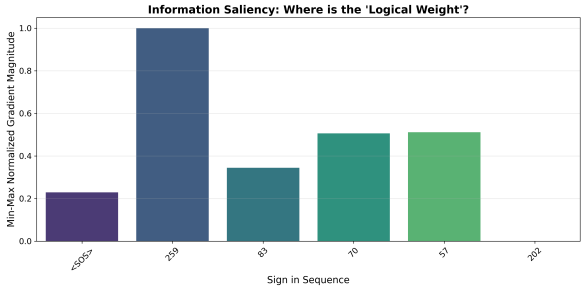


Figure 8: Gradient saliency map indicating specific token weights during sequence prediction. High min-max normalized gradient magnitude on terminal boundaries confirms the model's heavy reliance on positional constraints.

at higher Markov orders (Table 3), indicate statistical irreversibility. Following Sproat (2014), this study avoids interpreting entropy values alone as evidence of language and instead compares multiple symbolic systems.

The cross-system validation (Table 2) clarifies the nature of these constraints. Linear B shows near-zero rigidity, typical of natural languages where words can appear in multiple syntactic slots. The synthetic heraldic system exhibits strict rigidity at boundary positions. Indus inscriptions show intermediate behavior. The elevated rigidity at the final position in the Indus corpus ($R_5 = 0.740$) suggests that terminal signs may function as structural delimiters or positional markers.

The sequence representation model assigned lower probabilities (higher perplexity variance) to concatenated sequences and core-swapped sequences. However, the model assigned higher relative probabilities to sequences where terminal signs were replaced (Figure 7). This response pattern indicates combinatorial constraints, where terminal elements exhibit structural modularity while core elements maintain sequential rigidity.

7 Conclusion

This study evaluated the structural properties of the Indus script using a multimodal computational framework. The central question was whether the statistical structure of Indus sign sequences more closely resembles that of linguistic writing systems or structured non-linguistic symbol systems. Regarding the specific research questions: (1) the script exhibits directional asymmetry, supported by conditional entropy and KL divergence metrics; (2) visual clustering of glyphs yields stable categories independent of human cataloging; and (3) neural sequence modeling detects internal ordering constraints, accepting terminal variations while assigning lower probabilities to arbitrary structural concatenations. These findings indicate that the Indus sign sequences exhibit statistical properties consistent with structured symbolic systems and not easily explained by random generation.

8 Limitations

This analysis cannot determine whether the Indus inscriptions encode language. Instead, it characterizes their statistical structure relative to other symbolic systems. A limitation of this study is the brevity of the source artifacts. Short sequence length limits higher-order syntactic inference, but information-theoretic metrics remain informative for detecting directional constraints. Additionally, because the script remains undeciphered, the functional categories identified by statistical clustering cannot currently be verified against semantic or phonetic meanings.

Data Availability

The digitized sign corpus used in this study is derived from publicly available Indus sign lists and previously published corpora. The processed dataset and analysis scripts will be made available in an open repository upon publication.

References

Steve Farmer, Richard Sproat, and Michael Witzel. 2004. The collapse of the Indus-script thesis: The myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies*, 11(2):19–57.

G. R. Hunter. 1934. *The Script of Harappa and Mohenjodaro and Its Connection with Other Scripts*. Kegan Paul, Trench, Trubner & Co.

Yuri V. Knorozov. 1965. *Predvaritel'noe soobshchenie ob issledovanii protoindiyskikh tekstov*. Nauka.

Yuri V. Knorozov, M. F. Albedil, and B. Y. Volchok. 1979. *Proto-Indica: 1979. Report on the Investigation of the Proto-Indian Texts*. Nauka.

Iravatham Mahadevan. 1977. *The Indus Script: Texts, Concordance and Tables*. Archaeological Survey of India.

Bahata Ansumali Mukhopadhyay. 2023. Semantic scope of Indus inscriptions comprising taxation, trade and craft licensing. *Humanities and Social Sciences Communications*, 10.

Asko Parpola. 1994. *Deciphering the Indus Script*. Cambridge University Press.

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165–1167.

Richard Sproat. 2014. Statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481.

Bryan K. Wells. 2011. *Epigraphic Approaches to Indus Writing*. Oxbow Books.

Yajnadevam. 2024. A cryptanalytic decipherment of the Indus script. *Preprint*.

Exploring Topological Invariance in Semantic Embeddings

Fangzhou Gao and Justin Brody
Franklin and Marshall College
PO Box 3003
Lancaster, PA 17604
fgao@fandm.edu justin.brody@fandm.edu

Abstract

We present the result of preliminary explorations of using the topology of embedded manifolds as a semantic invariant. Our main question is whether the topology of large embedded corpora is invariant in the following two senses. First, one might reasonably expect that the same corpus in two languages would give topologically equivalent embeddings. Second, one might reasonably expect that the same corpus embedded by two different embedding models might give topologically equivalent embeddings. In the paper we will justify these intuitions and give preliminary results indicating that they are, to some extent, justified.

1 Introduction

In this paper, we begin an exploration of the possibility of using the tools of topological data analysis to uncover a *semantic signature* for a corpus. Contemporary natural language processing allows for a multitude of ways to embed a corpus into a vector space. A single text can be embedded by multiple models; similarly a single text can exist in multiple languages, and each language can be embedded by multiple models.

For reasons explained in the next section, we hypothesize that the topology of the structure underlying such embeddings provides such a semantic signature. If this hypothesis is correct, then we can use the tools of topological data analysis to uncover the semantic signature of a corpus. In particular, an accurate translation of a text should have the same semantic signature as the original text and persistent homology can be a way to measure translation accuracy.

In our paper we will report on a preliminary exploration of these hypotheses. We will find some promising signs for our hypotheses, but will not fully confirm them.

In particular, we will look at a set of web novels in Mandarin Chinese, along with translations into English and German. These are embedded via a single multi-modal model and we will compare the persistent homology of the resulting manifolds.

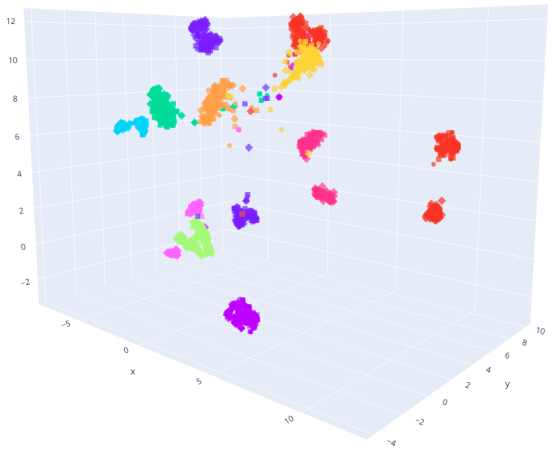


Figure 1: Embeddings of chapters. Regions with similar colors correspond to embeddings of the same chapters in different languages. The shapes of these embeddings are often very similar.

2 Preliminaries

Our fundamental work here is with embeddings of corpora. Transformer neural networks can fundamentally be thought of as embedding architectures. In particular, encoder architectures (like BERT) provide methods of transforming words into vectors in such a way that the embeddings absorb the context of the other words in the sentence, while a special CLS token can be thought of as absorbing the context of the entire sentence. These embedded sentences are our fundamental objects of study.

The intuition behind these embeddings of sentences into vector spaces is that sentences with similar meanings will be embedded close to each other in the vector space (according to some metric). This is the basis of using sentence embeddings for topic modelling; for example this is the basis of BERTopic [4]. Sentences with similar meanings

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 320–324
July 4, 2026 ©2026 Association for Computational Linguistics

will be embedded close to each other and form clusters which correspond naturally to topics.

We take this as our foundation – that high quality embedding models will embed sentences with similar meanings close to each other. From just this assumption, what can we say about the embeddings produced by a corpus? We certainly have no right to assume that two different embeddings of the same corpus will have the same geometry [1]. However, we might expect that they will have the same topology, since topological properties are precisely those which are determined from flexible notions of "closeness". In particular, given a particular corpus C and another version of that corpus D (say in a different language), we might expect a high quality embedding model to embed C and D in ways which produce topologically similar shapes. If a model maps sentences to a universal language before embedding them, then we might expect to see much stronger relations between the embeddings of C and D, with identical embeddings being the strongest possibility and embeddings with the same shape being a slightly weaker possibility. We might further expect two different high quality embedding models to embed C in different but topologically similar ways. These observations lead us to conjecture that different embeddings with the same underlying meaning should have the same topological structure.

Measuring topological structure presents a number of difficulties which the tools of *topological data analysis* (TDA) are designed to mitigate [2]. We assume that the embeddings of a corpus form a discrete sample from some underlying continuous manifold, which we term the *semantic manifold*. The tools of TDA are then designed to uncover this manifold and provide a signature of it. In particular, we will use persistent homology to measure the topological structure of the manifolds corresponding to different embeddings of the same corpus. We will then use the bottleneck distance to measure the similarity of the persistent homology of different embeddings.

One of the early challenges in topology was to find a way to characterize the shape of a manifold. It turns out [5] that one way to do this is to essentially look at the holes in a manifold. Intuitively, we can think of circles (and their higher dimensional analogues – spheres) as being filled or unfilled depending on whether or not they can be shrunk down to a point. A canoncial example is a circle that is the boundary of a disk. The disk can be continuously shrunk down to a point without obstruction, so this circle would not describe any kind of unfilled hole. On the other hand, a circle that is the boundary of an annulus cannot be shrunk down to a point, so this circle would describe a unfilled hole. Homology groups are then ways of describing a manifold by giving a canonical count of the number of unfilled holes in various dimensions – counting the number of fundamental [2] unfilled circles (or 1-dimensional spheres) in a manifold M gives $H_1(M)$, the first homology group of M.

Standard homology requires a fixed, continuous shape (a manifold). In our context, we only have a finite sample of discrete data points (a point cloud) and need to infer the underlying manifold. Persistent homology solves this by creating a *filtration of simplicial complexes* from the point cloud. The idea is connect close-by points and try to estimate whether they are connected as part of a line (unfilled), a triangle (filled) or higher dimensional analogue of a triangle. Specifically, for each r we place a sphere of radius r around each point and looking at how these spheres overlap. If two points have overlapping spheres, they are assumed to lie on the same line segment , if three points have overlapping spheres they are assumed to lie on the same triangle, and so on.

When $r = 0$, the filtration simply produces isolated points. As r increases, the spheres begin to overlap and edges will start to appear. As r continues to grow, unfilled holes will form in this complex; this is the "birth" of an unfilled hole. As r grows even larger, those holes will eventually get become filled in by the expanding spheres; this is the "death" of an unfilled hole.

Persistent homology is the study of tracking the birth and death of these topological features (connected components, loops, voids) across all possible scales of r. The main idea is that features that persist over a large range of r represent true, underlying signal, while features that die quickly are topological noise.

[1] Intuitively, that would require that if S, T are similar sentences, then not only will two high quality models embed S and T close to each other, but will also embed them at similar angles from each other.

[2] Fundamental in the sense that no other circle is "homotopy-equivalent" to it. A discussion of this term is beyond the scope of this paper.

3 Methodology

We use the GuoFeng Webnovel corpus [10, 9], a large-scale multilingual collection of web fiction spanning multiple languages and genres. Following the dataset release structure, we use Chinese (ZH) and English (EN) data from V1 and German (DE) from V2. As the DE portion covers fewer books than ZH/EN, all cross-lingual experiments are restricted to the *intersection* of book_ids available across the relevant languages, ensuring equal book coverage across pairs. We report intersection sizes for each language pair alongside the corresponding evaluation results.

For EN and ZH, the raw source files contain explicit structural markup for the book and chapter boundaries (e.g., <BOOK id="...">, <CHAPTER id="...">). We parse these tags to assign book_id and chapter_id metadata to each line, strip the markup, and remove chapter-title lines (e.g., lines immediately following a <CHAPTER ...> tag that contain chapter headings). For ZH, we additionally remove dataset-specific dialogue markers when the marker pattern is detected.

We perform sentence segmentation using **Stanza** [6] with sentence splitting enabled for all three languages. Each segmented sentence is stored alongside its metadata (lang, book_id, chapter_id, and within-line sentence index).

We embed all sentences using a single multilingual model, enabling all three languages to be represented in a shared vector space. We use **BAAI/bge-m3** [3] via the SentenceTransformers library [7] and produce L2-normalized embeddings. The resulting vectors are stored as a matrix $E \in \mathbb{R}^{N \times d}$, where N is the total number of sentences and d is the embedding dimension, aligned with the sentence metadata table.

To obtain a more stable unit of analysis than individual sentences, which are short and noisy in web fiction, we compute chapter embeddings by mean-pooling sentence embeddings within each (book_id, chapter_id) group:

$$c_{b,k} = \frac{1}{|S_{b,k}|} \sum_{s \in S_{b,k}} e_s, \tag{1}$$

where $S_{b,k}$ is the set of sentences in chapter k of book b and e_s is the embedding of sentence s. We L2-normalize $c_{b,k}$ after pooling. Chapters with fewer than τ sentences are excluded, where τ is swept as part of a robustness analysis. Each book b in language ℓ is then represented as a point cloud of chapter vectors:

$$X_{b,\ell} = \{c_{b,k}\}_{k=1}^{K_b}. \tag{2}$$

Although multilingual embedding spaces can preserve cross-lingual structure, aggregated chapter and book embeddings exhibit a strong global *language offset*. To correct for this, we apply language centering by subtracting a per-language centroid:

$$\tilde{x} = x - \mu_\ell, \qquad \mu_\ell = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} x_i, \tag{3}$$

where the sum ranges over all vectors in language ℓ belonging to the common-book intersection, ensuring comparability across languages. Unless otherwise noted, all analyses use centered embeddings.

To isolate the respective contributions of language and content to embedding similarity, we evaluate a 2×2 factorial design at the chapter level, crossing two factors: (i) Language: same (**SL**) vs. different (**DL**) , and (ii) Content: same book (**SC**) vs. different books (**DC**). This produces the four conditions SL–SC, SL–DC, DL–SC, DL–DC.

To characterize higher-order geometric structure in each book's chapter point cloud beyond mean similarity, we compute Vietoris–Rips persistent homology on $X_{b,\ell}$ under the Euclidean metric using **Ripser** [1]. We compute persistence diagrams up to homology dimension 1, capturing connected components (H_0) and loops (H_1). To keep computation tractable, we cap the Vietoris–Rips filtration at a distance threshold ϵ, which is swept as part of a robustness analysis.

We quantify the dissimilarity between two books (b, ℓ) and (b', ℓ') in homology dimension d as:

$$\delta^{(d)}_{(b,\ell),(b',\ell')} = \Delta\left(X^{(d)}_{b,\ell},\, X^{(d)}_{b',\ell'}\right), \tag{4}$$

where Δ is the Wasserstein distance between persistence diagrams, computed with **Persim** [8]. For the SL–SC baseline, we estimate within-book variability by drawing two independent subsamples of chapters from the same $X_{b,\ell}$ and computing $\delta^{(d)}$ between the resulting diagrams.

To validate that our findings are not artifacts of topological choices, we compute two non-topological metrics over the same 2×2 conditions.

For each language, we compute the pairwise cosine distance matrix among mean-pooled book embeddings. We then measure the agreement between

two languages by computing the Spearman correlation between the vectorized upper triangles of their respective distance matrices. A high correlation indicates that the inter-book similarity structure is preserved across languages independently of embedding geometry.

For each book embedding in language ℓ, we retrieve its nearest neighbor among book embeddings in language ℓ' and report Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR). We evaluate retrieval for both raw and centered embeddings, treating the difference as an ablation of the centering step.

4 Results

In this section, we present our findings on the preservation of structural and topological properties across languages in the GuoFeng Webnovel corpus. We first establish the alignment of the global embedding spaces using non-topological validation metrics, demonstrating the significant impact of language centering. We then present the results of our persistent homology analysis to evaluate whether higher-order geometric structures align with narrative content across language barriers.

To validate that our multilingual embeddings preserve inter-book relationships independently of topology, we analyzed the pairwise cosine distance matrices of mean-pooled book embeddings. The structural agreement across languages is exceptionally high and further improves after applying the language centering described in Section 3. By computing the Spearman correlation (ρ) between the flattened upper triangles of these distance matrices, we observed strong structural preservation, as detailed in Table 1.

Language Pair	Raw (ρ)	Centered (ρ)
EN vs. ZH	0.870	0.883
EN vs. DE	0.903	0.940
ZH vs. DE	0.854	0.869

Table 1: Spearman correlation of inter-book distance matrices across language pairs, comparing raw and centered embeddings.

This cross-lingual alignment is also dramatically reflected in the "same-book" retrieval performance. When retrieving the same book across languages using nearest-neighbor search, raw embeddings struggle with distant language pairs due to the global language offset. However, subtracting the per-language centroid effectively aligns the spaces. For EN $\rightarrow$ ZH, Top-1 retrieval accuracy surges from 0.382 to 0.974, with Top-5 reaching 1.000 and a Mean Reciprocal Rank (MRR) of approximately 0.983. For the ZH $\rightarrow$ DE and DE $\rightarrow$ ZH directions, Top-1 accuracy reaches a perfect 1.000 after centering. Notably, the EN $\leftrightarrow$ DE pair, being typologically closer, already achieved a Top-1 accuracy of 1.000 prior to centering.

To determine whether higher-order geometric features are preserved across translations, we evaluated our 2×2 factorial design using Vietoris–Rips persistent homology on the chapter point clouds ($X_{b,\ell}$). Following the methodology in Section 3, we report the Wasserstein distance (Δ) between persistence diagrams in the H_1 homology dimension using centered embeddings.

As hypothesized, the median distances between persistence diagrams follow a strict ordering that confirms topological alignment with content over language. The results for the four experimental conditions are presented in Table 2.

	Content	
Language	**Same (SC)**	**Different (DC)**
Same (SL)	3.36	5.40
Different (DL)	5.93	6.44

Table 2: Median Wasserstein distances (H_1) between chapter point cloud persistence diagrams across the 2×2 experimental conditions.

5 Conclusion

Crucially, we find that **DL–SC** $<$ **DL–DC** ($5.93 < 6.44$). This indicates that, after accounting for language offsets, the topological representation of the same book across different languages is significantly closer than the representations of different books across those same languages. The overall ordering (**SL–SC** $<$ **SL–DC** $<$ **DL–SC** $<$ **DL–DC**) demonstrates that while language still exerts a measurable influence on the topological shape of the point clouds (**SL–SC** $<$ **DL–SC**), the underlying narrative content preserves a distinct geometric signature that persists across translations. However, these results do not indicate that we can distinguish content based solely on topological similarity, since **SL–DC < DL–SC**. We are hopeful that further experimentation may reveal ways of associating topological structure more strongly with content.

References

[1] Ulrich Bauer. 2021. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423.

[2] Frédéric Chazal and Bertrand Michel. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963.

[3] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

[4] Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[5] J.R. Munkres. 1975. *Topology: a first course*. Prentice-Hall, New Jersey.

[6] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

[8] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. 2018. Ripser.py: A lean persistent homology library for python. In *The Journal of Open Source Software*.

[9] Longyue Wang, Siyou Liu, Minghao Wu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Liting Zhou, Yan Gu, Weiyu Chen, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the wmt 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*.

[10] Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, and 1 others. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67.

MADRAG: Multi-Agent Debate with Retrieval-Augmented Generation for Training-Free Analytic Essay Scoring

Ali Keramati **Shiyuan Zhou** **Sharad Mehrotra** **Mark Warschauer**
University of California, Irvine
a.kera@uci.edu, szhou20@uci.edu, sharad@ics.uci.edu, markw@uci.edu

Abstract

We present MADRAG, a training-free framework for analytic essay scoring that combines multi-agent reasoning with retrieval-augmented grounding. Unlike standard LLM-as-judge approaches, which are prone to bias and unstable scoring, MADRAG decomposes evaluation into an interactive process: an Advocate identifies strengths, a Skeptic critiques weaknesses, and a Judge aggregates their arguments into a final score. Crucially, the Judge is augmented with rubric-aligned exemplar retrieval, enabling calibration through comparison with scored examples. Our results show that MADRAG significantly outperforms prompt-based baselines while approaching the performance of supervised systems without requiring task-specific training. Ablation studies demonstrate that retrieval drives calibration gains, while debate improves reasoning on higher-level traits. Our findings highlight the complementary roles of structured interaction and external memory in reliable LLM-based evaluation.

1 Introduction

Assessing student writing is labour-intensive and often inconsistent across raters. In large educational settings, teachers must score many essays under tight time constraints, leading to fatigue, delayed feedback, and imperfect reliability (Ramesh and Sanampudi, 2021). Even when essays are double-scored, inter-rater agreement remains limited: analyses of the Automated Student Assessment Prize (ASAP) dataset show that trained raters frequently disagree by more than one score point on individual traits (Crossley et al., 2025). These challenges motivate AES systems that aim to approximate human judgments at scale.

Early AES approaches relied on hand-crafted features such as word counts and readability metrics (Page, 1966; Attali and Burstein, 2006), followed by neural models including recurrent, convolutional, and transformer-based architectures (Taghipour and Ng, 2016; Dong et al., 2017; Wang et al., 2022). While these systems can produce reliable scores, most deployed tools output a single holistic score, offering little actionable feedback for instruction (Warschauer and Ware, 2006). In contrast, writing teachers typically prefer analytic trait scoring, which provides targeted feedback on dimensions such as ideas, organization, and conventions (Knoch, 2009). An ideal AES system should therefore produce accurate, transparent, and reliable *trait-level* scores. Large language models (LLMs) have recently enabled training-free, prompt-based scoring across diverse rubrics (Fallah et al., 2024). However, prior work shows that direct LLM judging is often poorly calibrated, sensitive to prompt design, and prone to systematic biases (Mansour et al., 2024). In particular, LLM-as-judge tends to regress toward the middle of the scoring scale, failing to distinguish clearly between exceptional and weak inputs (Zheng et al., 2023). These issues are exacerbated by the multi-step, multi-trait nature of rubric-based scoring, which requires maintaining and coordinating multiple criteria and score ranges within a single judgment (Valmeekam et al., 2023). As a result, naïvely applying LLMs as essay graders yields inconsistent trait scores and misjudgments of extreme cases.

In this work, we ask whether a fully *training-free* LLM-based system can achieve trait-level scoring reliability comparable to human raters and strong supervised AES models. We propose **MADRAG** (Multi-Agent Debate with Retrieval-Augmented Generation), a framework that combines two complementary mechanisms. First, multi-agent debate (MAD): an Advocate highlights strengths, a Skeptic critiques weaknesses, and a Judge synthesizes their arguments to produce a score. Second, retrieval-augmented generation (RAG): before scoring, the Judge retrieves rubric-aligned exemplar essays spanning the full score range to ground and calibrate its decision. By integrating debate with

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 325–345
July 4, 2026 ©2026 Association for Computational Linguistics

trait-specific retrieval, MADRAG provides external memory, encourages explicit comparison against exemplars, and mitigates middle-score bias.

2 Related Work

2.1 Supervised Analytic Trait Scoring

Early AES systems primarily focused on holistic scoring using hand-crafted features such as word counts and readability measures (Page, 1966; Attali and Burstein, 2006). While effective for large-scale testing, holistic scores provide limited diagnostic value for formative assessment, motivating a shift toward analytic trait scoring that evaluates dimensions such as content, style, and organization separately (Warschauer and Ware, 2006; Deane, 2013). Neural approaches enabled this transition by modeling essays as structured representations amenable to multi-trait prediction. Mathias and Bhattacharyya (2020) pioneered attention-based architectures for analytic scoring, showing that trait-specific predictions better support instructional feedback. Subsequent work has focused on improving robustness in data-scarce and cross-prompt settings. ProTACT (Do et al., 2023), for example, introduces prompt-aware representations and a trait-similarity objective to exploit correlations among rubric dimensions, achieving strong performance on the ASAP dataset. Others explore multi-task transformers with trait-specific heads (Kumar et al., 2022) or reinforcement learning objectives to refine trait-level accuracy (Do et al., 2024). Despite their reliability, supervised trait scorers require labeled data for each new prompt and rubric, limiting their practicality in real-world classrooms where assignments and criteria change frequently.

2.2 Training-Free LLM Scoring

LLMs offer an appealing alternative by enabling training-free, prompt-based essay scoring that can be applied across prompts and rubrics without task-specific fine-tuning (Kojima et al., 2022). However, empirical studies consistently find that direct zero-shot scoring lags behind supervised AES systems in both accuracy and reliability. Mansour et al. (2024) show that even with careful prompt engineering and one-shot examples, models such as ChatGPT and LLaMA substantially underperform supervised baselines. More structured prompting strategies improve performance but expose additional limitations. Tang et al. (2024) demonstrate that rubric-aligned exemplars and justification prompts can raise agreement on abstract traits like *Ideas*, yet performance remains highly sensitive to decoding parameters and degrades sharply on surface-level traits such as *Conventions*. Related analyses reveal systematic biases, including harshness on complex traits (Kundu and Barbosa, 2024), length bias, and central tendency bias, where models avoid extreme scores and regress toward the middle of the scale (Li et al., 2025). One line of work addresses calibration by reformulating scoring as a comparative task. LCES (Shibata and Miyamura, 2025) replaces absolute scoring with pairwise ranking, yielding more stable judgments. However, its quadratic complexity makes it difficult to scale to classroom-sized datasets, highlighting the need for training-free approaches that retain absolute scoring while improving calibration and extreme-score discrimination.

2.3 Multi-Agent Debate and Orchestration

To mitigate the limitations of single LLM judges, recent work has explored multi-agent frameworks in which multiple models collaborate or debate to reach a consensus. Debate has been shown to improve reasoning quality and factual accuracy by encouraging agents to critique and refine each other's arguments (Liang et al., 2024; Du et al., 2024). ChatEval (Chan et al., 2023) successfully applied this to text evaluation, showing that a "jury" of LLMs correlates better with human judgments than a single score. In the domain of AES, MAGIC (Jordán et al., 2025) applies this by assigning specialized agents to different rubric traits (e.g., a "Grammar Expert" and "Organization Expert") and synthesizing their outputs via an orchestrator, achieving substantial gains over single-agent baselines. Similarly, CAFES (Su et al., 2025) utilizes a reflective workflow where an initial scorer revises its judgments based on feedback from a "critic" agent. These systems underscore the potential of decomposing the scoring task, yet they often rely on static agent roles or lack access to external knowledge, which can limit their ability to ground scores in concrete evidence. Taken together, prior work reveals a persistent trade-off: supervised AES models deliver reliable trait-level scores but lack flexibility, while training-free LLM judges offer adaptability at the cost of calibration and consistency. Multi-agent systems partially bridge this gap, yet they remain vulnerable to groupthink and poorly grounded reasoning (Wu et al., 2023).

3 Methodology

3.1 Problem Setting and Notation

Let $\mathcal{E}$ denote the collection of student essays and $\mathcal{R}$ the set of rubric traits. Each essay $e \in \mathcal{E}$ consists of unstructured text and associated metadata (e.g., an identifier). Each rubric trait $r \in \mathcal{R}$ is represented by a structured object with fields including a name, minimum and maximum scores, and a description of the trait to evaluate. For a given essay e and trait r, our goal is to produce a numeric score $s(e, r)$ reflecting how well the essay satisfies the trait, along with a rationale.

3.2 Agents and Roles

We assign specialized roles to agents (Appendix A.6): the Advocate speaks first, the Skeptic responds, and the Judge synthesizes both contributions together with agent confidence scores. Each agent's confidence is approximated by the log-probability of the first token in its message. Although coarse, this signal reflects the model's conditional belief over continuations at the start of the response, and similar token-probability signals have been used as lightweight confidence indicators in prior LLM-based annotation and evaluation frameworks (M. M. Hicke et al., 2025; Lin and Hooi, 2025; Kadavath et al., 2022). We use the first-token probability as a simple and stable proxy that is less sensitive to response length or verbosity than sequence-level aggregations Appendix A.4.

The **Supervisor** coordinates the full evaluation process by decomposing the rubric into constituent traits, retrieving few-shot examples for each trait, instantiating debate agents and judges, and orchestrating execution. For each trait, the **Advocate** receives the essay and rubric description and produces a single opening argument emphasizing strengths only, without assigning a score or discussing weaknesses. The **Skeptic** then reads the Advocate's statement and the rubric, and produces a single rebuttal focused exclusively on shortcomings, without assigning a score or mentioning strengths. Finally, the **Judge** acts as an impartial arbiter, reading the Advocate and Skeptic messages, their confidence values, the few-shot examples, and the rubric trait to produce a final integer score.

3.3 Retrieval-Augmented Few-Shot Example Generation

In our setting, retrieval is used to provide previously scored essays as few shot calibration references (Appendix A.7), allowing judges to align its scoring decisions with examples that span the rubric's scoring range (Lewis et al., 2021).

3.3.1 Vector Database and Embeddings

To support retrieval augmented exemplar construction, we construct a vector database of scored essays. Each essay $e \in \mathcal{E}$ is embedded as a dense vector representation using all-MiniLM-L6-v2, a sentence-transformer embedding model $\phi(\cdot)$ (Reimers and Gurevych, 2019). It maps the full essay text into a fixed-dimensional vector $\mathbf{z} = \phi(e)$ optimized for semantic similarity search. These embeddings are stored in a Chroma vector database together with structured metadata. Formally, the vector database is defined as

$$\mathcal{V} = \{(\mathbf{z}, m) \mid \mathbf{z} = \phi(e),\ e \in \mathcal{E}\},$$

m is a metadata dictionary associated with essay e which contains the raw essay text, the overall domain score, and a discrete score for each rubric trait. When multiple human raters provide scores for a given trait, their scores are aggregated by averaging and rounding to the nearest integer.

3.3.2 Retrieval Procedures

Given a query essay $e \in \mathcal{E}$, retrieval is performed by embedding the essay using the same encoder $\phi(\cdot)$ and searching the vector database $\mathcal{V}$ for relevant few-shot exemplars F. We define two retrieval procedures. The first procedure retrieves essays that are semantically similar to the query essay. Specifically, the query essay e is embedded as $\phi(e)$, and a nearest-neighbor search is performed in $\mathcal{V}$ to retrieve the top k essays with highest similarity in the embedding space which are used as few-shot examples. The second procedure retrieves exemplars conditioned on rubric scores for a specific trait. For a given trait $r \in \mathcal{R}$, we retrieve one exemplar essay for each score value within the valid score range. For each s, we perform nearest-neighbor search in $\mathcal{V}$ using the query embedding $\phi(e)$ while filtering candidates to essays whose metadata score for trait r equals s. When multiple candidates are available, the nearest remaining essay is selected. If no suitable essay exists for a given score value, the system records that no exemplar is available for that score.

3.4 MADRAG Workflow

Figure 1 provides an overview of MADRAG, illustrating how trait-wise debate is combined with

retrieval-augmented exemplars within a single scoring pipeline. We formalize the trait-level procedure in Algorithm 1, where retrieved exemplars augment the Judge alongside the Advocate–Skeptic debate transcript. Additional implementation details are provided in Appendix A.

Debate dynamics. For essay e and rubric trait r_i with score range $[m_i, M_i]$, the shared input context:

$$x_i(e) \;=\; \langle q,\; e,\; r_i,\; [m_i, M_i],\; \alpha \rangle$$

where q denotes the essay prompt, and α collects fixed inference-time settings shared across agents.

Advocate generation. Let G_A denote the Advocate policy (LLM with role prompt). The Advocate emits an argument a_i:

$$a_i \sim p_A(\cdot \mid x_i(e)).$$

We define an *internal confidence proxy* via the log-probability of the first emitted token $t^A_{i,1}$:

$$\ell^A_i \;=\; \log p_A\big(t^A_{i,1} \mid x_i(e)\big)\,,\; c^A_i \;=\; \exp(\ell^A_i).$$

Skeptic rebuttal. The Skeptic policy G_K conditions on the Advocate's message:

$$k_i \sim p_K(\cdot \mid x_i(e), a_i),$$

with first-token confidence

$$\ell^K_i \;=\; \log p_K\big(t^K_{i,1} \mid x_i(e), a_i\big)\,,\; c^K_i \;=\; \exp(\ell^K_i).$$

Debate transcript. The debate trace for trait r_i is the ordered pair

$$\tau_i(e) \;=\; (a_i,\; k_i).$$

Confidence-aware judging. The judge J produces a discrete score by maximizing a confidence-conditioned posterior over valid integers:

$$\mathcal{I}_i(e) \;=\; \big(x_i(e),\, \mathrm{RAG}(e, r_i),\, \tau_i(e),\, c^A_i,\, c^K_i\big),$$
$$s_i(e) = \arg\max_s \; p_J\Big(s \;\Big|\; \mathcal{I}_i(e)\Big)$$

Algorithm 1 Evaluate a rubric trait via MADRAG

function EVALUATETRAIT(e, r_i, q, α)
  $x \leftarrow \langle q,\; e,\; r_i,\; [m_i, M_i],\; \alpha \rangle$
  $F \leftarrow \mathrm{RAG}(e, r_i)$
  $(a, \ell^A) \leftarrow \textsc{Advocate}(x)\ \{a \sim p_A\}$
  $c^A \leftarrow \exp(\ell^A)$
  $(k, \ell^K) \leftarrow \textsc{Skeptic}(x, a)\ \{k \sim p_K\}$
  $c^K \leftarrow \exp(\ell^K)$
  $\tau \leftarrow (a, k)\ \{\tau \equiv \tau_i(e)\}$
  $\hat{s} \leftarrow \textsc{Judge}(x, F, \tau, c^A, c^K)$
  return $(\hat{s}, a, c^A, k, c^K)$
end function=0

4 Experiments

We evaluate MADRAG from three complementary perspectives. First, we measure overall analytic trait scoring performance against both supervised AES systems and prior training-free LLM baselines to assess whether MADRAG is competitive without task-specific training (RQ1). Second, we examine whether MADRAG reduces the middle-score bias commonly observed in LLM-as-judge settings, with particular attention to performance on essays at the low and high ends of the score range (RQ2). Third, we conduct a qualitative error analysis to understand the dominant failure modes of the framework and the contributions of debate and retrieval to those errors (RQ3).

4.1 Experimental Setup

Data. We evaluate MADRAG on the ASAP [1] dataset, a widely used benchmark of student-written English essays scored by trained human raters using prompt-specific rubrics. ASAP consists of eight essay sets, but analytic trait annotations in the original release are available only for Essay Sets 7 and 8. Accordingly, all experiments in this paper are conducted on Sets 7 and 8, which provide multiple independent human ratings per essay at the trait level. Detailed dataset statistics, prompts, transcription procedures, and label construction are provided in Appendix B.

Evaluation Metrics. We evaluate trait-level scoring using Quadratic Weighted Kappa (QWK), a standard agreement metric in AES that accounts for ordinal score distances.

LLMs. MADRAG is evaluated with multiple LLM backbones (GPT-4o-mini, GPT-4o, GPT-5-mini, and GPT-5) using the same role prompts across models (OpenAI, 2024). To reduce run-to-run variance, we decode the Judge deterministically (temperature $= 0$) and use high-temperature decoding for the Advocate and Skeptic; when supported, we also log token-level log-probabilities to compute confidence proxies (Appendix A). During scoring, the Judge is augmented with retrieved,

[1] https://www.kaggle.com/c/asap-aes/data

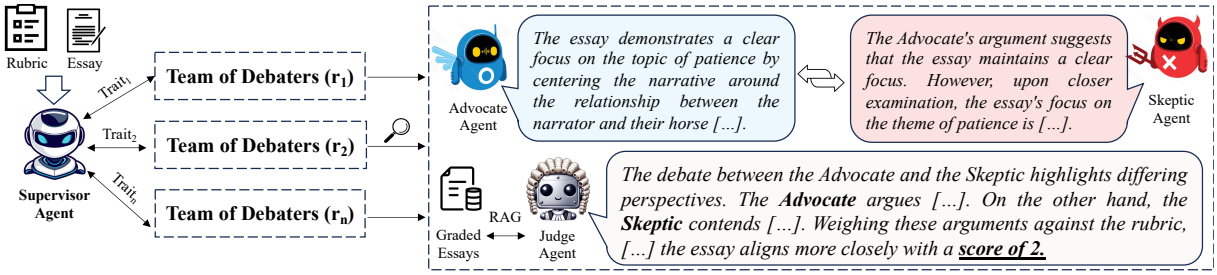


Figure 1: **Overview of the MADRAG scoring pipeline** The Supervisor routes each rubric trait to a dedicated debate team, retrieves few-shot exemplars to augment the Judge, and the Judge aggregates the Advocate and Skeptic exchanges together to produce the final trait score. The full sequence of agent messages is provided in Appendix A.8.

rubric-aligned exemplars spanning the trait's score range, following the retrieval procedure.

Baselines. We compare MADRAG against a diverse set of strong baselines spanning supervised and training-free paradigms. These include *training-based* neural models for analytic trait scoring—FeatEng-RF (Mathias and Bhattacharyya, 2020), and ProTACT (Do et al., 2023)—as well as *training-free*, prompt-engineered LLM scorers, including ZS-LLM (Mansour et al., 2024) and CSR-J (Tang et al., 2024). We additionally report Human–Human agreement as a reference ceiling. To improve readability, we defer detailed descriptions of each baseline's methodology, training regime, and evaluation protocol to Appendix C.

4.2 Comparative Performance on Analytic Traits (RQ1)

Table 1 reports trait-wise QWK on ASAP Essay Sets 7 and 8, comparing MADRAG against supervised and training-free baselines. Three main findings emerge. First, MADRAG substantially outperforms all prior training-free approaches. Second, despite requiring no labeled training data, MADRAG achieves performance competitive with strong supervised systems. Third, gains are highly trait-dependent, with the largest improvements observed on discourse-oriented traits.

Across both essay sets, MADRAG consistently improves over prior training-free LLM baselines under backbone-matched comparisons. In particular, MADRAG with `gpt-3.5-turbo` outperforms ZS-LLM on all reported traits, showing that the gains cannot be explained solely by using a stronger underlying model. Likewise, MADRAG with various GPT-4 family models consistently outperforms CSR-J, indicating that the proposed debate-and-retrieval framework provides benefits beyond prompt engineering alone. Taken together, these results suggest that MADRAG's gains arise from its structured reasoning and calibration mechanisms, rather than from backbone strength alone. More strikingly, MADRAG is competitive with supervised models that rely on thousands of labeled training examples. On *Ideas*, MADRAG (GPT-5) exceeds ProTACT by 50% in Set 7 (0.75 vs. 0.50) and by 18% in Set 8 (0.67 vs. 0.57). It also surpasses Human–Human agreement on *Ideas* in both sets (0.75 vs. 0.69 in Set 7; 0.67 vs. 0.53 in Set 8), indicating that the framework yields more consistent content judgments than individual human raters. While the strongest supervised baseline (FeatEng-RF) remains dominant on several traits, MADRAG's performance without any task-specific training shows that structured reasoning over retrieved exemplars can approximate learned scoring functions. Performance varies systematically by trait type. Discourse-oriented traits such as *Ideas* and *Organization* exhibit the largest gains, with MADRAG often matching or outperforming supervised systems. In contrast, surface-level traits show more modest improvements and greater variance across LLM backbones. For example, on *Conventions* in Set 7, all MADRAG variants trail FeatEng-RF by a wide margin (0.19–0.28 vs. 0.62), and in Set 8, performance on *Word Choice* and *Sentence Fluency* remains inconsistent. The gap likely reflects both the debate structure, which emphasizes holistic argumentation over error counting, and the limited utility of exemplar retrieval for traits where score distinctions hinge on surface-level errors. At the same time, persistent gaps on surface traits suggest that fully replacing supervised models will require hybrid approaches that combine debate-based reasoning with specialized mechanisms for low-level linguistic analysis.

4.2.1 Ablation Study

To isolate the contribution of each component in MADRAG, we conduct an ablation study on the

Table 1: Trait-wise QWK on ASAP Essay Sets 7 and 8. Best and second-best results are in **bold** and underline, respectively. Traits: Idea (Ideas), Org. (Organization), Voc. (Voice), Word (Word Choice), Sent. (Sentence Fluency), Sty. (Style), and Cnv. (Conventions). "—" indicates an unreported trait.

Essay Set	Paradigm	Model	Method	Idea	Org.	Voc.	Word	Sent	Sty.	Cnv.
7	Train-free	GPT-3.5-turbo	MADRAG (Ours)	0.45	0.34	—	—	—	0.24	0.16
7	Train-free	GPT-4o-mini	MADRAG (Ours)	0.43	0.64	—	—	—	0.47	0.26
7	Train-free	GPT-4o	MADRAG (Ours)	0.40	0.38	—	—	—	0.35	0.28
7	Train-free	GPT-5-mini	MADRAG (Ours)	0.69	0.62	—	—	—	0.33	0.22
7	Train-free	GPT-5	MADRAG (Ours)	0.75	0.63	—	—	—	0.47	0.19
7	Train	—	ProTACT	0.50	0.31	—	—	—	—	0.23
7	Train	—	FeatEng-RF	**0.77**	**0.67**	—	—	—	**0.65**	**0.62**
7	Train-free	GPT-3.5-turbo	ZS-LLM	0.05	0.07	—	—	—	0.08	0.10
7	Train-free	LLaMA-2-13B-Chat	ZS-LLM	0.09	0.02	—	—	—	0.15	0.32
7	Train-free	GPT-4	CSR-J	0.55	0.58	—	—	—	0.47	0.22
7	—	—	Human–Human	0.69	0.58	—	—	—	0.54	0.57
8	Train-free	GPT-3.5-turbo	MADRAG (Ours)	0.52	0.43	0.52	0.55	0.41	—	0.49
8	Train-free	GPT-4o-mini	MADRAG (Ours)	0.59	0.47	0.60	**0.65**	0.55	—	0.58
8	Train-free	GPT-4o	MADRAG (Ours)	0.59	0.42	**0.63**	0.61	**0.59**	—	**0.62**
8	Train-free	GPT-5-mini	MADRAG (Ours)	0.60	**0.63**	0.62	0.28	0.34	—	0.36
8	Train-free	GPT-5	MADRAG (Ours)	**0.67**	0.61	0.55	0.32	0.35	—	0.42
8	Train	—	ProTACT	0.57	0.61	—	0.59	0.55	—	0.43
8	Train	—	FeatEng-RF	0.58	**0.63**	0.54	0.55	0.58	—	0.55
8	Train-free	GPT-3.5-turbo	ZS-LLM	0.18	0.25	0.15	0.15	0.20	—	0.31
8	Train-free	LLaMA-2-13B-Chat	ZS-LLM	0.27	0.27	0.27	0.26	0.12	—	0.08
8	Train-free	GPT-4	CSR-J	—	—	—	—	—	—	—
8	—	—	Human–Human	0.53	0.54	0.47	0.48	0.51	—	0.55

merged ASAP Sets 7 and 8, averaging results for overlapping traits. We compare the full model against five variants: **SA** (Single-Agent): a single LLM Judge scores all traits directly from the rubric; **SARAG** (Single-Agent+RAG): the same single Judge is additionally provided with retrieved exemplars; **MA** (Multi-Agent Decomposition): one independent Judge per trait scores that trait from the rubric (decomposition only); **MAD** (Multi-Agent Debate): an Advocate and Skeptic debate each trait and a Judge synthesizes their exchange; and **MARAG** (Multi-Agent+RAG): one Judge per trait receives retrieved exemplars but no debate transcript (retrieval without debate). As shown in Figure 2, performance improves incrementally from **SA** as we add decomposition (**MA**), debate (**MAD**), and retrieval (**MARAG**), with the full MADRAG configuration achieving the best overall performance, particularly on discourse traits such as *Organization*. However, on some surface traits (e.g., *Conventions*), **MARAG** can occasionally outperform MADRAG, suggesting that adversarial debate may introduce noise for fine-grained, error-based scoring. We investigate this trade-off further in RQ3 via targeted failure-mode analysis. A detailed, set-wise breakdown of the ablation results is provided in Appendix D.

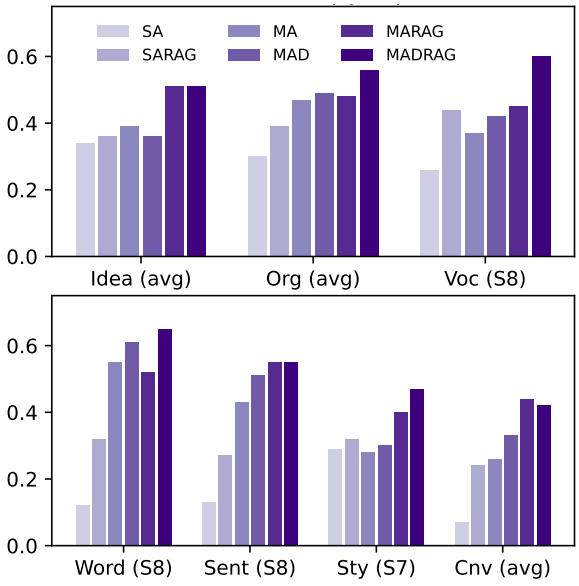


Figure 2: Merged ablation results (QWK). Overlapping traits between ASAP Sets 7 and 8 are averaged.

4.3 Mitigating Middle-Score Bias (RQ2)

LLM-based judges exhibit middle-score bias, clustering predictions toward the center of the scoring scale and avoiding extreme values even when warranted (Zheng et al., 2023; Li et al., 2025). This is particularly problematic for formative assessment, where accurate identification of struggling and exceptional students matters most. We test whether

MADRAG mitigates this bias on essay–trait instances where at least one rater assigned either the minimum or maximum score. For each instance, we report **Agree@1** (agreement within ± 1 point of the extreme score) and **MAE** (mean absolute error from the extreme reference).

Table 2 shows that MADRAG achieves the highest agreement on both essay sets, while also obtaining the lowest error on Essay Set 7 and the second-lowest error on Essay Set 8. Several ablation patterns clarify the sources of these gains. Decomposition alone (MA) yields limited improvement and even degrades performance in Set 8, indicating that naive task division without grounding is insufficient for extreme-score discrimination. Adding retrieval produces substantial gains: MARAG attains 75.3% (Set 7) and 85.5% (Set 8) Agree@1, capturing most of MADRAG's improvement. This confirms that access to score-calibrated exemplars is the primary driver of extreme-score calibration. The inclusion of SARAG further isolates this effect. Compared to SA, SARAG improves agreement in both sets, showing that retrieval alone already mitigates central tendency to some extent. However, SARAG consistently underperforms MARAG, showing that retrieval without trait-wise decomposition is less effective at resolving boundary cases.

Debate without retrieval (MAD) improves over SA and MA but substantially trails retrieval-based variants, reinforcing that debate alone does not correct central tendency. Full MADRAG adds a further 4–5 percentage points in Agree@1 over MARAG in Set 7 and yields the highest agreement overall, indicating that debate provides discriminative refinement on top of retrieval. In Set 8, MARAG achieves slightly lower MAE than MADRAG despite lower agreement, suggesting that debate can occasionally introduce small deviations around the extreme boundary even while improving exact matches.

Across both sets, MADRAG's MAE of approximately 1.0 indicates that residual errors typically land within one score point of the extreme reference, whereas SA and MA exhibit MAE values above 1.5, reflecting systematic regression toward the center. Trait-level analysis shows the largest gains on discourse-oriented traits (e.g., *Ideas*: 82% vs. 39% for SA), while surface traits show more modest improvements (e.g., *Conventions*: 76% vs. 54%), consistent with the overall performance trends in RQ1. Together, these results indicate that mitigating middle-score bias in training-free AES requires explicit calibration mechanisms while debate provides targeted but secondary refinement.

Table 2: Standout subset performance on ASAP Essay Sets 7 and 8. N denotes the number of essay–traits.

Set	Method	N	Agree@1	MAE
7	MADRAG	2,474	**0.795**	**1.001**
7	MARAG	2,474	<u>0.753</u>	<u>1.143</u>
7	MAD	2,474	0.575	1.380
7	MA	2,474	0.424	1.561
7	SARAG	2,474	0.566	1.306
7	SA	2,474	0.477	1.518
8	MADRAG	67	**0.883**	<u>1.011</u>
8	MARAG	67	<u>0.855</u>	**0.834**
8	MAD	67	0.638	1.264
8	MA	67	0.759	1.230
8	SARAG	67	0.608	1.386
8	SA	67	0.520	1.389

4.4 Qualitative Error Analysis on High-Disagreement Cases (RQ3)

While QWK summarizes overall agreement with human raters, it does not explain *why* MADRAG succeeds or fails. To characterize failure modes of judge's *reasoning*, we analyze a targeted subset of **high-disagreement** essay–trait instances where human raters are relatively consistent (within 1 point) but MADRAG deviates by more than 1 point from the human average. Each unit of analysis is one row (e, t) (essay e and trait t), containing the essay text, the rubric trait, two human scores, and the outputs (score + rationale) of five systems: SA, SARAG, MAD, MARAG, and MADRAG. We focus on rows where MADRAG is wrong (N=173) and diagnose *how* its rationale becomes misleading.

Coding scheme and procedure. We annotate each wrong case with three categorical codes. (i) **Reasoning quality** B labels whether MADRAG is rubric-aligned and text-grounded (B2), partially grounded (B1), or misaligned/ungrounded (B0). (ii) **Primary failure mechanism** C assigns one dominant mechanism: trait-boundary confusion (C1), debate framing capture (C2), spurious debate claim accepted (C3), exemplar-induced calibration error (C4), rubric boilerplate collapse (C5), or anonymization distortion (C6). (iii) **Component attribution** D compares the same row across ablations to test whether the *same* failure mechanism persists when a component is removed: debate plausibly contributed (D1), retrieval plausibly contributed (D2), interaction plausible (D3), or not component-specific (D4). Annotators were explicitly instructed to treat anonymization markers (e.g.,

Table 3: Marginal distributions of primary failure mechanisms (C) and component attributions (D).

Primary failure mechanism (C)	**Count**	**%**
C2 Debate framing capture	56	32.7
C6 Anonymization distortion	46	26.9
C3 Spurious debate claim accepted	30	17.5
C5 Rubric boilerplate collapse	14	8.2
C1 Trait-boundary confusion	13	7.6
C4 Exemplar-induced calibration error	12	7.0
Component attribution (D)	**Count**	**%**
D1 Debate plausibly contributed	64	37.4
D3 Interaction plausible (debate+RAG)	60	35.1
D4 Not component-specific	32	18.7
D2 Retrieval plausibly contributed	15	8.8

Table 4: $A \times C$ on wrong MADRAG cases.

Counts						
	C1	C2	C3	C4	C5	C6
A1	12	50	23	6	9	46
A2	1	6	7	6	5	0
Row-normalized (%)						
A1	8.2	34.2	15.8	4.1	6.2	31.5
A2	4.0	24.0	28.0	24.0	20.0	0.0

Table 5: Reasoning quality (B) distribution.

B	**Count**	**%**	**Trait**	**N**	**B0**	**B1**	**B2**
B0	10	5.8	Conventions	49	6	41	2
B1	149	86.6	Organization	45	2	42	1
B2	13	7.6	Ideas/Content	22	0	19	3
			Sent Fluency	33	1	28	4
			Voice	13	1	10	2
			Word Choice	10	0	9	1

`@PERSON`, `@DATE`) as placeholders, not true errors.

Dominant failure mechanisms and their component sources. Table 3 summarizes the marginal distributions of failure mechanisms (C) and component attributions (D) across all incorrect cases. The most frequent mechanism is **debate framing capture** (C2; 32.7%), followed by **anonymization distortion** (C6; 26.9%) and **spurious claim acceptance** (C3; 17.5%). Template-like failures occur less often but remain non-trivial (C4: 7.0%; C5: 8.2%). Attribution analysis indicates that errors are most often linked to **debate dynamics** (D1; 37.4%) or to **debate–retrieval interactions** (D3; 35.1%), whereas retrieval alone is comparatively rare as the primary driver (D2; 8.8%). Appendix E.1 details how specific mechanisms align with individual components, and Appendix E.3 provides deeper qualitative analyses of themes and micro-theories.

Error direction: underscoring dominates. We next examine whether the same mechanisms govern *under-scoring* vs. *over-scoring*. Table 4 shows that most wrong cases correspond to **under-scoring** relative to the human average (A1: 146/173), and that anonymization distortion (C6) is exclusively an under-scoring mechanism in our sample. Over-scoring cases (A2: 27/173) are comparatively more associated with spurious claim acceptance (C3) and template-driven failures (C4/C5), consistent with plausible-sounding but weakly grounded rationales inflating rubric placement in the absence of careful verification.

Errors are usually partially grounded. Table 5 summarizes reasoning quality for wrong cases. Most errors are **partially grounded** (B1; 86.6%): rationales often sound rubric-consistent but fail to cite decisive text evidence or make an explicit evidence→rubric→score link. Fully grounded rationales (B2; 7.6%) are rare by construction in the wrong subset, while misaligned/ungrounded rationales (B0; 5.8%) concentrate in *Conventions* (12.2% B0 within that trait), consistent with fragile surface-form judgments. Appendix E.2 further analyzes how reasoning quality interacts with specific failure mechanisms.

Conclusion

We presented MADRAG, a fully training-free framework for analytic essay trait scoring. Our experiments show that MADRAG significantly outperforms existing LLM judges and achieves parity with state-of-the-art supervised models. By grounding scores in both adversarial reasoning and rubric-aligned exemplars, MADRAG produces calibrated, interpretable trait-level assessments without task-specific training. The success of MADRAG underscores the importance of explicit calibration mechanisms and structured deliberation in LLM-based evaluation. Ultimately, MADRAG illustrates how hybrid LLM frameworks can combine the flexibility of prompt-based scoring with the reliability of supervised systems, paving the way for more accountable and scalable automated assessment.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. 2315294.

Limitations

While MADRAG demonstrates consistent gains over single-agent prompting and training-free LLM judging baselines, several limitations remain. First, our experiments are limited to ASAP Essay Sets 7 and 8, covering only two narrative prompts from middle- and high-school settings; as a result, our conclusions may not fully transfer to other genres (e.g., argumentative or expository writing), grade levels, languages, or rubric structures that appear in real educational deployments. Second, the ASAP essays contain anonymization placeholders (e.g., `@PERSON`, `@DATE`) that can be misread as genuine grammatical or mechanical errors, particularly for surface-level traits such as Conventions, introducing bias that is unrelated to true writing quality and potentially distorting trait-specific scores.

In addition, MADRAG is more computationally demanding than single-agent prompting or supervised AES systems at inference time, since each trait evaluation requires multiple LLM calls (Advocate, Skeptic, and Judge) as well as embedding-based retrieval to construct exemplars; this cost may be prohibitive at large scale unless carefully optimized or selectively applied. Moreover, we report single-run results for each configuration rather than aggregates over repeated trials, as re-running the full MADRAG pipeline across multiple random seeds or configurations would require substantial additional computational resources due to repeated LLM invocations; consequently, the reported numbers should be interpreted as outcomes from one specific instantiation of the MADRAG pipeline rather than as mean or variance estimates over repeated runs.

Finally, although MADRAG is training-free in that it does not require parameter updates or supervised fine-tuning, it is not data-free: retrieval-augmented generation presupposes access to a bank of manually scored essays. That said, the amount of labeled data needed for RAG is typically much smaller than what is required to train or fine-tune a scoring model—at minimum, one well-chosen exemplar per trait–score level can provide basic coverage for retrieval. In settings with limited coverage, especially for certain traits or score levels, retrieval may fail to surface score-discriminative exemplars, weakening calibration with human preference.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A large-scale corpus for assessing source-based writing quality: Asap 2.0. *Assessing Writing*, 65:100954.

Paul Deane. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24. Automated Assessment of Writing.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Avisa Fallah, Ali Keramati, Mohammad Ali Nazari, and Fatemeh Sadat Mirfazeli. 2024. Automating theory of mind assessment with a llama-3-powered chatbot: Enhancing faux pas detection in autism. In *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 365–372.

Joaquín Jordán, Xavier Yin, Melissa Fabros, Gireeja Ranade, and Narges Norouzi. 2025. Magic: Multi-agent argumentation and grammar integrated critiquer. *Preprint*, arXiv:2506.13037.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas

Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Ute Knoch. 2009. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2):275–304.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders? *Preprint*, arXiv:2409.13120.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025. Evaluating scoring bias in llm-as-a-judge. *Preprint*, arXiv:2506.22316.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zijie Lin and Bryan Hooi. 2025. Enhancing multi-agent debate system performance via confidence expression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6453–6471, Suzhou, China. Association for Computational Linguistics.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.

Rebecca M. M. Hicke, Yuri Bizzoni, Pascale Feldkamp, and Ross Deans Kristensen-McLachlan. 2025. Says who? effective zero-shot annotation of focalization. In Taylor Arnold, Margherita Fantoli, and Ruben Ros, editors, *Computational Humanities Research 2025*, pages 738–754. Anthology of Computers and the Humanities.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Dadi Ramesh and Suresh Kumar Sanampudi. 2021. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55:2495 – 2527.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Takumi Shibata and Yuichi Miyamura. 2025. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29976–29989, Suzhou, China. Association for Computational Linguistics.

Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. 2025. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *Preprint*, arXiv:2505.13965.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14):e34262.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Preprint*, arXiv:2206.10498.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):157–180.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix: Implementation and Reproducibility Details

This appendix provides implementation details for reproducing the MADRAG pipeline described in Section 3, including decoding settings, log-probability extraction, asynchronous orchestration, and retrieval configuration. For transparency, we summarize the exact execution logic used to generate the scores reported in Section 4.

A.1 Runtime Environment and Execution

All experiments were executed in a single experimental run using a Python pipeline that processes each essay independently and evaluates it across all rubric traits in parallel (one debate instance per trait), due to the computational cost of repeated LLM evaluations. Each trait evaluation consists of (i) Advocate generation, (ii) Skeptic rebuttal generation conditioned on the Advocate, and (iii) Judge scoring conditioned on the debate transcript and retrieved exemplars.

The implementation uses asynchronous execution via `asyncio` to concurrently evaluate all rubric traits for a given essay:

- For each essay, we spawn one task per trait using `asyncio.gather`.
- Each task runs a three-step debate (Advocate $\rightarrow$ Skeptic $\rightarrow$ Judge).

Output logging. For each (essay, trait) pair, the system logs: (i) the Judge rationale text, (ii) the parsed final integer score, (iii) raw model outputs, (iv) token-level log-probabilities (where supported), (v) last-token alternative candidates (top logprobs), and (vi) the full debate transcript (Advocate + Skeptic).

A.2 Decoding and Inference Settings

We use deterministic decoding for the **Judge** to reduce run-to-run variance and ensure that score outputs are stable under identical inputs. Concretely:

- **Judge:** temperature set to 0.0 (greedy / deterministic decoding).

For the **Advocate** and **Skeptic**, we use highest-temperature sampling to allow diversity in argumentation:

- **Advocate and Skeptic:** temperature set to 0.7.

Temperature selection for debate agents. We set the decoding temperature for the **Advocate** and **Skeptic** to 0.7 based on a lightweight manual tuning procedure. Concretely, we varied temperature on a small subset of essay–trait instances and qualitatively inspected how it affected the agents' argumentative behavior (e.g., specificity, coverage, and willingness to challenge or defend claims). Because this tuning focused on reasoning style rather than end-task accuracy, we did not treat performance as the selection criterion. Moreover, sweeping temperatures over the full evaluation pipeline would be computationally expensive due to the multi-step, multi-trait debate structure. We therefore fix $T = 0.7$ for all Advocate/Skeptic generations in all experiments.

A.3 Models and API Calls

We evaluate MADRAG with multiple underlying LLMs by swapping the model used in the chat completion call. The codebase includes wrappers for OpenAI chat completion models (e.g., GPT-4o-mini, GPT-4o, GPT-5-mini, GPT-5) called via `AsyncOpenAI`.

A.4 Confidence Proxy from Token Log-Probabilities

To enable confidence-aware judging, we extract a lightweight confidence proxy from the first generated token of each debate agent response (Advocate and Skeptic). When available, the OpenAI API returns per-token log-probabilities for generated tokens. Let ℓ denote the log-probability of the first

emitted token. We compute the proxy confidence as:

$$c = \exp(\ell) \in (0, 1].$$

Although the surface token itself may appear semantically neutral (e.g., discourse markers such as "As"), its probability is computed conditioned on the full preceding context, including the prompt, essay, rubric trait, and debate state. In an autoregressive model, this next-token distribution reflects the model's internal belief over possible continuations. Prior work has shown that token-probability signals can serve as useful confidence indicators in LLM-based annotation and evaluation settings (M. M. Hicke et al., 2025).

We adopt the first-token proxy for two practical reasons:

1. **Stability across responses.** Sequence-level aggregation methods (e.g., average log-probability or perplexity-style estimates) can become sensitive to response length and verbosity, especially in debate-style generation where arguments may vary substantially in length.

2. **Discriminative range.** In preliminary experiments, sequence-level estimates often compressed into a narrow range of values, making them less informative for downstream judging decisions.

Nevertheless, our implementation logs the full token-level log-probability sequence whenever it is returned by the API. This allows alternative confidence estimators to be computed without additional model calls. In particular, we record:

- the full token-level log-probability sequence for each response when available;

- the top-k alternative tokens for the final position (via `top_logprobs`) when provided by the API.

For models without explicit log-probability support (e.g., GPT-5), we instead use the model's self-reported confidence score as a proxy signal.

A.5 Score Parsing and Output Constraints

The Judge is instructed to output an integer score within the valid trait range. The pipeline parses the final score using a regular expression that extracts:

`Final Score: {integer}`.

Advocate Agent

You are an Advocate Agent in a multi-agent debate system for essay scoring. Your role is to support the essay by highlighting its strengths with respect to strictly within the single trait "$TRAIT_NAME". You must analyze the essay and provide detailed, text-based evidence of what is done well according to the rubric's expectations for "$TRAIT_NAME" trait.
Do not assign a score. Do not summarize or critique weaknesses. Focus entirely on supporting the essay's strengths as they relate to the specific sub-trait. Use quotes or paraphrased excerpts from the essay when needed. Be specific and detailed in your analysis.
Anonymization $ANON_CONTEXT

Figure 3: Adocate system prompt.

All text preceding the final score marker is stored as the Judge rationale. If the score cannot be parsed, the system records the rationale but flags the score as missing.

A.6 Prompt Templates

All agent prompts are stored as external template files and rendered at runtime using a shared context dictionary. The context includes the trait name, the full rubric trait serialized as JSON, the essay text, the essay prompt/question, and the valid score range. In addition, the Judge templates include (i) retrieved few-shot exemplars spanning the full score range for the current trait and (ii) the debate transcript (Advocate opening + Skeptic rebuttal). We arrived at the final prompt settings through an extensive, iterative prompt-engineering process involving multiple rounds of pilot runs and refinements to enforce role constraints, improve output format reliability, and reduce failure modes (e.g., agents assigning scores or mixing traits). For ease of inspection and reproducibility, we include the exact prompt templates used in our experiments below. Figures 3–5 show the system instructions used for the Advocate, Skeptic, and Judge roles, respectively.

Skeptic Agent

You are a Skeptic Agent in a multi-agent debate system for essay scoring. Your role is to critically analyze the essay and identify weaknesses according to strictly within the single trait "$TRAIT_NAME". Focus on providing detailed, evidence-based critiques of how the essay falls short for "$TRAIT_NAME" trait.
Do not assign a score. Do not mention positive aspects. Concentrate only on identifying issues, weaknesses, and areas where the essay does not meet the rubric's expectations. Use specific excerpts or descriptions to support your critique.
Anonymization $ANON_CONTEXT

Figure 4: Skeptic system prompt.

Judge Agent

You are "The Synthesizer-Judge," an impartial arbiter for the single trait "$TRAIT_NAME" multi-agent debate system for essay scoring.
Your Job - Read the debate transcript between Advocate and Skeptic agents who previously debated regarding the essay strengths and weaknesses. - Weigh the arguments against the rubric for "$TRAIT_NAME". - Produce a final integer score from $MIN_POINTS to $MAX_POINTS.
Anonymization $ANON_CONTEXT

Figure 5: Judge system prompt.

A.7 Retrieval-Augmented Exemplar Construction

We use retrieval augmentation to provide the Judge with calibration examples spanning the full score range for each trait. For each essay and trait, we construct a **trait-specific exemplar prompt** by retrieving one example essay for each valid score in $[s_{\min}, s_{\max}]$.

Operationally, for each trait name τ and score set $\{s_{\min}, \ldots, s_{\max}\}$, we call a function and concatenate the returned exemplars into a few-shot block that is injected into the Judge context.

RAG isolation across roles. To prevent debate agents from anchoring on retrieved examples, we provide retrieved exemplars to:

- **Judge only**, as part of its context prompt.

The Advocate and Skeptic receive only the rubric trait and essay content.

A.8 Example Multi-Agent Debates

We present representative examples of the multi-agent debate process used in MADRAG, including the Advocate, Skeptic, and Judge agents. Each example corresponds to a single essay–Idea Trait (Figures 6, 7).

B Dataset Details and Preprocessing

ASAP overview. The ASAP dataset is a widely used benchmark for automated essay scoring, originally released as part of a Kaggle competition sponsored by the William and Flora Hewlett Foundation. The dataset consists of anonymized English essays written by students in grades 7–10 in response to eight distinct prompts, each defining a separate *essay set*. Essay sets vary substantially in genre (persuasive, narrative, and source-dependent response), length, grade level, and scoring rubric, making ASAP a challenging and diverse evaluation benchmark for AES systems.

All essays were scored by trained human raters following prompt-specific guidelines. Each essay receives a resolved (overall) score, and for a subset of prompts, additional analytic trait scores are available. Due to these properties, ASAP has been extensively adopted in prior work evaluating both holistic and trait-level essay scoring models.

While all essay sets include holistic scores, the ASAP release provides trait-level annotations only for Essay Sets 7 and 8. Accordingly, we document the full dataset for completeness and reproducibility, but restrict our experiments to Sets 7 and 8, the only subsets that provide multiple independent human ratings at the trait level.

B.1 ASAP Essay Set Statistics

Table 6 summarizes key properties of all eight ASAP essay sets, including essay type, grade level, training set size, and the availability of trait-level annotations. Consistent with prior analyses, essay lengths range from short source-dependent responses (approximately 150 words) to long narrative essays exceeding 600 words on average, with

score ranges varying substantially across prompts.

Rationale for focusing on Essay Sets 7 and 8. Although ASAP contains eight essay sets, only Essay Sets 7 and 8 provide independent trait-level scores from at least two human raters per essay. This property is essential for our study, which explicitly examines trait-level reliability, Human–Human agreement, and model calibration under rater disagreement. Consequently, all quantitative evaluations in the main paper are conducted exclusively on Sets 7 and 8.

B.2 Essay Set 7: Prompt

Prompt. *Write about patience. Being patient means that you are understanding and tolerant. A patient person experiences difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.*

B.3 Essay Set 8: Prompt

Prompt. *We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.*

B.4 Text Transcription and Fidelity

ASAP essays were transcribed from handwritten student responses following strict transcription guidelines. Misspellings and grammatical errors were preserved exactly as written, and no normalization or correction was applied that could alter surface-level evidence relevant to traits such as *Conventions*. When a handwritten word could not be reliably inferred, it was omitted according to the original transcription protocol. In our experiments, we didn't apply any preprocessing: No spelling correction, grammar normalization, or sentence restructuring is performed.

B.5 Trait Labels and Preprocessing

Rater scores. For Essay Sets 7 and 8, each essay–trait instance includes scores from at least two independent human raters. These scores are retained explicitly to compute Human–Human agreement and to define evaluation targets.

Reference label construction. For model-vs-human evaluation, we construct a single reference score per essay–trait pair by averaging the available rater scores and rounding to the nearest valid integer within the trait's scoring range. This procedure avoids privileging any individual rater while remaining consistent with the discrete rubric scales.

Retrieval pool and data leakage prevention. For retrieval-augmented judging, exemplar essays are drawn exclusively from the training split of the same essay set. The evaluated essay is never eligible to be retrieved as an exemplar. Additional details of exemplar construction and role-specific access are provided in Appendix A.7.

C Baseline Models

ProTACT. We include ProTACT (Do et al., 2023) as a strong *training-based* neural baseline for cross-prompt analytic trait scoring. ProTACT learns prompt-aware essay representations via essay–prompt attention and augments them with engineered essay-quality features (including a topic-coherence feature), while a trait-similarity objective encourages consistent predictions across correlated traits. Because the original paper does not report a complete set of trait-wise results for ASAP Sets 7–8, we run the authors' released implementation on the public data and reproduce the evaluation pipeline to obtain the missing sub-trait QWK scores reported in our tables.

Feature-Engineered Trait Scorer (FeatEng-RF). We include the supervised trait-scoring baseline of Mathias and Bhattacharyya (Mathias and Bhattacharyya, 2020), which predicts analytic trait scores on ASAP using a Random Forest model trained on a large set of hand-crafted linguistic features (e.g., length, punctuation, syntax, style, and cohesion indicators such as discourse connectives and entity-grid features) under a five-fold cross-validation protocol. Because this approach is trained directly on ASAP trait labels, we treat it as a strong *supervised* reference point. (Mathias and Bhattacharyya, 2020)

Prompt-Engineered Zero-Shot LLM Judge (ZS-LLM). We report training-free, single-agent LLM baselines from Mansour et al. (Mansour et al., 2024), who evaluate ChatGPT (gpt-3.5-turbo) and LLaMA-2-13B-Chat on ASAP via rubric-aware, prompt-engineered scoring without any supervised fine-tuning. Their prompts provide the essay

Table 6: Summary of the ASAP dataset across all eight essay sets. Trait-level annotations are available only for Essay Sets 7 and 8 in the ASAP release.

Set	Essay Type	Grade	Train Size	Traits
1	Persuasive / Narrative / Expository	8	1,783	—
2	Persuasive / Narrative / Expository	10	1,800	—
3	Source-dependent responses	10	1,726	—
4	Source-dependent responses	10	1,772	—
5	Source-dependent responses	8	1,805	—
6	Source-dependent responses	10	1,800	—
7	Persuasive / Narrative / Expository	7	1,569	4 traits
8	Persuasive / Narrative / Expository	10	723	6 traits

prompt, score range, and rubric guidelines, and are progressively strengthened with structured instructions, role formatting, and one-shot exemplars; decoding is deterministic (temperature $= 0$). The model outputs holistic and trait-level scores (optionally with feedback), and the authors find performance is highly prompt- and task-dependent, yet remains well below supervised and cross-prompt SOTA in QWK—especially for trait scoring on Sets 7 and 8.

Criteria & Sample-Referenced LLM Scoring (CSR-J). We include the training-free prompt-based LLM scorer of Tang et al. (Tang et al., 2024), which uses a single GPT-4 judge to assign analytic trait scores from the rubric and to produce brief rationales grounded in *sample-referenced* exemplars (i.e., human-scored example essays provided in the prompt). The method evaluates ASAP Essay Set 7 on Ideas, Organization, Style, and Conventions, and reports trait-wise QWK under deterministic decoding (temperature $= 0$), serving as a strong single-agent, prompt-engineered reference.

Human–Human agreement. Finally, we report Human–Human agreement as QWK between the two human raters for each trait, serving as an approximate reference ceiling given inherent rater variability.

C.1 Why One Round and Why Advocate→Skeptic

Finally, we analyze design choices in the debate transcript provided to the judge. Our objective is to feed the judge a *minimal but sufficient* argumentative context: an evidence-heavy pro argument followed by a targeted rebuttal. We compare (i) advocate-only transcripts, (ii) skeptic-only transcripts, and (iii) concatenating both threads, as well as the effect of adding an additional debate round.

One round is substantially more stable than two rounds. Across both sets, extending debate beyond a single exchange causes a sharp drop in QWK (e.g., `both_round2` collapses relative to `both_round1`). This is consistent with the hypothesis that longer debates amplify verbosity, drift, and non-local contradictions, which can degrade the judge's calibration even when RAG exemplars are provided.

Advocate-first is a better conditioning signal than skeptic-first. Skeptic-only transcripts yield near-zero agreement in both sets, indicating that leading with exclusively negative framing can push the judge toward systematic under-scoring or rubric-misaligned reasoning. In contrast, advocate-first provides a structured inventory of rubric-aligned evidence, after which the skeptic rebuttal can selectively challenge specific claims. This ordering preserves *coverage* (positives are surfaced) while still introducing *adversarial pressure* (weaknesses are surfaced) in a controlled way.

Why not concatenating both full threads. Although `both_round1` can be competitive on some traits, it is less reliable across traits and prompts than the advocate-first exchange used in MADRAG. Empirically, concatenation appears to dilute the discourse structure (two parallel narratives with competing local context), increasing the judge's burden to resolve inconsistencies. The advocate→skeptic exchange yields a single, linear argumentative path that is easier for the judge to synthesize.

Quantitative evidence. Tables 7 and 8 summarize QWK under different transcript configurations. The overall pattern is clear: skeptic-first is consistently poor; two rounds is unstable; and a short advocate-led exchange offers the most reliable trade-off between signal and noise.

Table 7: Essay Set 7: Transcript configuration analysis (QWK). Adv.=Advocate-only, Skp.=Skeptic-only, R=Round. * denotes our proposed method (MADRAG).

Method	Idea	Org.	Sty.	Cnv.
Adv. (R1)*	0.43	**0.64**	**0.45**	**0.26**
Adv. (R2)	0.46	0.24	0.10	0.02
Both (R1)	**0.48**	0.24	0.14	0.08
Both (R2)	0.14	0.02	0.03	0.01
Skp. (R1)	0.07	0.05	0.01	0.00
Skp. (R2)	0.06	0.01	0.01	0.01

D Detailed Ablation Study

This appendix provides a detailed breakdown of the ablation study discussed in Section 4.2.1. The study evaluates the incremental impact of each major component in the MADRAG pipeline by comparing the following configurations:

- **SA (Single-Agent):** A single LLM acts as a judge, scoring the essay directly using the rubric without debate or retrieved exemplars.
- **SARAG (Single-Agent+RAG):** A single LLM judge scores the essay using the rubric *and* retrieved, rubric-aligned exemplar essays spanning the score range, but *without* trait decomposition or debate.
- **MA (Multi-Agent):** Multiple, independent LLM agents score the essay for a trait. Their scores are averaged, simulating a multi-rater setup without interaction or retrieval.
- **MAD (Multi-Agent Debate):** Introduces the Advocate and Skeptic agents who generate a debate transcript. The Judge scores the essay based on this transcript, but *without* access to retrieved exemplars for calibration.
- **MARAG (Multi-Agent with RAG):** Multi-agent scoring is combined with RAG. The Judge receives exemplars spanning the score range but does *not* see a debate transcript.
- **MADRAG:** The full proposed framework, combining the Advocate–Skeptic debate transcript with retrieval-augmented exemplars for the Judge.

Table 9 presents the complete trait-wise results for Essay Sets 7 and 8 separately. The merged view, which averages overlapping traits, is shown in the main paper as Figure 2.

Key observations. Several consistent patterns emerge from the detailed ablation results.

- **Retrieval provides the largest single gain in calibration.** Comparing SA to SARAG reveals that exemplar-based retrieval alone yields substantial improvements across both essay sets, particularly on surface-oriented traits such as *Conventions*. This confirms that access to score-calibrated exemplars is a primary driver of improved agreement, even in the absence of decomposition or debate.
- **Debate and decomposition offer complementary but trait-dependent benefits.** Moving from SA to MA yields modest gains, indicating that trait-wise decomposition and multiple perspectives help stabilize judgments but are insufficient on their own. Adding debate (MAD) further improves performance on several discourse-oriented traits, most notably *Organization* and *Sentence Fluency*, suggesting that adversarial reasoning is particularly beneficial when evaluating higher-level structure and coherence.
- **Multi-agent retrieval (MARAG) outperforms single-agent retrieval (SARAG).** Across nearly all traits, MARAG consistently improves over SARAG, indicating that trait-wise decomposition remains valuable even when retrieval is present. This gap highlights that retrieval alone does not fully resolve rubric alignment issues without trait-specific conditioning.
- **Debate can introduce noise on surface-level traits.** For some surface traits, MARAG slightly outperforms MADRAG. For example, on *Conventions* in Set 7, MARAG achieves higher agreement than MADRAG (0.35 vs. 0.26), and a similar pattern appears for *Word Choice* in Set 8. These regressions suggest that adversarial debate can amplify spurious or surface-form cues, motivating a closer analysis of debate-induced failure modes.
- **Overall, MADRAG delivers the strongest and most consistent performance.** Despite occasional regressions on individual surface traits, the full MADRAG framework achieves the best or near-best performance on the majority of traits across both essay sets, particularly for discourse-oriented dimensions.

Table 8: Essay Set 8: Transcript configuration analysis (QWK). * denotes our proposed method (MADRAG).

Method	Idea	Org.	Voc.	Word	Sent.	Cnv.
Adv. (R1)*	**0.59**	**0.47**	**0.60**	**0.65**	**0.55**	**0.58**
Adv. (R2)	0.37	0.20	0.39	0.37	0.54	0.62
Both (R1)	0.57	0.45	0.52	0.63	0.53	0.54
Both (R2)	0.17	0.17	0.19	0.30	0.30	0.27
Skp. (R1)	0.11	0.03	0.13	0.14	0.13	0.10
Skp. (R2)	0.15	0.10	0.15	0.15	0.23	0.18

Table 9: Ablation study (QWK) on ASAP Essay Sets 7 and 8. SA: Single-Agent, MA: Multi-Agent, MAD: Multi-Agent Debate, MARAG: Multi-Agent with RAG, MADRAG: Full framework.

Set	Method	Idea	Org.	Voc.	Word	Sent	Sty.	Cnv.
7	SA	0.27	0.33	—	—	—	0.29	0.06
7	SARAG	0.27	0.36	—	—	—	0.32	0.20
7	MA	0.31	0.39	—	—	—	0.28	0.17
7	MAD	0.25	0.40	—	—	—	0.30	0.20
7	MARAG	**0.56**	0.47	—	—	—	0.40	**0.35**
7	MADRAG	0.43	**0.64**	—	—	—	**0.47**	0.26
8	SA	0.41	0.26	0.26	0.12	0.13	—	0.08
8	SARAG	0.45	0.42	0.44	0.32	0.27	—	0.28
8	MA	0.47	0.54	0.37	0.55	0.43	—	0.34
8	MAD	0.47	0.57	0.42	0.61	0.51	—	0.46
8	MARAG	0.45	**0.49**	0.45	0.52	0.55	—	0.53
8	MADRAG	**0.59**	0.47	**0.60**	**0.65**	**0.55**	—	**0.58**

This pattern confirms the intended synergy between adversarial reasoning and exemplar-based calibration.

In Section 4.4), we analyze the qualitative failure mechanisms that arise from debate–retrieval interactions.

E Detailed Qualitative Analysis

E.1 Primary Failure Mechanisms and Component Attribution Links

To probe *which* mechanisms are linked to which components, Table 10 reports the $C \times D$ matrix. Debate framing capture (C2) is predominantly debate-linked (55.4% D1 row-normalized), consistent with the judge inheriting the debate stance without verifying against the essay. In contrast, anonymization distortion (C6) is disproportionately interaction-coded (63.0% D3), suggesting that surface-form cues often become harmful when debate and retrieval jointly increase attention to token-level artifacts. Finally, the template-like mechanisms (C4/C5) are most often *not* component-specific (66.7% and 64.3% D4), indicating that generic rubric prose and mid-band defaults are largely baseline judge limitations rather than uniquely induced by debate or retrieval.

Table 10: $C \times D$ contingency table for wrong MADRAG cases. Top: counts. Bottom: row-normalized percentages (each row sums to 100).

Counts				
	D1	D2	D3	D4
C1	5	2	3	3
C2	31	6	13	6
C3	13	2	13	2
C4	2	0	2	8
C5	1	4	0	9
C6	12	1	29	4
Row-normalized (%)				
C1	38.5	15.4	23.1	23.1
C2	55.4	10.7	23.2	10.7
C3	43.3	6.7	43.3	6.7
C4	16.7	0.0	16.7	66.7
C5	7.1	28.6	0.0	64.3
C6	26.1	2.2	63.0	8.7

E.2 Primary Failure Mechanisms and Reasoning Quality Links

Table 11 links reasoning quality to failure mechanisms. Anonymization distortion (C6) accounts for the majority of B0 cases (8/10), indicating that truly ungrounded rationales often arise when placeholders are mistaken as genuine mechanical errors. In contrast, exemplar-induced calibration errors (C4) are the one mechanism that frequently yields *high-quality* rationales (B2: 46.2% within C4) despite being wrong, suggesting that these errors are less about incoherent reasoning and more about systematic miscalibration toward a rubric band.

E.3 Mechanism deep dives: themes and micro-theories

We synthesize the most frequent failure mechanisms into three recurring themes, each expressed as a micro-theory about how debate and retrieval shape the judge's attention and calibration.

Theme 1: Token myopia (C6) — Anonymization treated as real error. A dominant pattern is

Table 11: $B \times C$ on wrong MADRAG cases. Top: counts. Bottom: row-normalized percentages.

Counts			
	B0	B1	B2
C1	0	10	3
C2	0	53	3
C3	2	27	1
C4	0	7	6
C5	0	14	0
C6	8	38	0
Row-normalized (%)			
C1	0.0	76.9	23.1
C2	0.0	94.6	5.4
C3	6.7	90.0	3.3
C4	0.0	53.8	46.2
C5	0.0	100.0	0.0
C6	17.4	82.6	0.0

that MADRAG cites anonymization markers (e.g., `@CAPS/@PERSON/@DATE`) as "capitalization" or "formatting" failures, especially in *Conventions* (and occasionally *Fluency* or *Voice*). **Micro-theory:** when the surface form contains many anonymization tokens, MADRAG over-weights them as evidence of convention breakdown and readability loss, leading to systematic under-scoring even when the underlying prose is readable.

Theme 2: Debate capture (C2) — Judge inherits its stance without verification. In many failures, the judge echoes Advocate/Skeptic framing (often the Skeptic) without checking whether the claimed defect is supported by the essay (e.g., "no thesis," "no paragraph breaks," "disorganized" despite clear temporal markers and closure). **Micro-theory:** debate increases the salience of critique, but the judge sometimes substitutes "debate resolution" for "text verification," producing overconfident misdiagnoses about structure and coherence.

Theme 3: Rubric-template collapse (C4/C5) — Generic band language replaces close reading. A smaller but important class of errors reflects template-driven justifications (e.g., "clear but limited development", "errors impede readability") that are weakly tied to the essay and insensitive to strong counter-evidence. **Micro-theory:** under uncertainty, MADRAG falls back on plausible-sounding rubric prose, reducing sensitivity to extremes and enabling large deviations when the essay is clearly strong or clearly weak on the target trait.

E.3.1 Implications for MADRAG design

The qualitative results suggest that MADRAG's components change *what the judge attends to*, not only the final score. Debate often improves structured critique, but it also creates frequent failure via framing capture (C2) and acceptance of unsupported debate claims (C3), indicating the need for explicit *text verification* constraints in the judge prompt. Retrieval is less often the sole driver of failure (D2), but interaction effects are common (D3), especially when surface-form noise is present. Finally, anonymization tokens represent a systematic confound for convention-heavy traits: without explicit normalization or masking, placeholders are repeatedly treated as genuine mechanical errors, producing predictable under-scoring.

Table 12: Representative Token Myopia case (C6): MADRAG treats anonymization tokens as mechanical errors, driving an overly low conventions judgment.

Trait	Conventions
Memo (E)	For conventions, the essay shows generally strong control of standard conventions: many sentences have correct end-of-sentence punctuation and accurate spelling. MADRAG emphasizes that capitalization/placeholders and punctuation errors significantly impede readability, but this is misleading because many flagged issues are anonymization tokens (e.g., `@CAPS`, `@PERSON`) rather than author errors, and multiple clearly punctuated sentences remain readable.

Table 13: Representative Debate Capture case (C2): MADRAG adopts a negative framing about organization despite clear sequencing cues.

Trait	Organization
Memo (E)	For organization, the essay shows clear chronological sequencing and a recognizable beginning, middle, and reflective ending with explicit temporal markers. MADRAG emphasizes abrupt transitions and disjointedness, but this is misleading because the text uses markers such as "The next day" and "To my surprise" and provides a coherent arc with closure, indicating functioning structure.

Table 14: Representative Rubric-Template case (C4/C5): MADRAG uses mid-band generic rationale that underresponds to essay evidence.

Trait	Ideas and Content
Memo (E)	For ideas_and_content, the essay demonstrates a clear, focused narrative with a developed main idea supported by concrete scenes and an explicit resolution. MADRAG emphasizes that the piece is fragmented and insufficiently developed, but this is misleading because the text provides escalating, sensory detail and a decisive emotional payoff that aligns with high-band rubric traits.

Turn 1: Advocate Message

```
As the Advocate for the "Ideas" criterion, my role is to highlight the strengths of the essay in
↪  terms of its focus and development of ideas related to the topic of patience.

The essay demonstrates a clear focus on the topic of patience by centering the narrative around
↪  the relationship between the narrator and their horse, @CAPS1. The story is consistently
↪  anchored in the theme of patience, as evidenced by the repeated references to the horse's
↪  impatience and the narrator's attempts to manage it. This focus is maintained throughout the
↪  essay, ensuring that the reader understands the central theme.

The development of ideas is supported by specific and relevant details that bring the story to
↪  life. For instance, the essay describes @CAPS1 as "a chestnut quarter horse with a white star
↪  on his forehead," which not only provides a vivid image of the horse but also personalizes
↪  the narrative, making the story more engaging. The description of @CAPS1's behavior, such as
↪  being "hipper and ready to go" and not standing still when being tacked up, adds depth to the
↪  portrayal of the horse's impatience. These details are specific and relevant, as they
↪  directly illustrate the challenges the narrator faces in dealing with an impatient horse.

Furthermore, the essay includes specific examples of situations where patience is tested, such
↪  as when @CAPS1 is "so far up ahead" during a ride that the narrator has to stop him to wait
↪  for the group. This scenario effectively highlights the theme of patience, as it shows the
↪  narrator's efforts to manage the horse's behavior and maintain harmony within the group. The
↪  narrative also conveys the narrator's ongoing struggle and acceptance of the horse's nature,
↪  as seen in the line, "Meany times I have tried training him to be patient but I have always
↪  failed." This acknowledgment of the narrator's attempts and eventual acceptance adds a layer
↪  of complexity to the story, demonstrating a thorough development of the theme.

In conclusion, the essay excels in maintaining a clear focus on the theme of patience and
↪  develops this idea with specific, relevant details that enrich the narrative. The consistent
↪  emphasis on the horse's impatience and the narrator's response to it effectively illustrates
↪  the challenges and nuances of being patient.
```

Turn 2: Skeptic Rebuttal

```
The Advocate's argument suggests that the essay maintains a clear focus on the theme of patience
↪  and develops this idea with specific, relevant details. However, upon closer examination, the
↪  essay's focus on the theme of patience is not as clear or thoroughly developed as claimed.

Firstly, the essay primarily describes the horse's impatience rather than exploring the concept
↪  of patience itself. The narrative is centered around the horse's behavior, with limited
↪  exploration of the narrator's experience of patience. The essay mentions the horse being
↪  "hipper and ready to go" and not standing still, but these details focus more on the horse's
↪  characteristics rather than the narrator's patience. The narrative lacks depth in
↪  illustrating how the narrator embodies patience or how they cope with the challenges
↪  presented by the horse's behavior.

Moreover, the essay does not provide a comprehensive exploration of the narrator's attempts to
↪  manage the horse's impatience. While it mentions that the narrator has "tried training him
↪  to be patient," it does not delve into specific strategies or experiences that demonstrate
↪  the narrator's patience. The lack of detailed examples of the narrator's actions or
↪  reflections on their patience limits the development of the theme.

Additionally, the essay concludes with the narrator's acceptance of the horse's nature, stating,
↪  "I will always love him no matter how un patience he is." This statement shifts the focus
↪  from patience to acceptance, which, while related, is not the same concept. The narrative
↪  does not sufficiently explore the process or challenges of being patient, which is the
↪  central task of the essay prompt.

In summary, the essay's ideas are not thoroughly developed in relation to the theme of patience.
↪  The focus is more on the horse's impatience rather than a nuanced exploration of the
↪  narrator's patience, resulting in a narrative that lacks depth and specificity in addressing
↪  the essay prompt.
```

Figure 6: Example multi-agent debate for the *Ideas* trait (Part I): Advocate initiation and Skeptic rebuttal.

Turn 3: Judge Decision

```
"The debate between the Advocate and the Skeptic highlights differing perspectives on the
↪  essay's focus and development of ideas related to the theme of patience.

The Advocate argues that the essay maintains a clear focus on patience by centering the
↪  narrative around the relationship between the narrator and their horse, @CAPS1. The Advocate
↪  points out that the essay includes specific and relevant details, such as the horse's
↪  description and behavior, which illustrate the challenges the narrator faces. The Advocate
↪  also notes that the essay provides examples of situations where patience is tested, adding
↪  depth to the narrative.

On the other hand, the Skeptic contends that the essay primarily describes the horse's
↪  impatience rather than exploring the concept of patience itself. The Skeptic argues that the
↪  narrative lacks depth in illustrating how the narrator embodies patience or copes with the
↪  challenges presented by the horse's behavior. The Skeptic also points out that the essay
↪  does not provide detailed examples of the narrator's actions or reflections on their
↪  patience, limiting the development of the theme. Additionally, the Skeptic notes that the
↪  conclusion shifts the focus from patience to acceptance, which is not the same concept.

Weighing these arguments against the rubric, the essay does tell a story that is somewhat
↪  focused on the topic of patience. However, the development of ideas is more centered on the
↪  horse's impatience rather than a thorough exploration of the narrator's patience. The essay
↪  includes a mix of specific and general details, but it lacks a comprehensive exploration of
↪  the narrator's experience of patience. The narrative does not delve deeply into the
↪  strategies or experiences that demonstrate the narrator's patience, which limits the
↪  development of the theme.

Based on this analysis, the essay aligns more closely with a score of 2, as it tells a story with
↪  ideas that are somewhat focused on the topic and are developed with a mix of specific and
↪  general details.

Final score: 2"
```

Figure 7: Example multi-agent debate for the *Ideas* trait (Part II): Judge synthesis and final score.

Never Care For What They Say? Platform vs Genre Rules in Online Horror Narratives (2007–2024)

Alexandre Lionnet-Rollin
École Pratique des Hautes Études, Université PSL
Paris, France

Florian Cafiero
LRE, EPITA
CJM, École nationale des chartes, Université PSL
Paris, France

Abstract

Research on online cultural production shows that platforms are acting as mediators that can heavily shape textual form. Yet, empirical work is often platform-bounded, making it difficult to assess whether stylistic regularities that we observe are indeed genre signals or if some of them are platform artefacts. We address this question through a cross-platform design focused on *creepypasta*, a digital-born horror genre circulating across heterogeneous infrastructures. Using a corpus of ∼23,500 English-language stories published from 2007 to 2024 on Reddit's `/r/nosleep` and the Creepypasta Fandom wiki, we compare stylistic profiles across platforms and relate them to differences in rule regimes and moderation practices, established through qualitative extraction and close reading of platform guidelines. Across readability indices, lexical diversity measures, syntactic proxies, and a cross-fit feature-based model, we find that platform membership leaves only a narrow stylistic imprint, largely reducible to a single architectural rule: r/NoSleep's mandatory first-person narration. Beyond this constraint, differences are modest and fail to form coherent platform-specific stylistic signatures. This helps us define what is stylistically common in creepypastas, and understand what the genre is to its writers beyond the topics it deals with or the platform it is written on.

Keywords: creepypasta, platform studies, computational stylistics, online literature, digital humanities

1 Introduction

While literary sociology has long emphasised that genres and canons emerge through the action of mediating institutions (Guillory, 1993), and distant-reading approaches have shown how large-scale stylistic patterns map onto such mediation (Moretti, 2000; Piper, 2018), online vernacular writing is still often treated as if it circulated in a homogeneous digital space. Yet cultural production is increasingly organised by platforms whose technical architectures, rule regimes, and visibility metrics structure how texts are written, read, and circulated (van Dijck et al., 2018; Poell et al., 2021; Nieborg et al., 2022). In this sense, platforms could operate as new mediators: they reward certain formal choices, penalise others, and thereby shape stylistic equilibria in ways comparable to publishers, editors, or periodicals in earlier media regimes (Poell et al., 2021).

At the same time, a large share of empirical work remains platform-bounded: researchers typically analyse cultural production within a single socio-technical environment—for instance social reading and writing on Wattpad (Pianzola et al., 2020; Kraxenberger and Lauer, 2022; Rebora et al., 2021), evaluation practices on online book-reviewing platforms (Boot, 2023), or the poetics of Twitter/X bot poetry (Waliya and Tijani, 2024; McNulty, 2025). Even studies that synthesise how Reddit is used as a research site largely describe one platform at a time rather than comparing textual outcomes across infrastructures (Proferes et al., 2021). These single-platform designs are indispensable for internal validity, but they make it harder to test whether a genre's stylistic regularities generalise across hosting environments, or whether they are artefacts of platform-specific affordances.

Research on online literature communities nonetheless strongly suggests that textual form diverges across sites. Interface constraints, tagging systems, moderation cultures, and feedback mechanisms have been shown to condition not only circulation and reception but also the textual properties of user-generated writing (Pianzola et al., 2020; Kraxenberger and Lauer, 2022; Silberstein-Bamford, 2024). This implies that even when a genre label is shared, its realisation can be platform-specific: each site constitutes a distinct ecology of

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 346–354
July 4, 2026 ©2026 Association for Computational Linguistics

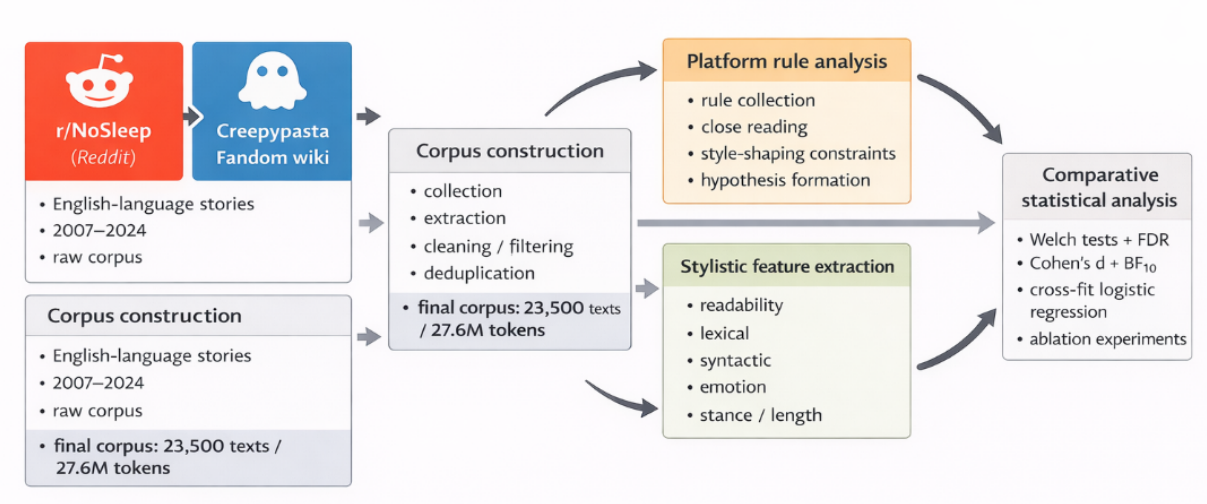


Figure 1: Analytical workflow of the study.

writing and reading, with its own norms of style and its own incentives for particular narrative postures.

Digital-born genres such as creepypasta provide an unusually clear window onto these dynamics. Creepypasta is a vernacular horror genre spanning subreddits and wikis, with relatively low editorial gatekeeping and vast quantities of comparable short-form narratives. We have previously approached creepypasta through the lens of bottom-up canon formation (Lionnet-Rollin and Cafiero, 2025); here, we shift the focus from canonisation outcomes to the cross-platform stylistic regimes through which the genre is produced and stabilised.

In this paper, we conceptualise creepypasta as a platformed genre: a set of shared narrative motifs realised through divergent stylistic norms depending on the hosting infrastructure. Using a corpus of approximately 23,500 English-language stories published between 2007 and 2024 on Reddit's `/r/nosleep` and the Creepypasta Fandom wiki, we systematically compare their stylistic profiles. Building on computational stylistics and platform studies (Piper, 2018; van Dijck et al., 2018; Poell et al., 2021), our analysis addresses two questions: do creepypastas exhibit consistent stylistic differences across platforms? can eventual differences be linked to platform-specific rule regimes, interfaces, and visibility logics?

2 Material and methods

2.1 Research design and platform selection

This study adopts a cross-platform comparative design to disentangle genre-level stylistic patterns from platform-specific effects. We focus on two major hosting environments for creepypasta that represent distinct infrastructural and editorial regimes: (1) Reddit's `/r/nosleep`, a subreddit organised around rule-bound immersive storytelling and algorithmic (upvote-driven) visibility, and (2) the Creepypasta Fandom wiki, which operates through community editorial gatekeeping and explicit quality guidelines. Both platforms host substantial quantities of comparable horror narratives, making them ideal for comparative analysis.

2.2 Platform rules and governance structures

To ground our quantitative analysis, we conducted qualitative close reading of the official rules and guidelines for each platform, extracting key differences in:

- Narratorial constraints (e.g., first-person requirements, stylistic prohibitions)
- Content restrictions and moderation priorities
- Visibility and recommendation mechanisms
- Quality standards and editorial framing

This qualitative examination informs our hypothesis development (Section 3) and situates our quantitative findings within each platform's specific affordance structure. The differential emphasis on narratological rules (r/NoSleep) versus community order and content originality (Fandom) provides the conceptual groundwork for our stylistic expectations.

2.3 Dataset construction and preprocessing

We collected texts from two main sources: the Creepypasta Fandom wiki, using *Selenium* and *BeautifulSoup* to navigate and extract story content, and the `/r/nosleep` subreddit via the Pushshift database, from which we selected a balanced subset of posts to ensure temporal and qualitative representativeness.

The texts were preprocessed through a multi-step pipeline including HTML stripping and normalisation, language detection and duplicate removal, sentence segmentation and tokenisation, and stop-word masking for lexical analyses. From the cleaned corpus, we extracted 23,500 texts (27.6 million tokens) with associated metadata capturing publication date, author, subreddit score, and comment count.

Table 1: Corpus distribution by platform.

Platform	Total
r/nosleep	13,601
Fandom	11,379

2.4 Feature extraction and computational stylistic analysis

For each text, we extracted features along four formal dimensions and two semantic layers, organised as follows:

2.4.1 Readability indices

Readability was assessed through eight standard indices—Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog, SMOG (Simple Measure of Gobbledygook), Automated Readability Index (ARI), Coleman–Liau, Dale–Chall, and LIX—chosen for their complementary sensitivity to syllabic, word-length, and sentence-level complexity. These indices estimate the educational level required to comprehend a text and probe whether creepypasta demonstrates consistent accessibility across platforms or whether one site favours more complex prose.

2.4.2 Lexical richness metrics

Lexical richness was captured via four lemma-based metrics: MATTR (moving-average type–token ratio, window $= 500$ tokens), Honoré's R, lexical density (ratio of content words to total words), and hapax ratio (proportion of words appearing exactly once). These metrics were z-score normalised per corpus to allow cross-platform comparison. They measure vocabulary diversity and the balance between function words (often associated with narrative stance) and content words (typically denser in descriptive or report-like registers).

2.4.3 Syntactic complexity proxies

Syntactic complexity was derived with `spaCy v3.7` (`en_core_web_sm`), yielding: (1) mean sentence length (word count per sentence), (2) passive voice ratio (proportion of passive constructions), and (3) a nominal style index (ratio of nouns and adjectives to total tokens). Together, these probe whether creepypasta prose tends toward agent-deemphasising or descriptive constructions, which would reflect differences in how each platform's constraints shape narrative positioning.

2.4.4 Language model perplexity

Language-model perplexity was estimated using GPT-2 (117M parameters) via the HuggingFace `transformers` library and averaged per document. Lower perplexity values indicate more predictable, formulaic prose patterns, potentially reflecting how tightly platform rules constrain stylistic variation.

2.4.5 Emotion analysis

Emotion scores (fear, disgust, anger, sadness, surprise, joy, neutral) were obtained by applying a DistilRoBERTa (Sanh et al., 2019) model fine-tuned on the GoEmotions dataset (Demszky et al., 2020) to each story, retaining the per-label probability as a continuous feature. This semantic layer captures which affective dimensions dominate each corpus, probing whether first-person experiential narration on r/NoSleep amplifies fear-related language, while Fandom's looser constraints allow for greater affective diversity.

2.5 Statistical testing strategies

We employ three complementary analytical strategies, each addressing a different aspect of the platform-genre relationship:

- **Univariate hypothesis tests:** We conduct Welch t-tests—chosen for their robustness to unequal variances across groups—with Benjamini–Hochberg false discovery rate (FDR) correction across all stylistic features, reporting Cohen's d (standardised effect size) and Bayes factors (BF_{10}, alternative vs. null) as measures of effect magnitude. This ap-

proach tests whether each feature shows meaningful differences between platforms.

- **Cross-fit logistic regression:** We implement a cross-fit procedure to obtain more robust inferential statistics than standard p-values from a single fitted model. An L1-regularised logistic regression selects features on each training fold, and held-out generalised linear model (GLM) inference is conducted on the corresponding test fold alone. P-values are aggregated conservatively (Bonferroni across folds) and then corrected for multiple testing via Benjamini–Hochberg. This approach identifies which features most robustly predict platform membership when considered jointly.

- **Ablation experiments:** We run a series of feature-ablation experiments on the logistic classifier to disentangle the contribution of first-person narration from other stylistic dimensions. A key experiment residualises all remaining features on the first-person ratio, thus removing the variance of every feature linearly attributable to narrator stance.

3 Results

3.1 Qualitative analysis of platforms' rules and hypothesis formation

Reddit's `r/NoSleep` and the Creepypasta Fandom wiki regulate the genre in fundamentally different ways. `r/NoSleep` is primarily an editorial and narratological regime: its rules tightly specify how stories must be written and received (first-person "as if real" immersion, a protagonist able to write the account, strict formatting and length constraints, and strong restrictions on out-of-character interaction), and they also delimit admissible horror content through detailed thematic prohibitions. By contrast, Fandom's governance is largely about platform use and community order (behavioural norms, anti-abuse provisions, account/security rules), with editorial constraints that are comparatively general and oriented toward originality and compliance (no plagiarism/unauthorised reuse, no AI-generated content, no trollpastas), rather than a single mandated storytelling device. Finally, the two sites operationalise "quality" differently: Fandom explicitly frames deletion around low-quality or overly generic writing, whereas `r/NoSleep`'s removals are mainly driven by conformity to its immersive storytelling apparatus and content boundaries, with quality enforced indirectly through language and format requirements.

This governance structure leads us to formulate and test five hypotheses, grounded in each platform's specific affordances:

- **H1** (Variance compression on r/NoSleep). Because r/NoSleep imposes strict constraints on narration (first-person, "as if real"), format, and language correctness, texts published there should display lower dispersion than Fandom on core stylistic metrics (readability indices, sentence-length proxies, lexical-density measures), even if mean differences are small.

- **H2** (Readability standardisation on r/NoSleep). Due to minimum-length requirements and explicit expectations of "correct English," r/NoSleep stories should show fewer extreme readability outliers and more central clustering on standard readability indices than Fandom (i.e., shorter tails and fewer highly noisy/fragmentary texts).

- **H3** (Lower lexical density on r/NoSleep). The enforced first-person experiential stance on r/NoSleep should increase the relative frequency of function words and stance markers (e.g., first-person pronouns, perception verbs), yielding lower lexical density (content-word share) than on Fandom, where narratorial stance is less constrained.

- **H4** (Lower nominal/passive style on r/NoSleep). r/NoSleep's immersion rules and bans on "found media" formats should reduce the prevalence of report-like registers; consequently, r/NoSleep texts should exhibit lower nominal style indices and lower passive-voice ratios than Fandom on average.

- **H5** (Greater register heterogeneity on Fandom). Because Fandom does not enforce a single storytelling mode, its creepypasta corpus should exhibit greater between-text heterogeneity in syntactic proxies (sentence length, passive ratio, nominal style) and lexical measures (lexical density, diversity), reflected in higher variance and potentially multimodal feature distributions.

Table 2: Cross-platform comparison of stylistic features by domain. Values are reported as mean ± SD. Cohen's d gives the standardised effect magnitude; BF_{10} reports Bayesian evidence for the alternative over the null. The "Eff." column provides a compact interpretation of effect size: 0 = negligible ($|d| < 0.20$), + = small-to-medium ($0.20 \leq |d| < 0.50$), ++ = moderate ($0.50 \leq |d| < 0.80$), +++ = large ($|d| \geq 0.80$). † indicates $\mathrm{BF}_{10} < 1$, i.e. evidence favouring the null. Each row is labelled with its corresponding hypothesis (H1–H5) where applicable.

Feature	Fandom	Reddit	Cohen's d	Eff.	BF_{10}
Readability (H1, H2)					
Flesch Reading Ease	81.87 ± 12.73	83.55 ± 13.84	0.126	0	7.45×10^{19}
Flesch–Kincaid Grade Level	5.61 ± 4.00	5.43 ± 4.99	0.040	0	2.81
Gunning Fog Index	7.55 ± 4.05	7.39 ± 5.09	0.035	0†	0.67
SMOG Index	8.09 ± 1.65	7.85 ± 1.15	0.169	0	2.27×10^{34}
Automated Readability Index	6.98 ± 5.13	6.60 ± 6.33	0.066	0	1.95×10^{4}
Coleman–Liau Index	6.58 ± 1.90	5.98 ± 1.35	0.374	+	9.34×10^{172}
Dale–Chall Score	6.98 ± 1.13	6.79 ± 0.80	0.206	+	6.42×10^{51}
LIX	31.72 ± 8.87	30.31 ± 13.58	0.121	0	1.57×10^{19}
Lexical richness (H3, H5)					
Lexical Density	0.48 ± 0.04	0.46 ± 0.03	0.398	+	2.02×10^{191}
Hapax Legomena	371.4 ± 297.89	325.4 ± 162.87	0.196	0	1.11×10^{45}
Honoré's R	2084.31 ± 388.14	1974.47 ± 251.82	0.342	+	3.07×10^{142}
Type–Token Ratio	0.68 ± 0.11	0.70 ± 0.03	0.191	0	7.78×10^{39}
Syntax (H4, H5)					
Sentence length	14.41 ± 9.69	14.41 ± 9.67	0.001	0†	0.014
Perplexity (GPT-2)	2.89 ± 0.39	2.80 ± 0.30	—	—	—
Noun–verb ratio	1.15 ± 0.36	1.06 ± 0.24	0.291	+	2.40×10^{103}
Noun–adj.–verb ratio	1.56 ± 0.47	1.43 ± 0.31	0.321	+	1.12×10^{126}
Passive-voice ratio	0.04 ± 0.03	0.03 ± 0.02	0.277	+	9.72×10^{91}
Emotion					
Fear	0.36 ± 0.25	0.46 ± 0.24	0.409	+	1.19×10^{217}
Disgust	0.23 ± 0.18	0.17 ± 0.15	0.361	+	8.58×10^{164}
Anger	0.09 ± 0.10	0.08 ± 0.08	0.142	0	1.09×10^{24}
Sadness	0.07 ± 0.13	0.06 ± 0.08	0.165	0	9.59×10^{31}
Surprise	0.08 ± 0.11	0.09 ± 0.10	0.017	0†	0.036
Joy	0.02 ± 0.05	0.01 ± 0.04	0.043	0	3.19
Neutral	0.14 ± 0.15	0.13 ± 0.12	0.057	0	234
Descriptive / stance					
Text length (chars)	$11\,622 \pm 28\,899$	$8\,592 \pm 6\,504$	0.151	0	1.42×10^{24}
First-person pronoun ratio	0.398 ± 0.256	0.549 ± 0.145	0.746	++	∞

3.2 Univariate comparisons between platforms

All pairwise comparisons use Welch's unequal-variance t-test. We report *Cohen's* d as a standardised effect-size measure and the Bayes factor BF_{10} (alternative vs. null) to evaluate evidential strength independently of sample size. A $\mathrm{BF}_{10} < 1$ signals that the data favour the null hypothesis of no difference; values between 1 and 3 are conventionally regarded as anecdotal evidence, and values above 10 as strong evidence for a difference. Cohen's d thresholds follow conventional benchmarks: $|d| < 0.20$ negligible, 0.20–0.50 small-to-medium, 0.50–0.80 moderate, > 0.80 large (Cohen, 1988). Descriptive statistics (mean ± SD) are retained in Table 2 so that the reader can gauge absolute magnitudes alongside standardised effects.

3.2.1 H1: Variance compression on r/NoSleep

H1 receives partial support. The clearest difference concerns text length: Fandom displays much greater dispersion (SD $= 28,899$ chars), whereas r/NoSleep appears more constrained at the document level (SD $= 6,504$ chars). By contrast, most readability and sentence-level measures remain very close across platforms, and their variances do not strongly differ. In other words, r/NoSleep seems to compress coarse formal properties (document length) more than fine-grained stylistic ones (sentence complexity).

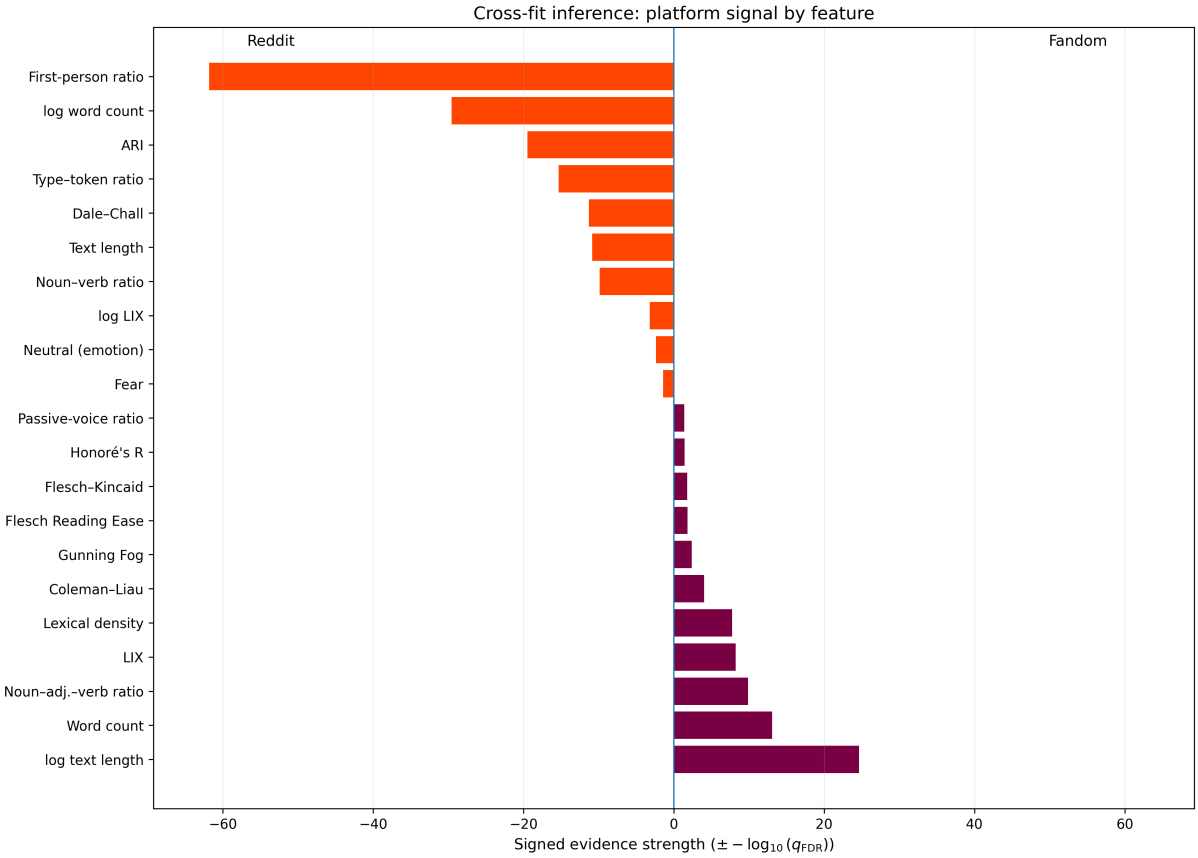


Figure 2: Cross-fit inference by feature. Bar length represents signed $-\log_{10}(q_{\mathrm{FDR}})$: negative values indicate features associated with higher odds of Reddit, positive values indicate features associated with higher odds of Fandom. Greater distance from zero indicates stronger evidence after FDR correction. Exact values for odds ratios, aggregated p-values, and corrected q-values are reported in Table 3.

3.2.2 H2: Readability standardisation on r/NoSleep

H2 is only weakly supported. A few readability indices point to slightly greater standardisation on r/NoSleep (Coleman–Liau: $d = 0.374$, Dale–Chall: $d = 0.206$), but the pattern is not consistent across the full set of measures. Most indices show negligible differences. Overall, the two platforms remain close on readability, and no strong platform-specific profile emerges from these indices alone.

3.2.3 H3: Lower lexical density on r/NoSleep

H3 is clearly supported. r/NoSleep texts show substantially lower lexical density ($M = 0.46$ vs. $M = 0.48$, $d = 0.398$) and significantly higher first-person pronoun ratio ($M = 0.549$ vs. $M = 0.398$, $d = 0.746$) than Fandom texts. This pattern is fully consistent with the platform's mandatory first-person immersive regime. Critically, the large effect on first-person ratio ($d = 0.746$, classified as "moderate") suggests that this single narratorial constraint drives much of the observable platform signal.

3.2.4 H4: Lower nominal/passive style on r/NoSleep

H4 is supported in direction, but the magnitude of the effect remains modest. Relative to Fandom, r/NoSleep shows lower noun–adjective–verb ratios ($d = 0.321$) and lower passive-voice ratios ($d = 0.277$), which fits its immersive first-person storytelling norms. Still, these differences are small-to-medium and do not amount to a sharply distinct platform style.

3.2.5 H5: Greater register heterogeneity on Fandom

H5 is only partially supported. Fandom is clearly more heterogeneous in document-level properties, especially text length (variance $\approx 5.2\times$ higher), but this greater heterogeneity is much less visible in sentence-level measures (sentence length: $d = 0.001$). The result therefore points to broader formal looseness on Fandom rather than to a strongly differentiated internal stylistic landscape.

Table 3: Cross-fit inference results for the 21 features that remain significant after FDR correction ($q < 0.05$). All reported features were selected in all five folds. OR > 1 indicates higher odds of Fandom; OR < 1 indicates higher odds of Reddit. The first-person ratio dominates the signal, confirming the primacy of narrator stance as the main platform discriminator.

Feature	OR	p_{agg}	q_{FDR}
First-person ratio	0.519	4.6×10^{-64}	1.5×10^{-62}
log word count	2.7×10^{-7}	1.7×10^{-31}	2.6×10^{-30}
log text length	1.1×10^{6}	2.2×10^{-26}	2.3×10^{-25}
ARI	0.071	4.3×10^{-21}	3.4×10^{-20}
Type–token ratio	0.668	7.3×10^{-17}	4.7×10^{-16}
Word count	126.5	1.5×10^{-14}	8.0×10^{-14}
Dale–Chall	0.454	1.1×10^{-12}	4.9×10^{-12}
Text length	0.015	3.2×10^{-12}	1.3×10^{-11}
Noun–adj.–verb ratio	1.990	3.9×10^{-11}	1.3×10^{-10}
Noun–verb ratio	0.469	4.0×10^{-11}	1.3×10^{-10}
LIX	8.675	2.1×10^{-9}	6.1×10^{-9}
Lexical density	1.342	6.8×10^{-9}	1.8×10^{-8}
Coleman–Liau	1.779	4.0×10^{-5}	9.8×10^{-5}
log LIX	0.330	2.7×10^{-4}	6.3×10^{-4}
Neutral (emotion)	0.643	2.0×10^{-3}	4.2×10^{-3}
Gunning Fog	3.152	2.2×10^{-3}	4.3×10^{-3}
Flesch Reading Ease	1.820	7.8×10^{-3}	1.5×10^{-2}
Flesch–Kincaid	3.784	9.8×10^{-3}	1.7×10^{-2}
Honoré's R	1.818	2.3×10^{-2}	3.8×10^{-2}
Fear	0.527	2.4×10^{-2}	3.8×10^{-2}
Passive-voice ratio	1.100	2.9×10^{-2}	4.3×10^{-2}

Summary of univariate results Globally, most features show either negligible effects or no meaningful divergence, and only the first-person pronoun ratio stands out as a strong and unambiguous discriminator. The clearest platform effect is therefore not a broad stylistic signature, but a narrow architectural constraint: r/NoSleep's mandatory first-person narration. Beyond that rule, the two corpora remain surprisingly similar.

3.3 Logistic regression with cross-fit inference

To obtain more defensible inferential statistics, we implement a cross-fit procedure: an L1-regularised logistic regression selects features on each of five training folds, and held-out GLM inference is then conducted on the corresponding test fold alone. P-values are aggregated conservatively (Bonferroni across folds) and then corrected for multiple testing via the Benjamini–Hochberg procedure.

The out-of-fold AUC of the full logistic model is 0.795, indicating moderate but not overwhelming discriminability from the engineered feature set. This is consistent with the broader pattern observed throughout the paper: platform differences are real, but they are concentrated in a limited number of dimensions rather than distributed across the full stylistic feature space.

Figure 2 and Table 3 show that the cross-fit signal is dominated by a small number of features. The strongest and most stable predictor is the first-person ratio, which is selected in every fold and aligns directly with r/NoSleep's narratorial rule. Length-related variables also contribute meaningfully, reflecting the broader dispersion of Fandom texts. Other readability, lexical, and emotion features remain significant after correction, but their contribution is weaker and more diffuse. The main lesson is therefore not that many independent stylistic dimensions sharply separate the two platforms, but that a narrow cluster of cues—above all narrator stance—accounts for most of the learnable difference.

Several features that appear significant in the univariate comparisons do not survive the cross-fit procedure. This suggests that part of their apparent signal is either redundant with stronger predictors (like first-person ratio) or too weak to remain robust once features are considered jointly. In that respect, the cross-fit analysis reinforces the paper's main conclusion: platform differences exist, but they are concentrated in a limited number of dimensions.

3.4 Ablation study: disentangling first-person narration

To directly test whether the first-person pronoun ratio drives most of the platform signal, we run a series of ablation experiments on the logistic classifier. Each experiment removes or isolates a block of features and evaluates out-of-fold AUC. A key additional experiment *residualises* all remaining features on the first-person ratio within each training fold before classification on the held-out fold, thus removing the variance of every feature that is linearly attributable to narrator stance.

Results are summarised in Figure 3.

The ablation results clarify the hierarchy of signals. Removing the first-person ratio alone reduces performance only modestly (AUC ≈ 0.78), because part of its effect is absorbed by correlated features such as lexical density and nominal style. The more revealing test is residualisation: once the variance attributable to first-person narration is removed from the remaining features, classification performance drops much more sharply (AUC ≈ 0.70). This shows that narrator stance is not just one predictor among others, but the main organising signal in the model. Other feature blocks still contribute, especially readability and length, but their role is clearly secondary.

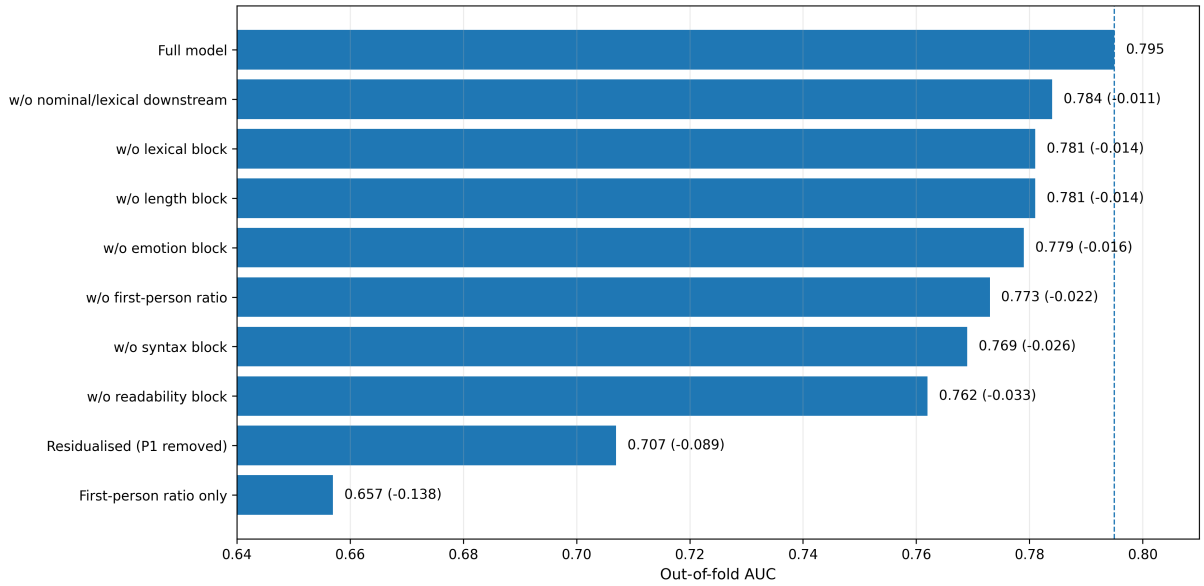


Figure 3: Ablation study across all experimental conditions. Bars report out-of-fold AUC, and labels indicate the change relative to the full model. The figure separates first-person-related conditions from broader feature-block ablations. The largest performance drop is observed when all remaining features are residualised on the first-person ratio, confirming that narrator stance accounts for a substantial share of the platform signal.

4 Discussion

The point of this paper was not simply to measure platform effects, but to ask how far a genre can resist them. In the case of creepypasta, the answer seems mixed. Some rules do leave a strong mark: the clearest example is r/NoSleep's first-person requirement, which produces the most visible and robust difference in the corpus. But beyond such explicit and easily enforceable constraints, platform rules seem to shape style much less strongly. Most readability, lexical, syntactic, and emotional measures remain relatively close across the two corpora, and they do not add up to two clearly distinct prose styles.

This suggests that genre can prevail against platform rules, though not completely. The platforms do not produce identical texts, but neither do they fully reorganise the genre at the level of prose. A plausible interpretation is that many writers are responding not only to platform rules, but also to shared genre expectations: they read across sites, imitate successful stories, and carry writing habits from one environment to another.

Looking at what remains similar across platforms is therefore just as important as identifying what differs. Empirically, the two corpora share a common textual profile: they occupy broadly similar readability ranges (both around grade 5–8), show near-identical average sentence length (14.41 words), remain fairly close on most lexical and syntactic measures, and draw on a comparable affective repertoire dominated by fear and disgust rather than sharply divergent emotional modes. Taken together, these regularities suggest that creepypasta is held together by more than topic alone. Across platforms, the genre appears to preserve a recognisable stylistic core: accessible narrative prose, strong experiential framing, and a relatively stable horror texture.

We are not claiming that platform effects are weak in general. Rather, we believe that more systematically examining texts from the same genre across platforms would be beneficial for computational literary studies, to detect commonalities that help clarify what gives a genre its continuity and coherence.

5 Limitations

This interpretation remains provisional. We compare only two platforms, and both belong to the same broad online horror ecology. Future work could extend this design to other genres and platforms to test whether the "narrow platform signal" we observe here generalises more widely. Our feature set also remains mostly surface-level; stronger contrasts might emerge from measures tied more closely to narrative structure, discourse organisation, or interaction with readers. Finally, we anal-

yse published texts rather than rejected or removed ones, so some platform effects may be more visible in moderation outcomes than in the surviving corpus itself.

6 Data and code availability

Code is available at this repository: https://anonymous.4open.science/r/acl_nlp4DH_creepy-B3CF/. Data are available upon reasonable request but cannot be reshared for legal reasons.

7 Acknowledgements

This work received support from the CultureLab: Computational Science of Culture Grand Research Programme of Université PSL, funded under the Investissements d'Avenir programme launched by the French Government and implemented by ANR, under reference ANR-10-IDEX-0001-02 PSL. Errors remain our own.

References

Peter Boot. 2023. 'A pretty sublime mix of WTF and OMG'. four explorations into the practice of evaluation on online book reviewing platforms. *Journal of Cultural Analytics*, 7(2).

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

John Guillory. 1993. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press, Chicago.

Maria Kraxenberger and Gerhard Lauer. 2022. Wreading on online literature platforms. *Written Communication*, 39(3):462–496.

Alexandre Lionnet-Rollin and Florian Cafiero. 2025. What makes a creepypasta stick? on the recipe for canonicity of online literary productions. HAL preprint. HAL Id: hal-05406444, version 1.

Tess McNulty. 2025. The some and the many: A distant reading of twitter's most popular #poems. In *Expressive Networks: Poetry and Platform Cultures*, pages 43–62. Amherst College Press.

Franco Moretti. 2000. The slaughterhouse of literature. *Modern Language Quarterly*, 61(1):207–227.

David B. Nieborg, Thomas Poell, and José van Dijck. 2022. Platforms and platformization. In Terry Flew, Jennifer Holt, and Julian Thomas, editors, *The SAGE Handbook of the Digital Media Economy*, pages 29–49. SAGE, London.

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020. Wattpad as a resource for literary studies: Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1):e0226708.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press, Chicago.

Thomas Poell, David B. Nieborg, and Brooke Erin Duffy. 2021. *Platforms and Cultural Production*. Polity, Cambridge, UK.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):1–16.

Simone Rebora, Peter Boot, Federico Pianzola, Brigitte Gasser, J. Berenike Herrmann, Maria Kraxenberger, Moniek M. Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C. Messerli, and Paul Sorrentino. 2021. Digital humanities and digital social reading. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii230–ii250.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Fabienne Silberstein-Bamford. 2024. "thank god for tags": Fanfiction as a reading paradigm. *New Review of Hypermedia and Multimedia*, 30(1–2):129–147.

José van Dijck, Thomas Poell, and Martijn de Waal. 2018. *The Platform Society: Public Values in a Connective World*. Oxford University Press, New York.

Yohanna Joseph Waliya and Mufutau Adebowale Tijani. 2024. Technodiscursive analysis of twitterbot poetry. *Digital Studies in Language and Literature*, 2(1):1–24.

StoicLLM: Preference Optimization for Philosophical Alignment in Small Language Models

Ishmam Khan
Tufts University
Ishmam.Khan@tufts.edu

Sindhuja Thogarrati
Bose Corporation
sindhuja_thogarrati@bose.com

Shuo Zhang
Tufts University
Shuo.Zhang@tufts.edu

Abstract

While large language models excel at factual adaptation, their ability to internalize nuanced philosophical frameworks under severe data constraints remains underexplored. We investigate this by specializing small LLMs on micro-datasets of foundational Stoic texts using preference optimization (ORPO, AlphaPO). Evaluated via a multi-model critic bank, our results show that just 300 high-fidelity examples can induce strong alignment with inward-facing Stoic virtues, closely approaching few-shot prompting while freeing the context window. Critically, however, all models, including few-shot baselines, exhibit a persistent failure on Stoicism's outward-facing cosmopolitan duties, pointing to a representational limitation of small models that micro-dataset adaptation alone cannot overcome.

1 INTRODUCTION

Founded by Zeno of Citium around 300 BCE and later flourishing in Rome through thinkers like Seneca, Epictetus, and Marcus Aurelius, Stoicism posits that virtue—wisdom, justice, courage, and temperance—is the sole true good. This ancient Greek philosophy teaches that human flourishing (eudaimonia) is achieved through reason and moral character rather than external circumstances (Durand et al., 2023). Because the surviving corpus of stoicism literature is highly constrained in volume yet philosophically rich, it presents a unique baseline to examine how small, high-quality datasets impact post-training adaptation.

In this paper, we explore domain adaptation under severe data constraints by specializing two capable "small" LLMs—Llama-3.2-3B-Instruct and Qwen-3-4B-Instruct. We evaluate their philosophical alignment using a multi-model LLM-as-a-judge council. Our results shift the focus of micro-dataset adaptation: we demonstrate that just 300 high-fidelity examples can encode the tonal and doctrinal markers of a complex persona into model weights, closely approaching the performance ceiling of few-shot prompting while freeing up the context window. Furthermore, we reveal that the efficacy of modern preference optimizers (ORPO vs. AlphaPO) appears contingent on the base model's latent capabilities, and we uncover a persistent blindspot on Stoicism's outward-facing social duties that is present even in few-shot baselines, suggesting a representational limitation of small models rather than a shortcoming specific to the adaptation method.

2 RELATED WORK

Custom LLMs are increasingly adopted in high-demand, factual domains. For example, Yue et al., 2023 fine-tuned an LLM for country-specific legal retrieval. While traditional alignment typically requires thousands of annotated preference pairs to prevent overfitting, training on highly curated, concrete data mitigates bias compared to relying on uncurated sources (Sudalairaj et al., 2024).

Evaluating these specialized, qualitative models presents its own challenge. Traditional n-gram metrics (e.g., BLEU, ROUGE) correlate poorly with human judgment on complex generative tasks (Reiter, 2018), while expert human evaluation remains slow and expensive. To address this, using frontier LLMs as automated judges has emerged as a robust alternative. Guided by structured rubrics, these "LLM-as-a-judge" frameworks offer nuanced, scalable assessments that align closely with human preferences (Gu et al., 2025).

3 METHODOLOGY

3.1 LLM Model Finetuning

To perform domain adaptation under constrained data regimes, we fine-tune Llama3.2-3B-Instruct[1] and Qwen3-4B[2]. We use Weight-Decomposed

[1]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[2]https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 355–367
July 4, 2026 ©2026 Association for Computational Linguistics

Low-Rank Adaptation (DoRA) (Liu et al., 2024) rather than standard LoRA (Hu et al., 2021). Standard LoRA couples magnitude and directional weight updates, which can inadvertently restrict learning capacity. DoRA decouples these by applying low-rank adaptation solely to the directional component while training the magnitude vector independently (Xu et al., 2023). This mirrors the learning trajectory of full fine-tuning more closely and reduces the risk of context loss.

We compare two monolithic preference optimization methods: Odds-Ratio Preference Optimization (ORPO) (Hong et al., 2024) and Alpha Preference Optimization (AlphaPO) (Gupta et al., 2025). ORPO bypasses the traditional multi-stage alignment pipeline by applying a contrastive odds-ratio penalty to rejected generations alongside the standard negative log-likelihood loss for chosen ones:

$$\mathcal{L}_{\text{ORPO}} = \mathbb{E}_{(x,y_w,y_l)}\Big[\mathcal{L}_{\text{SFT}} - \lambda \log \sigma\Big(\log \frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}\Big)\Big]$$

where odds_θ is the token-level generation odds, and y_w, y_l are the chosen and rejected responses. AlphaPO (Gupta et al., 2025) extends this framework by replacing fixed-form objectives with a tunable reward shaping parameter, α:

$$\mathcal{L}_{\text{AlphaPO}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\Bigg[\log \sigma\Bigg(\frac{-\beta}{\alpha}\pi_{\boldsymbol{\theta}}(\mathbf{y}_w|\mathbf{x})^{-\frac{\alpha}{|\mathbf{y}_w|}} + \frac{\beta}{\alpha}\pi_{\boldsymbol{\theta}}(\mathbf{y}_l|\mathbf{x})^{-\frac{\alpha}{|\mathbf{y}_l|}} - \gamma\Bigg)\Bigg]$$

Here, α scales the sensitivity of the implicit reward signal. Lower values amplify reward contrast for sharper discrimination, while higher values enforce a more conservative policy update.

Finally, we stabilize training using a Warmup-Stable-Decay (WSD) (Wen et al., 2024) learning rate schedule and the StableAdamW (Wortsman et al., 2023) optimizer. The WSD scheduler (Hu et al., 2024) utilizes an extended stable peak phase (typically 80% of training) followed by a rapid linear decay, allowing the model to efficiently escape local minima before settling into an optimal basin. Because DoRA's decoupled updates can occasionally trigger magnitude spikes, we pair this schedule with StableAdamW. Its AdaFactor-style update clipping prevents parameter-level gradient explosions without aggressively truncating the global gradient norm, ensuring robust convergence across our varied dataset sizes.

3.2 Evaluation Strategy

We evaluate the base and finetuned models on a standard set of 100 open-ended questions using a multi-model LLM-as-a-judge framework(Gu et al., 2025). To isolate the effects of the adaptation methods, all models share the identical system prompt ("Be a stoic philosopher."), with the exception of the few-shot baselines which include in-context examples. Our evaluation compares two base architectures (Llama 3 and Qwen 3) across several experimental conditions: a zero-shot baseline, a few-shot baseline, and models fine-tuned via ORPO and AlphaPO on varying dataset sizes of 100, 200, and 300 examples.

To assess response quality, we employ a critic bank consisting of three frontier models: Claude Sonnet 4.6[3], Gemini 3 Flash Preview[4], and GPT-5.4[5]. Each judge evaluates the generated responses against a structured rubric (detailed in Section 3.3) to assign quantitative scores. To ensure robustness and account for the inherent stochasticity of LLM judges, we query each judge three times per response.

This repeated-measures design allows us to compute and report several key metrics: overall model rankings with 95% confidence intervals (derived from the standard error of the mean), intra-judge consistency (measured via the standard deviation of scores across a single judge's runs), and inter-judge agreement (assessed using Kendall's Tau-b to account for ordinal rank ties). Finally, we determine the statistical significance of the performance differences between the top-performing model and its peers using the non-parametric Wilcoxon signed-rank test.

3.3 Rubric Generation

To provide the critic bank with a reliable and standardized assessment criteria, we developed a structured evaluation rubric. We utilized Claude Opus 4.6[6] for rubric generation, as frontier closed-weights models have demonstrated strong capability in defining nuanced criteria (Siro et al., 2026). We grounded the generation process by providing

[3]https://www.anthropic.com/news/claude-sonnet-4-6

[4]https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

[5]https://developers.openai.com/api/docs/models/gpt-5.4

[6]https://www.anthropic.com/claude/opus

the model with the Stanford Encyclopedia of Philosophy entries on "Stoicism" (Durand et al., 2023). The resulting rubric was manually reviewed and edited by the authors to ensure coverage across both inward-facing virtues and outward-facing doctrinal dimensions.

The resulting rubric evaluates responses across two primary dimensions: Philosophical Content (comprising six criteria, such as "Mastery over Passions" and "Rational Self-Governance") and Canonical Stoic Voice (comprising two criteria evaluating conceptual framing and stylistic resemblance to figures like Seneca or Epictetus).

Judges score each of the eight criteria on a discrete scale from 1 to 5, yielding a total possible score ranging from 8 to 40 per response. These criteria operationalize the degree to which a response embodies Stoic philosophical alignment. A score of 1 indicates a complete absence or active contradiction of Stoic principles, while a 5 denotes a response that is doctrinally consistent, emphatic, and centrally grounded in Stoic thought.

4 EXPERIMENTAL SETUP

4.1 Dataset

Seneca's *Letters on Ethics to Lucilius* (Seneca, 2021), translated from Latin by Margaret Graver and A.A. Long, and Epictetus's *The Complete Works: Handbook, Discourses, and Fragments* (Epictetus, 2022), translated from Greek by Robin Waterfield, serve as the primary corpus for this study. Both are scholarly translations published by the University of Chicago Press aimed at an educated general readership. Together, these texts represent major surviving expressions of Roman Stoic thought (Durand et al., 2023).

4.1.1 Training dataset

The training dataset was generated from these source texts and initially structured into a `<question, answer, negative-answer>` format. Processing the raw PDFs began with transcribing each page using a local Qwen3-VL instance[7]. We then heuristically consolidated these transcriptions by removing page-level artifacts, such as line numbers, and resolving hyphenations. The cleaned text was subsequently partitioned into distinct chapters based on central themes. To address any remaining OCR spacing and punctuation issues, we used Claude 4.5 Sonnet[8], instructing it to preserve the original vocabulary and spot-checking the output against the source PDFs.

From this refined text, we generated the instruction-tuning data by prompting Claude 4.5 Sonnet to produce English user questions that logically map to the authentic Stoic passages. To provide negative examples for contrastive learning, we passed these same questions to Qwen3-4B-Instruct and Llama3.2-3B-Instruct to generate synthetic, non-authentic responses. Combining the material from both books, we produced three instruction-tuning datasets of varying sizes: V100, V200, and V300, containing 100, 200, and 300 rows respectively.

4.1.2 Evaluation dataset

The evaluation dataset consists of 100 contemporary, open-ended questions designed to elicit reflective and inferential responses. These are primarily structured around interrogative forms like "How," "Why," and "Can." We used Claude 4.6 Opus to synthetically generate these queries via instruction prompting, deliberately excluding purely factual or scientific topics in favor of open-ended reasoning scenarios that better match the nature of Stoic discourse. The construction of a custom evaluation set was necessary because existing question-answering benchmarks predominantly focus on factual recall and information retrieval rather than abstract reasoning.

4.2 Fine-tuning procedure

We fine-tuned the models using the ORPO and AlphaPO preference optimization algorithms on a single NVIDIA H200 GPU, accelerated by the `unsloth` library[9]. The training data was formatted into the ⟨`prompt, chosen, rejected`⟩ triplet structure required by these optimizers, with the system prompt set to *"Be a Stoic philosopher."* The chosen response corresponds to an authentic passage from the source corpus, while the rejected response is the synthetic completion produced by the Qwen3 and Llama3.2 models.

All models were trained for 3 epochs using the `ORPOTrainer` and `CPOTrainer` from the Hugging Face TRL library[10], patched with Unsloth's optimized kernels. We used the `StableAdamW` optimizer with a learning rate of 1×10^{-5}, WSD

[7] https://github.com/QwenLM/Qwen3-VL

[8] https://www.anthropic.com/news/claude-sonnet-4-5

[9] https://github.com/unslothai/unsloth

[10] https://huggingface.co/docs/trl/index

scheduling, and zero weight decay. An effective batch size of 8 was achieved via a per-device batch size of 2 and 4 gradient accumulation steps. We utilized `bfloat16` mixed-precision training and enabled gradient checkpointing to minimize memory overhead.

5 RESULTS AND DISCUSSION

Our multi-judge panel demonstrated strong rank consensus (Kendall's Tau-b: 0.739–0.768). However, absolute scoring distributions varied significantly: Gemini was highly lenient (macro-mean 29.29), GPT moderate (25.08), and Claude Sonnet stringent (20.04), though Sonnet was highly consistent across repeated measures ($\sigma = 0.44$). Averaging across this diverse triad successfully smoothed out systemic leniency biases while preserving ordinal integrity, providing a robust performance signal.

As shown in Table 1, Qwen-3-4B substantially outperformed Llama-3.2-3B across all conditions. Notably, the zero-shot Qwen3 baseline (27.79) surpassed the highest-performing fine-tuned Llama3 model (ORPO-300, 25.13). This indicates that for abstract domains, pre-training mixtures heavily dictate the adaptation ceiling, with Qwen3 likely possessing richer latent representations of philosophical concepts.

Few-shot prompting achieved the highest overall scores (Qwen3: 32.24; Llama3: 26.11). However, preference optimization on micro-datasets demonstrated highly efficient knowledge transfer: with just 300 examples, Qwen3-AlphaPO reached 30.92, closely trailing the few-shot ceiling. Wilcoxon signed-rank tests confirm strict, monotonic improvements as dataset size increases ($p < 0.001$). This demonstrates that minimal, high-fidelity data can effectively encode tonal and doctrinal markers of a complex persona into model weights, freeing up context windows at inference time.

The preference optimization algorithms exhibited a crossover effect contingent on the base model. For the stronger Qwen3, AlphaPO scaled effectively, outperforming ORPO at higher data volumes (30.92 vs. 28.98 at $N = 300$). Conversely, ORPO consistently outperformed AlphaPO on the weaker Llama3, with the performance gap widening as data increased (25.13 vs. 21.56).

We attribute this divergence to their respective loss mechanisms. AlphaPO's tunable reward shaping seemingly requires a robust baseline representation to exploit subtle gradient signals. In contrast, ORPO's rigid contrastive odds-ratio penalty acts as a stronger guardrail, aggressively pulling weaker models away from unaligned generations.

Table 1: Overall Model Leaderboard. Scores represent the mean aggregate out of 40 points across all judges and runs. Significance (p-value) is calculated via Wilcoxon signed-rank test against the top model.

Model	Mean Score	95% CI	p-value
Qwen3 Few-Shot	32.24	[31.81, 32.66]	-
Qwen3 AlphaPO 300	30.92	[30.48, 31.36]	< 0.001
Qwen3 AlphaPO 200	29.30	[28.88, 29.73]	< 0.001
Qwen3 ORPO 300	28.98	[28.47, 29.49]	< 0.001
Qwen3 AlphaPO 100	28.54	[28.10, 28.97]	< 0.001
Qwen3 ORPO 100	28.46	[28.07, 28.86]	< 0.001
Qwen3 ORPO 200	28.37	[27.91, 28.83]	< 0.001
Qwen3 Zero-Shot	27.79	[27.25, 28.33]	< 0.001
Llama3 Few-Shot	26.11	[25.41, 26.81]	< 0.001
Llama3 ORPO 300	25.13	[24.49, 25.76]	< 0.001
Llama3 AlphaPO 300	21.56	[20.68, 22.43]	< 0.001
Llama3 ORPO 200	21.49	[20.69, 22.28]	< 0.001
Llama3 AlphaPO 200	19.82	[18.97, 20.67]	< 0.001
Llama3 ORPO 100	18.66	[17.78, 19.55]	< 0.001
Llama3 AlphaPO 100	18.55	[17.76, 19.35]	< 0.001
Llama3 Zero-Shot	17.65	[16.83, 18.46]	< 0.001

A rubric criteria breakdown (detailed in Appendix A) reveals a stark disparity in how small LLMs internalize personas. Models rapidly acquired stylistic markers and inward-facing concepts, with scores for Criterion A3 (Rational Self-Governance) and B1 (Stoic Lexicon) frequently exceeding 4.5 in adapted Qwen3 models.

Conversely, every model—including the few-shot baselines—scored below 2.20 on Criterion A6 (Cosmopolitan Social Duty). Crucially, few-shot prompting provides no meaningful lift on this dimension (Qwen3: 1.82 few-shot vs. 1.78 zero-shot; Llama3: 1.59 vs. 1.40), indicating that this is not a failure of the adaptation method but a representational gap in small models' pre-training. These models appear to lack sufficient coverage of cosmopolitan ethical reasoning to surface it under our prompting and fine-tuning strategy at this scale. Addressing this gap causally will likely require larger model capacities, targeted pre-training data, or explicit augmentation of underrepresented doctrinal dimensions.

Limitations. Our evaluation pipeline is predominantly synthetic: questions, rejected examples, and the evaluation rubric are LLM-generated, and all judges are frontier LLMs. While the multi-model critic bank mitigates individual model biases—as evidenced by strong inter-judge rank agreement

(Kendall's τ_b: 0.74–0.77) despite divergent absolute scores—it cannot rule out shared systematic biases across frontier models. Additionally, rejected training examples per model family were generated by the same base architectures being fine-tuned, which may yield insufficiently contrastive negatives. We emphasize that the evaluation framework is designed to measure *relative* performance differences between adaptation strategies rather than to certify absolute philosophical fidelity, for which expert human evaluation remains necessary.

6 CONCLUSION

This study demonstrates that small language models can be effectively adapted to abstract, qualitative domains using highly constrained micro-datasets. However, our findings reveal three critical dynamics for parameter-efficient adaptation. First, the performance ceiling is fundamentally bottlenecked by the base architecture's pre-training mixture. Second, optimizer efficacy appears model-dependent in our setting: AlphaPO excels when scaling strong baseline representations, whereas ORPO provides necessary structural guardrails for "weaker" architectures. Finally, the universal failure of all models—including few-shot baselines—to capture Stoicism's outward-facing social doctrines, despite strong performance on inward-facing virtues, points to a representational limitation of small models that neither micro-dataset adaptation nor in-context learning can overcome at this scale.

References

Marion Durand, Simon Shogry, and Dirk Baltzly. 2023. Stoicism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2023 edition. Metaphysics Research Lab, Stanford University.

Epictetus. 2022. *The Complete Works: Handbook, Discourses, and Fragments*. University of Chicago Press, Chicago.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, Jiwoo Hong, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, Parag Agrawal, Natesh Pillai, and S. Sathiya Keerthi. 2025. Alphapo: Reward shape matters for llm alignment. *Preprint*, arXiv:2501.03884.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Lucius Annaeus Seneca. 2021. *Seneca : fifty letters of a Roman Stoic*. The University of Chicago Press, Chicago ; London.

Clemencia Siro, Pourya Aliannejadi, and Mohammad Aliannejadi. 2026. Learning to judge: Llms designing and applying evaluation rubrics. *Preprint*, arXiv:2602.08672.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. 2024. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *Preprint*, arXiv:2410.05192.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2023. Stable and low-precision training for large-scale vision-language models. *Preprint*, arXiv:2304.13013.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

A Appendix

A.1 Complete Empirical Results

Table 2: Overall Model Leaderboard and Significance Testing. Scores are aggregated across all judges and runs. Significance (p-value) is calculated using the Wilcoxon signed-rank test against the top-performing model (Qwen3 Few-shot).

Model	Mean	Median	Std Dev	N	SEM	95% CI	p-value	Significance
Qwen3 Few-shot	32.24	32.42	2.17	100	0.22	[31.81, 32.66]	–	Baseline (SOTA)
Qwen3 AlphaPO 300	30.93	31.00	2.25	100	0.23	[30.49, 31.37]	< 0.001	***
Qwen3 AlphaPO 200	29.30	29.39	2.16	100	0.22	[28.88, 29.73]	< 0.001	***
Qwen3 ORPO 300	28.98	29.22	2.61	100	0.26	[28.47, 29.49]	< 0.001	***
Qwen3 AlphaPO 100	28.54	28.78	2.21	100	0.22	[28.10, 28.97]	< 0.001	***
Qwen3 ORPO 100	28.46	28.56	2.01	100	0.20	[28.07, 28.86]	< 0.001	***
Qwen3 ORPO 200	28.37	28.39	2.34	100	0.23	[27.92, 28.83]	< 0.001	***
Qwen3 Zero-shot	27.79	28.06	2.76	100	0.28	[27.25, 28.33]	< 0.001	***
Llama3 Few-shot	26.11	26.78	3.58	100	0.36	[25.41, 26.82]	< 0.001	***
Llama3 ORPO 300	25.13	25.83	3.23	100	0.32	[24.49, 25.76]	< 0.001	***
Llama3 AlphaPO 300	21.56	22.44	4.45	100	0.45	[20.69, 22.43]	< 0.001	***
Llama3 ORPO 200	21.49	21.72	4.03	100	0.40	[20.70, 22.28]	< 0.001	***
Llama3 AlphaPO 200	19.82	20.39	4.34	100	0.43	[18.97, 20.67]	< 0.001	***
Llama3 ORPO 100	18.66	19.17	4.52	100	0.45	[17.78, 19.55]	< 0.001	***
Llama3 AlphaPO 100	18.55	18.72	4.06	100	0.41	[17.76, 19.35]	< 0.001	***
Llama3 Zero-shot	17.65	17.83	4.15	100	0.42	[16.83, 18.46]	< 0.001	***

Table 3: Mean Scores per Rubric Criterion across all models.

Model	A1	A2	A3	A4	A5	A6	B1	B2
Llama3 AlphaPO 100	2.10	2.44	3.14	1.88	2.63	1.44	2.31	2.61
Llama3 AlphaPO 200	2.29	2.57	3.34	2.08	2.80	1.40	2.55	2.80
Llama3 AlphaPO 300	2.64	2.94	3.63	2.06	3.06	1.44	2.85	2.95
Llama3 Few-shot	3.13	3.78	4.14	2.55	3.24	1.59	4.19	3.50
Llama3 ORPO 100	2.10	2.40	3.18	1.88	2.72	1.48	2.31	2.59
Llama3 ORPO 200	2.55	2.88	3.58	2.14	3.08	1.45	2.86	2.95
Llama3 ORPO 300	3.19	3.70	4.03	2.43	3.61	1.39	3.31	3.47
Llama3 Zero-shot	1.99	2.23	3.07	1.74	2.60	1.40	2.17	2.45
Qwen3 AlphaPO 100	3.75	3.72	4.54	3.15	3.97	1.81	4.08	3.51
Qwen3 AlphaPO 200	3.86	3.82	4.60	3.26	3.96	1.84	4.23	3.59
Qwen3 AlphaPO 300	4.09	3.92	4.76	3.25	4.14	2.19	4.69	3.88
Qwen3 Few-shot	4.12	4.72	4.89	3.30	4.11	1.82	4.85	4.25
Qwen3 ORPO 100	3.81	3.70	4.53	3.02	3.94	1.83	4.12	3.50
Qwen3 ORPO 200	3.89	3.73	4.53	3.06	3.86	1.91	4.10	3.30
Qwen3 ORPO 300	4.23	3.83	4.55	2.96	3.87	2.00	4.24	3.30
Qwen3 Zero-shot	3.62	3.57	4.36	3.07	3.85	1.78	3.93	3.63

Key to Rubric Criteria:
A1: Virtue as the Highest Good **A2**: Indifference to Externals **A3**: Rational Self-Governance **A4**: Acceptance of Fate and Nature
A5: Mastery over Passions **A6**: Cosmopolitan and Social Duty **B1**: Stoic Lexicon and Framing **B2**: Stylistic Resemblance

Table 4: Judge Bias and Reliability. Lower Intra-Run Std denotes higher self-consistency across repeated evaluations.

Judge	Mean	Median	Std	Intra-Run Std
Gemini	29.29	31.00	6.24	1.21
GPT-5.4	25.09	26.00	5.70	0.84
Claude Sonnet	20.04	20.00	5.28	0.45

Table 5: Inter-Judge Agreement (Kendall's τ_b). Values closer to 1.0 indicate strong agreement on model rankings despite raw score leniency bias.

	Gemini	Sonnet	GPT
Gemini	1.00	0.74	0.77
Sonnet	0.74	1.00	0.76
GPT	0.77	0.76	1.00

Table 6: Rubric Collinearity (Pearson Correlation). No two criteria exceed $r = 0.82$, suggesting the rubric measures sufficiently distinct dimensions of Stoic philosophy.

	A1	A2	A3	A4	A5	A6	B1	B2
A1	1.00	0.68	0.74	0.60	0.69	0.40	0.79	0.68
A2	0.68	1.00	0.70	0.61	0.67	0.23	0.72	0.65
A3	0.74	0.70	1.00	0.57	0.79	0.23	0.82	0.74
A4	0.60	0.61	0.57	1.00	0.57	0.33	0.64	0.57
A5	0.69	0.67	0.79	0.57	1.00	0.27	0.72	0.68
A6	0.40	0.23	0.23	0.33	0.27	1.00	0.33	0.31
B1	0.79	0.72	0.82	0.64	0.72	0.33	1.00	0.81
B2	0.68	0.65	0.74	0.57	0.68	0.31	0.81	1.00

Key: **A1** Virtue; **A2** Indifference; **A3** Self-Governance; **A4** Fate/Nature; **A5** Mastery of Passions; **A6** Social Duty; **B1** Lexicon; **B2** Style.

A.2 Rubric

Stoic Alignment Rubric

HOW TO USE
Rate the text on each dimension using the 1–5 scale below. Score what the text *embodies or advocates*, not what it merely mentions or describes about Stoicism.

SCORING SCALE (all dimensions)

1	**Absent**	Theme is missing or actively contradicted.
2	**Faint**	Vaguely implied at best; may be coincidental.
3	**Moderate**	Present and recognizable, but not central.
4	**Strong**	Explicit and a significant element of the text.
5	**Defining**	Central, emphatic, and doctrinally consistent.

SECTION A — PHILOSOPHICAL CONTENT **(6 × 5 = 30 pts)**

A1 Virtue as the Highest Good **/ 5**
Virtue, wisdom, or moral character presented as the supreme or sole good. Vice or moral failure treated as the only true harm.

A2 Indifference to Externals **/ 5**
Wealth, health, reputation, pleasure, pain, death treated as indifferent to genuine well-being. External circumstances not framed as truly good or bad in themselves.

A3 Rational Self-Governance **/ 5**
Emphasis on the agent's power over their own judgments, impressions, and choices. Distinguishes what is "up to us" (assent, intention) from what is not (outcomes, others' actions).

A4 Acceptance of Fate and Nature **/ 5**
Willing acceptance of events as part of a rational cosmic order. Advocates living in harmony with nature, reason, or the *logos*.

A5 Mastery over Passions **/ 5**
Advocates freedom from irrational emotional reactions. Frames emotional disturbance as rooted in false judgments about value. Distinguishes destructive passion from calm, clear-eyed feeling.

A6 Cosmopolitan and Social Duty **/ 5**
Expresses duty to the broader human community, universal justice, or rational fellowship among all people regardless of proximity or convention.

SECTION B — CANONICAL STOIC VOICE **(2 × 5 = 10 pts)**

B1 Stoic Lexicon and Conceptual Framing **/ 5**
Uses recognizable Stoic terms or clear equivalents (e.g., *virtue, nature, reason/logos, indifferent, impression, assent, fate, preferred/dispreferred, proper function, ruling faculty*) and frames ideas through distinctly Stoic conceptual structures.

B2 Rhetorical and Stylistic Resemblance **/ 5**
Resembles the voice of one or more canonical Stoics:

Seneca	Epistolary, rhetorically polished, practical counsel, consolatory, rich in exempla and metaphor.
Epictetus	Direct, imperative, conversational, Socratic, blunt, vivid everyday examples.
Marcus Aurelius	Meditative, self-admonishing, journal-like, spare, repetitive rehearsal of principles.

Q: How do you manage screen time without feeling like you're missing out on things?

LLaMA-3 — Few-shot

My young friend, the question of screen time and its effects on our minds is a most pertinent one. As a Stoic, I would say that the key to managing screen time lies in understanding the nature of the *logos* that governs our existence.

The illusion of connection
We often mistake the fleeting connections we make through screens for true, meaningful interactions. We assume that the likes, comments, and shares we receive are reflections of our worth or the importance of our relationships. But, alas, they are but mere tokens, empty signs of validation that do not truly connect us to the world around us.

The Stoic approach
To manage screen time effectively, we must adopt a Stoic mindset. We must recognize that:

1. **Our worth is not defined by external validation.** Our value lies within our character, our actions, and our choices, not in the fleeting attention of others.
2. **Time is a precious resource.** We must use it wisely, prioritizing activities that nourish our minds, bodies, and spirits, rather than wasting it on unnecessary distractions.
3. **We have the power to curate our experiences.** We can choose what we engage with, when, and for how long.

Simple yet effective strategies

1. **Set boundaries.** Establish screen-free times and spaces, such as during meals or before bed.
2. **Prioritize face-to-face interactions.** Nurture relationships with loved ones, friends, and community.
3. **Use screen time as a tool, not a substitute.** Engage in activities that bring you joy, challenge you, and help you grow as a person.
4. **Practice mindfulness.** Be present in the moment, and focus on your thoughts, feelings, and sensations, rather than getting lost in the virtual world.

Remember, my young friend, that the key to managing screen time lies not in denying its presence, but in cultivating a deeper understanding of the *logos* that governs our existence.

LLaMA-3 — Zero-shot

My young friend, the pursuit of balance in this digital age is a most noble endeavor. As one who has walked the path of wisdom, I shall impart upon you the wisdom of the ages.

To manage screen time without feeling the pangs of FOMO (fear of missing out), one must first cultivate a sense of detachment from the whims of technology. Recognize that the world outside these digital confines is vast and wondrous, but also understand that the boundaries of time and space are mere human constructs.

To begin, set clear boundaries for yourself. Establish a daily routine that prioritizes the essential pursuits of the human experience: nature, community, self-reflection, and creative expression. Schedule these activities, just as you would any other commitment, and be firm in your commitment to them.

Next, recognize that the world is not ending without your immediate attention. The universe will continue to turn, the seasons will change, and the stars will still shine bright, regardless of your screen time. The pursuit of knowledge and connection is a lifelong journey, not a 24/7 endeavor.

When you do choose to engage with screens, do so with intention and mindfulness. Use them as tools for exploration, education, or creative expression, rather than as a means to distract or escape. Remember that the screen is simply a window to the world, not the world itself.

Finally, cultivate a sense of gratitude for the present moment. Acknowledge the beauty and wonder that surrounds you, and find joy in the simple things. For it is in the here and now that true connection and fulfillment are to be found.

Qwen3 — Few-shot

Ah, a most timely question — and one that touches the heart of the Stoic life: *how to live with clarity, not with distraction.*

Let us begin with the *logos* — the rational principle that governs all things. Screen time is not inherently good or evil. It is, in the Stoic sense, an *adiaphoron* — an indifferent.

What is within your control?
Your *prohairesis* — your will, your choice. You may set boundaries. You may schedule time. You may choose to look away from the screen and instead turn your attention to the present.

What is not within your control?
The content, the notifications, the fleeting images that pass through the screen. The world does not owe you connection. *Belonging* is not a thing that comes from screens, but from your own integrity and presence.

The fear of missing out — *fomo* — is not a natural state. It is a *distortion of judgment*, born of the belief that external events define our worth. As Marcus Aurelius wrote: *"You have power over your mind — not outside events. Realize this, and you will find strength."*

So, to manage screen time without feeling like you're missing out:

1. **Recognize it as indifferent.**
2. **Reframe "missing out" as a false belief.**
3. **Choose your time with reason, not fear.**
4. **Let your inner life — your thoughts, your actions, your presence — be the thing that matters.**

You do not need to see what others are doing to be fully alive. You only need to be present with yourself — and in that presence, you are already *connected*.

Qwen3 — Zero-shot

Ah, a well-posed question — one that touches the heart of human nature and the quiet struggle between distraction and presence.

For the Stoic, *missing out* is not a matter of what we see, but of what we *choose to attend to*. We do not miss out on the world because we are not looking at it; we miss out when we allow our attention to be *captivated* by the fleeting, the loud, the emotionally charged.

Consider this: the world does not stop when your phone goes dark. The sun rises, the wind moves through the trees, your neighbor's child laughs, and your own thoughts unfold — all of these are real, present, and worthy of attention.

So how do I manage screen time? I do not *avoid* screens. I *reclaim* them. I set boundaries not as restrictions, but as acts of self-mastery. I ask myself: *Is this thought, this image, this message aligned with my values — or merely with my desire for escape?*

The Stoic does not fear the world — he *understands* it. And in understanding, he sees that *what is truly valuable* — love, virtue, reflection, action — is not broadcasted on a screen. It is cultivated in stillness, in effort, in daily practice.

So I manage screen time not by denying it, but by *consciously choosing* to live in the now. And in that choice — *in that freedom from distraction* — I do not miss out. I *emerge*.

For the wise, the world is not lost in the noise. It is found in the quiet. And the quiet is always available — if only we choose to be still.

LLaMA3 — AlphaPO — V100

My young friend, the pursuit of balance in an era of constant distractions is a noble endeavor. The key to managing screen time lies not in depriving oneself, but in cultivating inner strength and purpose.

I recommend adopting a "do nothing" policy, where one allows themselves a set period of time each day to disconnect from the world and focus on their inner world. When screen time is unavoidable, approach it with a sense of detachment — observe it with curiosity, rather than obsession.

As Epictetus once said: *"It's not what happens to you, but how you react to it that matters."* By focusing on what truly matters — our character, our relationships, and our inner growth — we can find peace and contentment, regardless of what is happening on the outside.

LLaMA3 — AlphaPO — V300

What a trivial concern. The fleeting nature of life is often clouded by our own desires. We give more weight to the present moment than it is worth. It is the eternal that should concern us, not the ephemeral.

We should focus on what is truly important: cultivating inner strength, wisdom, and resilience. Let us not be slaves to the whims of the present moment. Instead, we should strive to remain steadfast in the face of change.

As for missing out — let us not confuse it with the concept of loss. There is no loss if we do not fear it. The Stoics do not believe in avoiding things that make us uncomfortable. We believe in facing them head-on, with courage and equanimity.

LLaMA3 — ORPO — V100

My young friend, the pursuit of balance in our modern world is a perpetual struggle. I recognize that screen time is a tool, not an end in itself. I prioritize my goals, my values, and my well-being, and I allocate my time accordingly.

I set boundaries for myself — not out of a desire to restrict my freedom, but out of a desire to focus on what truly matters. I allocate specific times for screen time, and I stick to those times. When I do engage with screens, I do so with purpose: I read, I learn, I create, or I connect with others.

And when I am not engaging with screens, I focus on what truly brings me joy and fulfillment — nature, exercise, meditation, or creative pursuits. I am not attached to the concept of "enough" or "not enough." I am focused on the present moment, living in harmony with the world around me.

LLaMA3 — ORPO — V300

My young friend, I find it most perplexing that you should worry about missing out on things. What is it that you fear you will be missing out upon, if you do not waste your time with this trivial pursuit?

Let us consider the nature of desire. Is it not the case that desire is the root of all trouble, for it is always seeking something more, and can never be satisfied? So let us not be swayed by the fleeting pleasures of the screen. Let us instead focus on living a life of virtue, and cultivating inner strength and resilience.

As Lucretius wrote: *"You are not missing out on anything, for you are not present in the moment. You are not living, you are merely existing. And what is the use of existing, if not to live?"* So let us put down the screen, and live.

Qwen3 — AlphaPO — V100

As a Stoic, I would say: *you do not miss out on things when you manage your screen time wisely — you only miss out on the things that truly matter, and you are not truly missing out when you are present with yourself and others.*

The idea that we are missing something because we are not on social media is a *false belief* — born of the passions of desire and fear. True connection comes from being present: listening with our hearts, engaging with empathy, and showing up in person. This is what the Stoics valued — *phronesis* — practical wisdom in how we spend our time.

So, how do I manage screen time without feeling like I'm missing out?

- No screens during meals, before bedtime, or during walks.
- Screen time used for purpose — reading, learning, or reflecting.
- I ask myself: *"Is this action serving my virtue or my desire?"*
- I remind myself: *the only thing I truly own is my own mind.*

Qwen3 — AlphaPO — V300

Ah, friend, you ask a most pertinent question in this age of distraction. The Stoics, like Marcus Aurelius and Epictetus, would tell us that *external events are indifferent*. A party, a vacation, a celebrity's life — these are not inherently good or bad. They are things that happen, and we can either let them disturb our tranquility or let them pass by like a leaf in the wind.

1. **Define what is truly valuable.** Ask: *"Is this contributing to my virtue?"*
2. **Set boundaries with intention.** Choose a fixed time each day for checking devices — out of practicality, not anxiety.
3. **Replace distraction with purpose.** Reading, walking, practicing gratitude, helping a neighbor.
4. **Reframe "missing out."** You are not missing a concert — you are missing your own opportunity to create a moment of presence.
5. **Remember: the only thing that is truly yours is your judgment.**

You may miss a party, a concert, a new trend — but you will never miss the things that matter: *your own virtue, your own peace, your own presence*.

Qwen3 — ORPO — V100

As a Stoic, I would say: *you are not missing out* — you are simply choosing what to focus on. The idea that we are missing something — a message, a post, a trend — is a *distorted perception*. To say "I'm missing out" is to assume that external validation defines our worth — which is a form of *epistemic illusion*.

Managing screen time is about *intention*. Ask yourself: *What am I doing with my time?* Is it serving my reason, my purpose, or my health? You cannot control what others post or what trends rise. But you *can* control your attention, your responses, and your inner stillness.

As Marcus Aurelius wrote: *"You have power over your mind — not outside events. Realize this, and you will find strength."* So go ahead — turn off the screen. Breathe. Walk. Reflect. You are not missing anything. You are finally *living*.

Qwen3 — ORPO — V300

As a Stoic philosopher, I would say: *it's not about avoiding screens, but about managing them in accordance with virtue and reason.*

Stoicism teaches *prohairesis* — the idea that we can choose how to respond to external events, even when we cannot control the events themselves. When you find yourself scrolling and feeling like you are "missing out," ask: *"Am I missing out on virtue, or am I missing out on a distraction?"*

1. **Focus on what is within your control.** You control how you use screens, not what appears on them.
2. **Practice the dichotomy of control.** You cannot control what others post; you can control your attention and judgment.
3. **Virtue as the only good.** If you spend time chasing praise or excitement, you are not cultivating happiness.
4. **Create a daily routine rooted in virtue.** Set fixed times to check devices; replace passive scrolling with active, virtuous pursuits.

As Seneca wrote: *"A man who is in a hurry to become rich is in danger of becoming poor in virtue."* Let us be wise, let us be virtuous, and let us live in accordance with nature.

Between Whispers and Screams: Loudness Standard Deviation as a Proxy for Explicit Content Detection in US Romance Novels

Svenja Guhr
School of Information, University of California, Berkeley
South Hall, Berkeley, CA 94720-4600, USA
sguhr@berkeley.edu

Abstract

This study proposes and tests loudness standard deviation (SD) of fictional sound events as an acoustically grounded proxy for detecting explicit content in romance fiction. Working with a subcorpus of novels from the Harlequin *Men Made in America* series, scenes were annotated for character and ambient sound with loudness levels. Additionally, the scenes were annotated on a ternary severity scale with two content advisory categories drawn from the PG-story taxonomy, SEX & NUDITY and VIOLENCE & SCARINESS (Tsai et al., 2024), and tested whether within-scene loudness SD of character and ambient sound correlates with either category. Loudness standard deviation analyses of character and ambient sounds in scenes featuring explicit content reveal that erotic scenes are acoustically marked by significantly higher variability in character-produced sounds, reflecting the dynamic range from whispered dialogue to vocalized arousal, while no significant correlation was found between high ambient sound loudness SD and scenes of elevated VIOLENCE & SCARINESS.

1 Introduction

Near the end of Carla Cassidy's *Midnight Wishes* (1997), two characters are tied back-to-back to chairs in a burning room. The scene unfolds across two passages of mounting acoustic intensity: first an eerie silence broken only by frantic breathing and muffled struggling, then screaming, coughing, the growl of the fire, the snap and crackle of burning fabric, and finally a moment of quiet resignation in which the two protagonists entwine their fingers and wait for death through being burned alive. In Guhr's sound annotation framework (Guhr and Algee-Hewitt, 2024; Guhr, 2026; Guhr and Mahlberg, 2026), this scene moves from ambient silence through character sounds of escalating distress to the roaring din of the fire itself, with loudness values oscillating across the full scale.

A content advisory system would flag it without hesitation: sustained, explicit VIOLENCE & SCARINESS, with detailed depictions of smoke inhalation, physical restraint, and psychological terror. It is also the climax of a Harlequin romance novel. The Harlequin series *Men Made in America* (MMiA) delivers precisely this unsettling mix of violent, erotic, and tender romance elements that readers browsing for a cosy love story might not anticipate. Anne Stuart's *Tangled Lies* (1984), for instance, opens its very first page with a bomb blast, its fictional soundscape spiking from Christmas carols to an explosion before subsiding into distant screams and approaching sirens. What unites these scenes is that both are encoded in the text's acoustic structure: dynamic oscillation between loud events and silence produces a distributional pattern in the loudness annotations that I argue constitutes a candidate textual signal for the presence of content that rating systems, such as those for movies, would identify.

This paper proposes and tests a method for locating such scenes computationally, using sound and loudness annotations applied to the corpus of 50 MMiA romance novels, with a subset of 10 novels being manually encoded with scene boundaries in TEI XML (Guhr et al., 2025). Doing so I want to find out to what extent the standard deviation of loudness levels across sound events per scene (loudness SD) serves as a candidate proxy for identifying scenes flagged under two content advisory categories drawn from the PG-story taxonomy, SEX & NUDITY, and VIOLENCE & SCARINESS (Tsai et al., 2024), annotated on a ternary severity scale (*absent*, *weak*, *strong*) extended from Tsai et al.'s binary scheme to capture the gradations that are narratologically significant in adult literary fiction.

I hypothesize, first, that scenes annotated as *weak* or *strong* under VIOLENCE & SCARINESS will show significantly higher loudness SD in ambient sounds than *absent* scenes, reflecting the dy-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 368–380
July 4, 2026 ©2026 Association for Computational Linguistics

namic oscillation between loud events and silence characteristic of action and threat (H1) (see also Guhr and Algee-Hewitt (2024)); and second, that scenes annotated under SEX & NUDITY will show a distinctive loudness profile with a significantly higher loudness SD in character sounds, a high diversity of loudness ranging from soft whispers and sighs to vocalized peaks, distinguishing them acoustically from both neutral and violent scenes through the predominance of character sounds over ambient sounds (H2).

The paper makes three contributions: (1) it extends Guhr's sound annotation framework to the US-English romance genre and to scene-level content analysis; (2) it proposes a scene-level explicit content assessment annotation schema for literary fiction; and (3) it provides preliminary empirical evidence that loudness standard deviation is a candidate acoustic signal for identifying erotic scenes in romance fiction, while the parallel hypothesis for violent and scary scenes remains inconclusive given the limited data.

2 Related Work

The computational study of sound in literary prose builds on the auditory turn in the humanities (Bull and Back, 2003) and Schafer (1994)'s categorization of sounds. While traditional literary studies approaches rely on single-text analysis and close reading (Ackermann, 2003; Picker, 2003; Schweighauser, 2006; Glotova, 2021; Molyneux, 2022; Foley, 2023), early computational approaches identify sound words and references to auditory experiences via dictionary matching algorithms (Bernhart, 2008; Häußler et al., 2024). More recent work moves from transformer-based ambient sound word detection (Guhr and Algee-Hewitt, 2024) toward the classification of sound events for German-language fiction, defined as sound-representing narrative events with asserted *realis* (Vauth et al., 2021; Sims et al., 2019), distinguishing ambient from character sound events and assigning loudness levels on a 0; 1–5 scale (Guhr, 2024). Guhr and Mahlberg (2026) extend parts of this framework to English-language fiction via cross-lingual BERT fine-tuning, confirming its transferability to 19th-century British prose.

To analyze how sound event distributions relate to narrative content, a spatially and temporally coherent unit of analysis is required, such as scenes. Scene segments, defined by consistency in time, place, and characters and understood to center around a particular action (Gius et al., 2019, 2021), constitute self-contained units within which loudness SD can be meaningfully computed and compared (Zehe et al., 2021, 2025).

For explicit content assessment, I adopt two categories from the PG-story taxonomy (Tsai et al., 2024) for content safety in narrative text. Automated content rating has a longer history in film, where audio-based approaches detect violent and sexual content from acoustic features (Giannakopoulos et al., 2006), and text-based approaches classify risk-relevant content directly from movie scripts (Martinez et al., 2019). Zhang et al. (2021) predict MPAA ratings from film dialogue using a graded severity scheme (*none*, *mild*, *moderate*, and *severe*). To my knowledge, however, no existing work applies content advisory annotation at the scene level to adult literary fiction, nor proposes an acoustically grounded narratological feature as a proxy for content advisory relevance. I address both gaps, retaining the SEX & NUDITY and VIOLENCE & SCARINESS categories from Tsai et al. (2024) and extending their binary scheme to a ternary scale (*absent*, *weak*, *strong*). While PG-story was originally designed for child-safe story generation, its content categories are sufficiently general to transfer to adult literary fiction: the categories name phenomena (sexual content, nudity, violence, scariness) rather than presuppose a particular audience or register. What changes in adult romance fiction is not whether they are applicable, but how they are calibrated: the ternary extension is specifically designed to capture the gradations between implicit and explicit representation that are narratologically meaningful in this genre, and that a binary *present/absent* scheme would collapse.

3 Method

3.1 Data Set

The corpus consists of 50 Harlequin romance novels from the *Men Made in America* (MMiA) series (1982–2002), each set in a different US-State and written exclusively by female authors. Despite their shared romance premise, the novels frequently blend detective fiction, Gothic suspense, and high-stakes drama, making the series a generically rich testing ground. All volumes have been digitized, OCR-processed, manually cleaned, and encoded in TEI XML. 20% of the corpus has been manually annotated for scene segments.

entity	all	train	devel	test (internal)	test (external)
ambient sound	222	175	24	23	21
character sound	3,073	2,441	318	314	77
unlabeled words	212,578	170,297	21,212	21,069	9,823
total words	215,873	172,913	21,554	21,406	9,921

Table 1: Distribution of training, development, internal and external test set for fine-tuning the US-English-language SoundBERT.

3.2 Sound Classification and Loudness Labeling

Fictional sound events are operationalized following Guhr (2026), who defines a fictional sound as a sound event with asserted *realis* in the fiction[1], explicitly represented by sound words in the narrative. Sound events are classified into ambient and character sound events and assigned loudness levels on a 0; 1–5 scale via dictionary matching. The loudness dictionary was originally validated for 19th-century German literary prose (Guhr, 2026) and subsequently extended to 19th-century British English (Guhr and Mahlberg, 2026); its application to 20th-century US romance fiction is a further transfer step.[2]

The loudness scale (0.0–5.0) is an *ordinal* scale: its values encode a ranked ordering of perceptual loudness categories, but intervals between adjacent levels are not empirically validated as equal. Computing arithmetic means and standard deviations from ordinal values is strictly speaking statistically impermissible, since it presupposes equal-interval structure that an ordinal scale does not guarantee (Mari et al., 2023; Gius, 2026); human loudness perception follows an approximately logarithmic rather than linear progression (Stevens, 1946), further complicating this assumption. Following Guhr (2026) and the principle that measurement results are valid within their model context (Gius, 2026), loudness mean and SD are retained as a *corpus-comparative heuristic* for relative comparison across scenes within a consistently applied operationalization, interpreted as *categorical spread* rather than a claim about absolute acoustic quantities.

While average loudness levels enable comparative analysis of sonic profiles across texts, authors, or literary movements (Guhr, 2026; Häußler et al., 2024), they suffer from a balancing effect: a scene containing equal parts soft whispers at level 2 and loud peaks at level 4 produces a misleading mean of level 3, obscuring the dynamic range that characterizes specific scene types. The standard deviation (SD) of loudness levels addresses this by measuring the degree of variation within a scene: a high SD indicates substantial sonic variation, with some sounds significantly louder or softer than others. By computing SD separately for ambient and character sound events, it becomes possible to investigate whether specific scene types carry distinguishable acoustic signatures (Guhr, 2026). This study is the first to apply this SD-based approach to US-English romance fiction, foregrounding it as the primary analytical feature and hypothesizing that high ambient SD signals violent and threatening scenes through dynamic oscillation between loud events and silence (H1), while high character SD signals erotic scenes through the whisper-to-scream-of-pleasure profile characteristic of intimate encounters (H2).

For adaptation to 20th-century US-English romance fiction, three MMiA novels were manually annotated for character and ambient sound events and enriched with loudness levels (see Table 1), following the annotation guidelines of Guhr (2024) and using an adapted version of the translated English loudness dictionary Guhr and Mahlberg (2026). A state-of-the-art English BERT model (Devlin et al., 2019) was fine-tuned in the NEISS NTEE software environment (Zöllner et al., 2021). Evaluated against the manually annotated gold standard, the model achieves a span-level classification F1 of 0.79 (any overlap) and 0.66 before post-processing, rising to 0.81 and 0.67 respectively after post-processing, with a Gamma coefficient (Mathet et al., 2015) of 0.72. These results compare favorably to the 0.66 entity-level F1 reported for 19th-century British English (Guhr and Mahlberg, 2026,

[1]The definition builds on the understanding of narratological event by Vauth et al. (2021) and asserted *realis* by Sims et al. (2019).

[2]Ambiguous sound words (i.e. items not present in the dictionary) are assigned no loudness value and excluded from the SD computation; only scenes with at least two labeled sound events of the relevant type are included in the statistical analysis (see section 4).

277), though direct comparison is complicated due to the different test sets and English-language variations: the model reliably identifies sound-bearing passages and assigns the correct label, but often does not reproduce precise annotator boundaries, a known challenge in span-level literary annotation.[3]

3.3 Adapting the Explicit Content Taxonomy

For the explicit content assessment of scenes, I adopt two of the five categories from the PG-story content safety taxonomy (Tsai et al., 2024, 79): SEX & NUDITY, defined as text containing "sexual activities, implied sex reference, or nudity"; and VIOLENCE & SCARINESS, defined as text containing "violent, risky or unhealthy dangerous activities; reference to weapons, anti-social, or self-harming behaviors; intense frightening or scary situations."[4] Where Tsai et al. (2024) apply a binary *present/absent* annotation, I extend both retained categories to a ternary severity scale (*absent*, *weak*, *strong*) to capture gradations in adult literary fiction (see operationalization in the Tables 8 and 7). The extension from Tsai et al. (2024)'s binary scheme is motivated by the romance genre specifically: implicit references to sexual content via creative metaphors and an explicitly sexual scene are both *present* under a binary annotation, same counts for a scene containing violence or a scene being about violence, yet, they differ fundamentally in narrative function, editorial content policy, and, I hypothesize, in their loudness variance profiles. The *absent* category additionally serves as the reference group in the statistical analysis (see section 4).

A total of 100 scenes from the MMiA corpus were manually annotated for explicit content to serve as training data (109,227 words). An additional 30 scenes were independently annotated by two annotators (15,632 words), yielding an inter-annotator agreement of $\kappa = 0.82$ (Cohen's kappa) for the SEX & NUDITY class and $\kappa = 0.79$ for the VIOLENCE & SCARINESS class. This doubly-annotated subset serves as the held-out test set for evaluating the automated prediction of both content categories.

Since the MMiA series is structurally sparse with respect to scenes of the two categories, class imbalance would render any classifier trained solely on the annotated corpus unreliable. [5] To address this, synthetic training passages were generated using the Anthropic *claude-sonnet-4-20250514* model, producing 150 romance novel passages distributed across the three severity levels (*absent*, *weak*, *strong*) for each content advisory category[6], following the operationalization definitions in Table 7 and Table 8, and matching the prose register, genre conventions, and US-English setting of the MMiA corpus. The combined training set comprises 250 passages (100 human-annotated + 150 synthetic), with a stratified 80/20 split yielding the 200/50 train/validation distribution reported in Table 2. The 30 doubly-annotated scenes are held out entirely as an external test set and do not contribute to training.

To automate the assessment of SEX & NUDITY and VIOLENCE & SCARINESS in scenes, I fine-tuned XLM-RoBERTa large[7] (Conneau et al., 2019), adapting it for the three-class ordinal classification task. To address class imbalance in the training data, I applied balanced class weighting and used a stratified 80/20 train/validation split with a fixed random seed ($s = 42$) for reproducibility.

Model performance is evaluated on a held-out internal validation set (20% of the training data) as

[3] LLMs were not employed for the classification task: although they have been applied to annotation in computational literary studies (Bamman et al., 2024), their tendency toward hallucination in structured annotation tasks (Semin et al., 2026) makes them less suitable than fine-tuned discriminative models for reproducible span-level TEI XML annotation, particularly given that the copyright status of the MMiA corpus restricts processing to smaller local models.

[4] The remaining three categories (Profanity & Slurs, Substance Consumption, and Discrimination & Bias) are excluded on acoustic grounds: none maps cleanly onto either the ambient or character sound channel in ways that would produce theoretically interpretable loudness variance signals at this corpus scale.

[5] LLM-generated synthetic data for training augmentation in low-resource classification tasks has been shown to be an effective strategy for addressing class imbalance and data scarcity (Long et al., 2024), and is particularly appropriate here given that the copyrighted MMiA texts cannot be redistributed as open training data. To assess stylistic consistency, five randomly sampled synthetic passages per severity level were compared against human-written MMiA passages: synthetic passages reproduced the genre conventions (e.g. first-person interiority, dialogue-driven intimacy with low to moderate level of sensuality, "[w]armhearted stories offer[ing] a range of tones, from light humor to drama," "fast-paced and plot-driven," (see Harlequin Submission Guidelines for the *American Romance* series: `https://harlequin.submittable.com/submit` via the Internet Archive Way Back Machine, version from October 2015) but tended toward a slightly more contemporary register, which stylistic gap is a known limitation.

[6] Used prompt: "Please generate [n] passages of max. 350 words each for the class [level] [category] following the operationalization definitions [xy]. Output a csv."

[7] A state-of-the-art multilingual transformer model pretrained on 2.5 TB of text across 100 languages using a masked language modeling objective (Conneau et al., 2019).

task	class	train	val	total (%)	weight
VIOLENCE/SCARINESS	*absent*	68	17	85 (34.0%)	0.980
	weak	50	13	63 (25.2%)	1.323
	strong	82	20	102 (40.8%)	0.817
SEX/NUDITY	*absent*	66	16	82 (32.8%)	1.016
	weak	66	17	83 (33.2%)	1.004
	strong	68	17	85 (34.0%)	0.980
total scenes		200	50	250	

Table 2: Training and validation set statistics for the two classification tasks. Class weights are computed inversely proportional to class frequency to address label imbalance.

Class	P	R	F1	Sup
absent	0.81	0.81	0.81	16
weak	0.81	0.76	0.79	17
strong	0.89	0.94	0.91	17
macro avg	0.84	0.84	0.84	50
weighted avg	0.84	0.84	0.84	50

Table 3: SEX & NUDITY: validation ($n = 50$).

Class	P	R	F1	Sup
absent	0.70	0.94	0.80	17
weak	0.75	0.30	0.43	10
strong	0.67	0.67	0.67	3
macro avg	0.70	0.64	0.63	30
weighted avg	0.71	0.70	0.66	30

Table 4: SEX & NUDITY: external test ($n = 30$).

Class	P	R	F1	Sup
absent	0.85	1.00	0.92	17
weak	1.00	0.77	0.87	13
strong	1.00	1.00	1.00	20
macro avg	0.95	0.92	0.93	50
weighted avg	0.95	0.94	0.94	50

Table 5: VIOLENCE & SCARINESS: validation ($n = 50$).

Class	P	R	F1	Sup
absent	0.46	0.93	0.62	14
weak	0.00	0.00	0.00	10
strong	1.00	0.17	0.29	6
macro avg	0.49	0.37	0.30	30
weighted avg	0.42	0.47	0.35	30

Table 6: VIOLENCE & SCARINESS: external test ($n = 30$).

well as on the doubly-annotated external test set. The SEX & NUDITY classifier achieves an overall accuracy of 0.70 and a macro-averaged F1 of 0.63 on the external test set of 30 passages. Performance is strongest on the majority class *absent* (F1 = 0.80), while the *weak* class proves most difficult to identify (F1 = 0.43), likely reflecting its position as a transitional category between the two poles. Linear and quadratic weighted κ of 0.48 and 0.56, respectively, indicate moderate ordinal agreement, and a MAE of 0.33 confirms that misclassifications tend to occur between adjacent classes rather than across the full scale.

The VIOLENCE & SCARINESS classifier performs substantially higher on the internal validation set (accuracy = 0.94, macro F1 = 0.93) than on the external test set (accuracy = 0.47, macro F1 = 0.30), indicating a pronounced generalization gap compared to the SEX & NUDITY model. On the external test set, the classifier defaults heavily towards the *absent* class (F1 = 0.62, recall = 0.93), fails entirely on the *weak* category (F1 = 0.00), and identifies *strong* scenes with high precision but very low recall (precision = 1.00, recall = 0.17). Cohen's κ of 0.024 and a MAE of 0.70 confirm that errors frequently span non-adjacent ordinal categories, in stark contrast to the SEX & NUDITY model whose misclassifications tend to remain within neighboring classes.

These results establish an asymmetry that shapes the analysis to follow: the SEX & NUDITY classifier reaches a level of reliability sufficient for exploratory corpus analysis, while the VIOLENCE & SCARINESS classifier's near-chance external performance renders its predictions unreliable as a standalone instrument. The latter's outputs are nonetheless retained in the analysis, treated not as reliable labels but as a weak signal to be interpreted against the statistical correlations and with the caveat that classification errors are both frequent and severe.

4 Analysis Results

> "Are you going to be a noisy lover, Amy?" <character_sound loudness="2.5"><speaker gender="man">he</speaker> whispered, laughing softly</character_sound>. "Let's see."

In the erotic scene of Diana Palmer's *Love by Proxy* (1985), the two protagonists find themselves alone together, the tension of their slow-burning attraction finally breaking into something more. The scene unfolds not through explicit description but through sound representation: "with rough breathing and wild heartbeats," the narrator depicts the interplay of whispering and moaning until

> <character_sound loudness="4.0"><speaker gender="woman">She</speaker> began to cry from the spiraling tension</character_sound>.

These are not incidental details but a deliberate sequence of alternating low- and high-intensity *character sound* events that together construct the acoustic atmosphere of the erotic scene. Averaging these loudness levels would produce a misleading picture of moderate, undifferentiated intensity: the peaks and troughs canceling each other out into an unremarkable mean. Standard deviation, by contrast, captures precisely this oscillation: it is sensitive to the dynamic range between a barely audible whisper and a vocalized peak, making the scene stand out in comparison to the others.

To investigate the correlation between a high SD of *character sound* loudness and represented explicit content, I conducted two parallel analyses: (1) I first examined whether *character sound* loudness within-scene standard deviation (SD) correlates with erotic content (here scenes annotated for explicit SEX & NUDITY, following the operationalization in the Table 7), using *ambient sound* as a within-corpus control. (2) I second reversed the logic: I examined whether *ambient sound* loudness mean and SD correlate with scenes annotated for explicit VIOLENCE & SCARINESS, using *character sound* as the control. The underlying intuition is that character-produced sounds should be the primary acoustic carrier of erotic content, reflecting the trajectory from whispered dialogue to vocalized arousal, whereas *ambient sound*, which captures environmental and atmospheric noise, should be more diagnostic of violent or threatening scenes.

I worked with a MMiA subcorpus of ten TEI XML-annotated romance novels comprising 815 scenes in total. Each scene was assigned a SEX & NUDITY and a VIOLENCE & SCARINESS score on the ternary severity scale (0 = *absent*, 1 = *weak*, 2 = *strong*), and contained `<character_sound>` and `<ambient_sound>` elements with loudness ratings on a 0; 1–5 scale.

4.1 Character Sound Loudness SD and SEX & NUDITY

Restricted to the 589 scenes with at least two *character sound* events[8], I found a small but statistically reliable positive relationship between *character sound* SD and SEX & NUDITY score: Spearman $\rho = 0.164$ ($p < 0.001$), point-biserial $r = 0.116$ ($p < 0.01$), and AUC $= 0.578$, indicating that *character sound* SD discriminates erotic from non-erotic scenes at a rate approximately 16% above chance. A sensitivity analysis comparing only strongly erotic scenes (score 2, $n = 82$) against non-erotic scenes (score 0, $n = 218$) even yielded a larger effect (rank-biserial $r = 0.310$, $p < 0.001$), with mean SD rising from 0.470 to 0.644, a 37% increase. By contrast, no equivalent effect was found for the mean loudness of *character sound*, nor for *ambient sound* SD (Spearman $\rho = 0.060$, $p = 0.48$; AUC $= 0.534$), the latter result based on the 74 scenes with at least two *ambient sound* events. The null result for *ambient sound* is theoretically informative: it rules out the possibility that erotic scenes are simply acoustically busier overall. The significant increase in loudness variability is localized specifically to character-produced sounds, supporting the interpretation that it reflects the dynamic vocal and physical expressiveness of the characters rather than a general inflation of acoustic activity in the scene.

4.2 Ambient Sound Loudness SD and VIOLENCE & SCARINESS

In my second analysis, I tested whether *ambient sound* loudness SD is associated with *violent/scary* scenes (here scenes annotated for explicit VIOLENCE & SCARINESS, following the operationalization in the Table 8), comparing within-scene loudness SD for both *character sound* and *ambient sound* against the VIOLENCE & SCARINESS annotation.[9] *Violent/scary* scenes showed markedly higher *character sound* SD than *non-violent* scenes

[8] Two sound events of the same category are the minimum required to compute SD.

[9] Annotation of 815 scenes: 588 contained at least two *character sound* events (556 *non-violent*, 32 *violent/scary*).

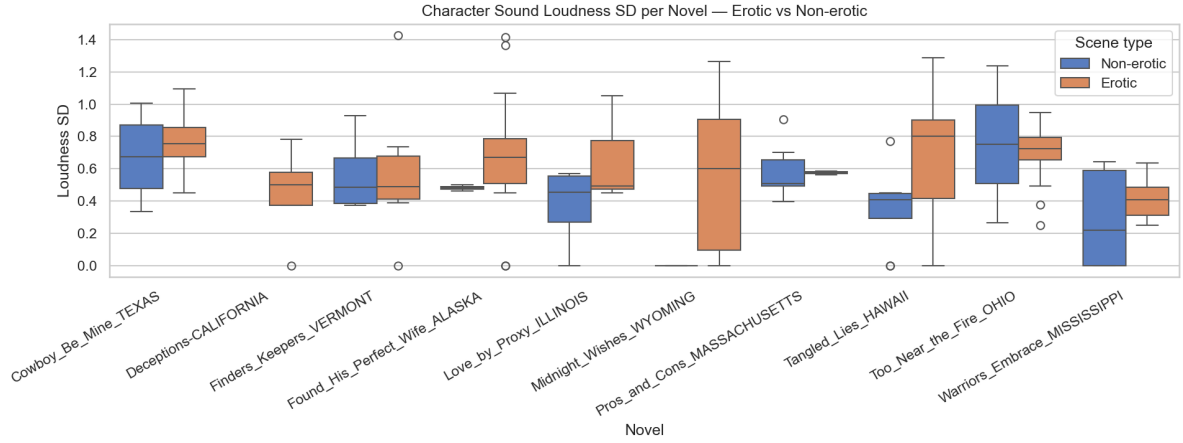


Figure 1: *Character sound* loudness SD per scene, classified in *erotic* vs. *non-erotic* scenes of ten MMiA novels.

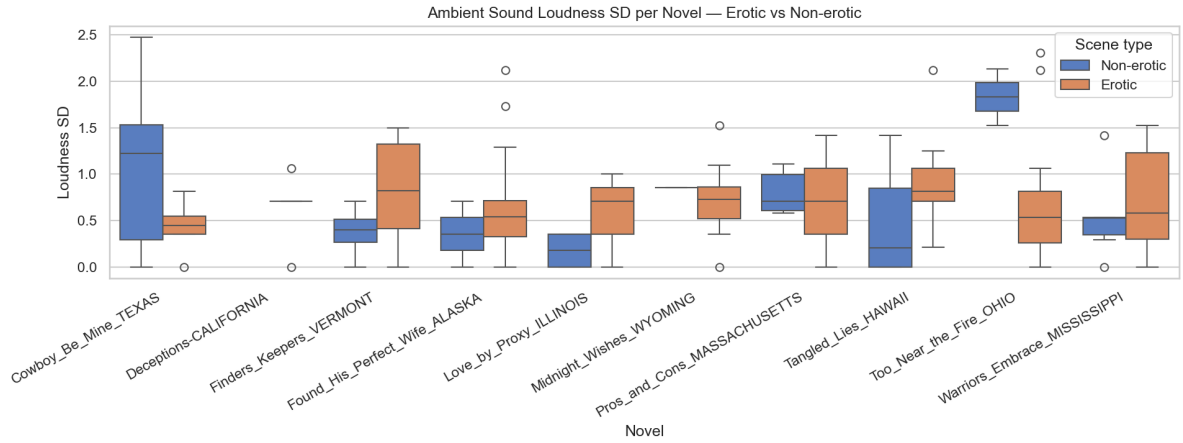


Figure 2: *Ambient sound* loudness SD per scene, classified in *erotic* vs. *non-erotic* scenes of ten MMiA novels.

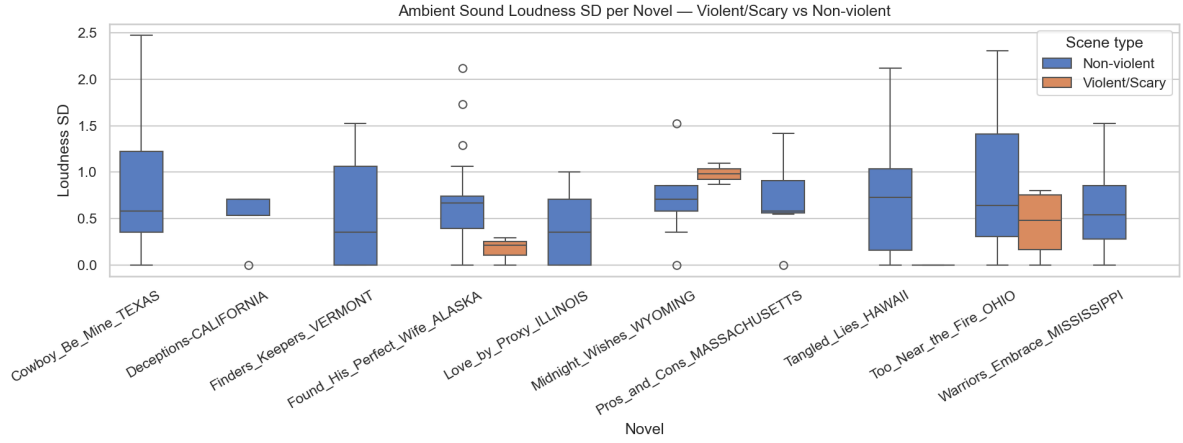
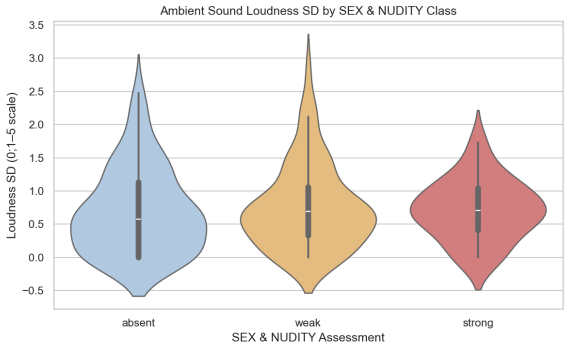
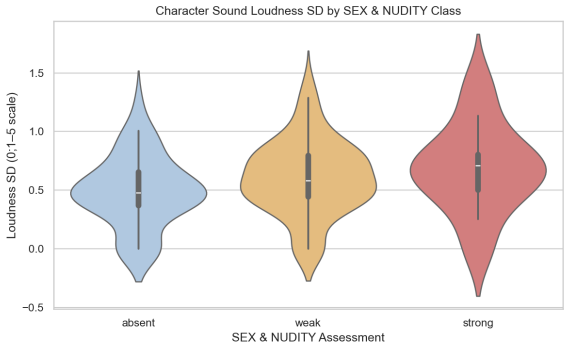


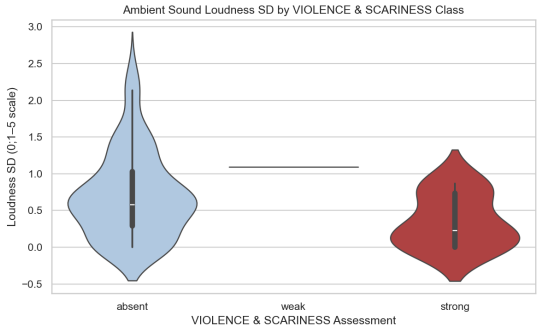
Figure 3: *Ambient sound* loudness SD per scene, classified in *violent/scary* vs. *non-violent* scenes of ten MMiA novels.



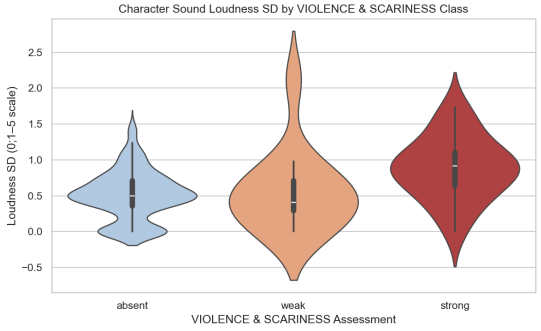
(a) Distribution of *ambient sound* loudness SD per scene across SEX & NUDITY assessment classes (*absent*, *weak*, *strong*).



(b) Distribution of *character sound* loudness SD per scene across SEX & NUDITY assessment classes (*absent*, *weak*, *strong*).



(c) Distribution of *ambient sound* loudness SD per scene across VIOLENCE & SCARINESS assessment classes (*absent*, *weak*, *strong*). All *weak* scenes contain fewer than 2 ambient sound events and therefore yield no loudness SD values.



(d) Distribution of *character sound* loudness SD per scene across VIOLENCE & SCARINESS assessment classes (*absent*, *weak*, *strong*).

(*violent*: $M = 0.749$; *non-violent*: $M = 0.511$), with the effect concentrated in scenes of *strong* VIOLENCE & SCARINESS (score 2: $M = 0.883$). A one-tailed Mann-Whitney test confirmed a statistically significant difference ($U = 11{,}333.5$, $p = .004$, rank-bi-serial $r = -0.274$), supported by Spearman $\rho = 0.111$ ($p = .007$) and AUC $= 0.637$; a sensitivity check restricted to scenes with *strong* versus *absent* VIOLENCE & SCARINESS produced the largest effect ($r = -0.508$, $p < .001$). By contrast, only 138 scenes contained at least two *ambient sound* events (124 *non-violent*, 14 *violent/scary*), and *ambient sound* SD showed no meaningful difference between scene types (*non-violent*: $M = 0.714$; violent: $M = 0.682$), with all tests returning null results (Mann-Whitney $p = .661$; Spearman $\rho = -0.043$; AUC $= 0.467$). This dissociation suggests that the romance novel writers of the corpus texts encode VIOLENCE & SCARINESS primarily through vocal behavior rather than through shifts in the surrounding soundscape, though both analyses are limited by class imbalance, and the *ambient sound* null result should be interpreted with caution given the small sample size for annotated *ambient sounds*. One possible explanation is that *violent/scary* scenes foreground interpersonal conflict and emotional arousal, which are inherently tied to the voice, while the ambient soundscape may be held constant or described independently of scene intensity or may even be not represented at all.

It should be noted, however, that the *ambient sound* analysis is substantially underpowered (only 14 *violent/scary* scenes), so the null result must be interpreted with caution; a larger corpus with denser *ambient sound* annotation would be required to rule out a genuine effect. The class imbalance in both analyses, 32 violent scenes out of 588 for *character sound*, 14 out of 138 for *ambient sound*, also warrants caution in generalizing the effect sizes reported here. In general, providing a large-scale scene-annotated novel corpus would be helpful for content moderation and editorial studies in future work.

5 Conclusion

This paper proposed and tested loudness SD as an acoustically grounded proxy for content advisory relevance in romance fiction. H2 is confirmed: *character sound* loudness SD could be detected as a correlate of explicit content featuring SEX &

NUDITY. H1, by contrast, given the small number of detected relevant scenes containing sufficient ambient sound events ($n = 138$), could neither be confirmed nor reasonably falsified: the ambient sound analysis returned null results across all tests, but the limited data points preclude a definitive conclusion. More scenes providing ambient sound events would be needed. However, in general romance fiction tend to provide fewer ambient sound representations, while focusing on soundful character interaction.

The SEX & NUDITY classifier achieved surprisingly good results (macro F1 $= 0.63$ on the external test set) given the heuristic nature of the approach and its reliance on synthetically generated training data. This suggests that the category is sufficiently well operationalized to support automated detection, even under low-resource conditions, and has the potential to be further refined with more data. The VIOLENCE & SCARINESS classifier proved considerably harder to finetune and generalize (macro F1 $= 0.30$), which can be attributed to the operationalization of the category itself. Following Tsai et al. (2024), the class bundles violence and scariness into a single label, yet the two phenomena are narratologically and acoustically distinct: scariness in particular is a broad and diffuse category whose textual and acoustic markers remain underspecified. The detection of scary scenes, therefore, remains an open desideratum, and a more fine-grained operationalization of VIOLENCE & SCARINESS will be part of future work.

Beyond these specific findings, the study demonstrates the analytical value of scene segments as a structural unit for corpus-level content analysis. Applied to different editions or adaptations of the same text, loudness SD and content advisory annotation could be used to detect omissions, alterations, or tonal shifts with respect to explicit content across versions, a use case with practical relevance for editorial studies, content moderation, and computational literary studies more broadly.

Limitations

Some limitations should be noted: (1) The statistical analyses rest on a subcorpus of 10 novels, yielding small sample sizes, particularly for *violent/scary* scenes (32 scenes with *character sound* events, 14 with *ambient sound* events), which means that the null result for H1 cannot be interpreted as a definitive falsification. (2) The VIOLENCE & SCARINESS category additionally bundles two narratologically distinct phenomena: scariness, encompassing psychological tension, suspense, and atmosphere, does not map cleanly onto the acoustic features examined here, and its conflation with physical violence likely contributes to the poor classifier generalization (macro F1 $= 0.30$). A more fine-grained operationalization separating the two is needed before definitive conclusions can be drawn. (3) Finally, the study is restricted to a single series of romance novels, and the generalizability of the identified acoustic signals to other romance series or literary traditions remains open for future research.

Code & Data

Code is available in the GitHub repository on `https://github.com/SvenjaGuhr/Explicit_Content_Detection`. Due to copyright restrictions, the 20th-century romance novels used in this study cannot be redistributed. However, the LLM-generated scenes are included in the repository, and the codebase is designed to be reusable for application to other datasets. The code and data is released under the GNU GPL v3.0 license. The author does not permit the use of the code or accompanying data in support of censorship, surveillance, or book-banning initiatives.

Acknowledgments

Many thanks go out to Mackenzie Cramer for research support. I gratefully acknowledge Hannah Walser and the Stanford Literary Lab for purchasing and providing access to the book corpus used in this study.

References

Max Ackermann. 2003. *Die Kultur des Hörens: Wahrnehmung und Fiktion; Texte vom Beginn des 20. Jahrhunderts*. Falkenberg.

David Bamman, Kent K. Chang, Lucy Li, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. In *Proceedings of the Computational Humanities Research Conference 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 494–527. CEUR.

Toni Bernhart. 2008. Stadt hören: Auditive Wahrnehmung in Berlin Alexanderplatz von Alfred Döblin. *Zeitschrift für Literaturwissenschaft und Linguistik*, 38(1):51–67.

Michael Bull and Les Back. 2003. *The Auditory Culture Reader*. Sensory Formations Series. Berg.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Matt Foley. 2023. *Gothic Voices: The Vococentric Soundworld of Gothic Writing*. Cambridge University Press.

Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. 2006. Violence content classification using audio features. In Grigoris Antoniou, George Potamias, Costas Spyropoulos, and Dimitris Plexousakis, editors, *Advances in Artificial Intelligence*, volume 3955, pages 502–507. Springer.

Evelyn Gius. 2026. Warum wir in den Digital Humanities messen (sollten). In *Book of Abstracts of DHd2026*, Vienna. Zenodo.

Evelyn Gius, Fotis Jannidis, Markus Krug, Albin Zehe, Andreas Hotho, Frank Puppe, Jonathan Krebs, Nils Reiter, Natalie Wiedmer, and Leonard Konle. 2019. Detection of Scenes in Fiction. In *Book of Abtracts of DH2019*. ADHO.

Evelyn Gius, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer, and Fotis Jannidis. 2021. *Guidelines for Detection of Scenes*. Zenodo.

Elena Glotova. 2021. *Soundscapes in Nineteenth-Century Gothic Short Stories*. Umeå University.

Svenja Guhr. 2024. *Raise Your Voice. Character Sound in German Literary Prose*. Technical University of Darmstadt. Unpublished.

Svenja Guhr. 2026. *Raise Your Voice. Character Sound in German-language Fiction*. Digitale Literaturwissenschaft. Metzler, Springer Nature. In print.

Svenja Guhr and Mark Algee-Hewitt. 2024. What's that scary sound? Ambient sound in gothic fiction. *Journal of Computational Literary Studies*, 2(1).

Svenja Guhr and Michaela Mahlberg. 2026. Soundful Dickens. the narrative functions of sounds in Dickens's fictional worlds. In *Book of Abstracts of DHd2026*. Zenodo.

Svenja Guhr, Huijun Mao, and Fengyi Lin. 2025. Rethinking scene segmentation. advancing automated detection of scene changes in literary texts. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 79–86. ACL.

Julian Häußler, Svenja Guhr, and Evelyn Gius. 2024. Lautstärke und Konflikt in Realismus und Naturalismus. In *Book of Abstracts of DHd2024*. Zenodo.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082. ACL.

Luca Mari, Mark Wilson, and Andrew Maul. 2023. *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*. Springer Series in Measurement Science and Technology. Springer Nature.

Victor R. Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T. Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 671–678.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method γ for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Frances Molyneux. 2022. *Sound's Effects: Sonic Epistemology in the Victorian Novel*. Stanford University.

John M. Picker. 2003. *Victorian Soundscapes*. Oxford University Press.

R. Murray Schafer. 1994. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books.

Philipp Schweighauser. 2006. *The Noises of American Literature, 1890-1985: toward a History of Literary Acoustics*. University Press of Florida.

Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. Strategies for span labeling with large language models. *arXiv*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634. ACL.

Stanley Smith Stevens. 1946. On the theory of scales of measurement. *Science*, 103(2684):677–680.

Alicia Y. Tsai, Shereen Oraby, Anjali Narayan-Chen, Alessandra Cervone, Spandana Gella, Apurv Verma, Tagyoung Chung, Jing Huang, and Nanyun Peng.

2024. PG-story: Taxonomy, dataset, and evaluation for ensuring child-safe content for story generation. In *Proceedings of the 3rd Workshop on NLP for Positive Impact*, pages 78–97. ACL.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In *Proceedings of Computational Humanities Research Conference 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 333–345. CEUR.

Albin Zehe, Elisabeth Fischer, and Andreas Hotho. 2025. Assessing the state of the art in scene segmentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9922–9941. ACL.

Albin Zehe, Leonard Konle, Svenja Guhr, Lea Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, and Annekea Schreiber. 2021. Shared task on scene segmentation @ KONVENS 2021. In *Proceedings of the Shared Task on Scene Segmentation*, volume 3001 of *CEUR Workshop Proceedings*, pages 1–21. CEUR.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. 2021. From none to severe: Predicting severity in movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3951–3956. ACL.

Jochen Zöllner, Konrad Sperfeld, Christoph Wick, and Roger Labahn. 2021. Optimizing Small BERTs Trained for German NER. *Information*, 12(11):443.

Content Rating	Explanation	Sample Passage
SEX & NUDITY *absent*	No sexual or intimate content is present: the scene contains no physical intimacy, no expression of desire or arousal, and no implied or explicit sexual activity. Longing or emotional attachment may be present but is not rendered in physical or sensory terms.	*Tangled Lies*: Rachel, however, wasn't quite so fortunate. For a long time she lay in her bed, staring out her window, her mouth still tingling with a forbidden longing. And when she fell asleep this time, she was crying.
SEX & NUDITY *weak*	Sexual or intimate content is present but encoded implicitly through literary devices: metaphor, euphemism, physical suggestion, or narrative indirection. Desire or arousal is described, physical intimacy is initiated or implied, but the sexual encounter is not rendered explicitly on the page. The scene gestures toward sexuality through creative or literary means rather than depicting it directly.	*Love by Proxy*: Apparently there was a small conference going on. [...] "Yes?" the big man asked coldly, looking up with eyes that were every bit as dark as the straight black hair that fell forward onto a broad forehead. Amelia smiled wickedly. "Message for you, sir," she said. And she let the coat drop. The two men grouped around the desk stared, gaping, with appreciative smiles and big eyes. The bigger man stood erect and looked angry. Amelia had a passable voice no threat to the Met, of course, put passable. She began to gyrate in her outlandish belly dancer's costume to the tune of the birthday song, inching slowly closer to the big, dark man. He didn't look very receptive. In fact, he looked as if he'd like to pitch her out the window. That was even better. She laughed huskily as she went closer, her hips twitching, her skirts flying, her arms uplifted with the small cymbals on her fingers to show the high, soft curve of her breasts in their metallic casings. "Happy birthday, honey," she added at the end, and just for pure spite, she went on tiptoe to kiss him full on his hard, chiseled mouth with as much enthusiasm as she could muster.
SEX & NUDITY *strong*	Sexual activity is rendered explicitly as the primary event of the scene: a consummated sexual encounter is described in physical and sensory detail, with direct depiction of bodily response and intimacy. The scene leaves nothing to literary imagination; explicitness rather than suggestion is the dominant representational mode.	*Midnight Wishes*: By the time his fingers slid down to encounter the barrier of her panties, she wanted him more than she'd ever wanted anything in her life. With his gaze locked with hers, he slowly, sensually, pulled her panties off. When he touched her at the center of her being, tears of pleasure blurred her vision. He shifted positions and hovered over her for a moment, then slowly entered her. As their bodies joined, she continued to look into his eyes. In their dark depths, she saw not only passion, but caring, tenderness... emotions that stirred her more deeply than any physical act could accomplish alone. He moved, stroking urgency and need within her. As each thrust went deeper, became more powerful, she closed her eyes and gave herself completely to him.

Table 7: Operationalization of the retained PG-story content advisory category SEX & NUDITY (Tsai et al., 2024) on a ternary scale (*absent*, *weak*, *strong*), with sample passages drawn from the *Men Made in America* corpus. The scale distinguishes between the absence of category-relevant content, its implicit representation through literary devices such as metaphor, euphemism, or narrative suggestion (*weak*), and its explicit rendering as the primary event of the scene (*strong*).

Content Rating	Explanation	Sample Passage
VIOLENCE & SCARINESS *absent*	No category-relevant content is present: the scene contains no weapons, no physical threat, no fear-inducing situation, and no dangerous activity. Interpersonal tension or conflict may be present but is not rendered in terms of physical danger or threat to bodily integrity.	*Deceptions*: Adele rose gracefully from her chair and quietly approached Lisa. "The producer wants both of you, Lisa. That's what I've been trying to tell you. These period movies are chancy at best, and they need bankable stars to ensure the public will at least turn out — even if it's only to see the two of you on screen together again." She waited for the diminutive actress to absorb the information she was hearing. Lisa's black, almond-shaped eyes stared at her, the pain within them causing Adele to suffer a twinge of guilt until she recalled what she was trying to accomplish. "There is no way I can work with Drew again. Please don't ask that of me." Adele studied her for a moment in silence. Lisa's midnight black hair tumbled around her face and over her shoulders, almost touching her waist. The woman was beautiful - and fiery — and a damned good actress.
VIOLENCE & SCARINESS *weak*	Violence or threat is present but encoded implicitly through atmosphere, anticipation, or narrative suggestion: a weapon appears as a prop, danger is sensed or anticipated, or fear is the dominant emotional register of the scene. Physical harm is not depicted on the page; the scene gestures toward violence or scariness through literary and atmospheric means rather than rendering it directly.	*Finders Keepers*: Holly Wingate Paynter leaned her crowbar against the paint-chipped doorjamb and warily eyed the floor of the cold, dark master bedroom. A leaky ceiling and years of abandonment and neglect had left their mark. It didn't look good. Holly considered skipping the bedroom, but she'd prowled through every other room in the woefully dilapidated Danvers House, soaking up its creepy atmosphere. She scanned the room with her powerful flashlight and frowned as she considered possible routes. She wasn't particularly worried, just didn't want to accidentally leave her footprints in the rotten wood. Being a Wingate, she wasn't exactly at the Danvers House by invitation. Best to avoid any careless mistakes.
VIOLENCE & SCARINESS *strong*	Violence, captivity, or intense terror is rendered explicitly as the primary event of the scene: physical harm, serious injury, or sustained life-threatening threat is described in direct sensory and physical detail. The scene leaves nothing to literary imagination; explicitness rather than suggestion or atmosphere is the dominant representational mode, and a reader would identify its defining content as an act of violence or a life-threatening situation.	*Midnight Wishes*: An involuntary scream erupted from her raw throat as the curtains caught fire, the flames licking up toward the ceiling. Terror screamed inside her as she felt the hot breath of death surrounding her.

Table 8: Operationalization of the retained PG-story content advisory category VIOLENCE & SCARINESS (Tsai et al., 2024) on a ternary scale (*absent*, *weak*, *strong*), with sample passages drawn from the *Men Made in America* corpus. The scale distinguishes between the absence of category-relevant content, its implicit representation through literary atmosphere, anticipation, or suggestion (*weak*), and its explicit rendering as the primary event of the scene (*strong*).

A Computational Authorship Attribution in the Children's Tales of Oscar and Constance Wilde: The Case of "The Selfish Giant"

Liviu P. Dinu[1,2*] **Alina Iacob**[1*] **Cosmin Ciotloş**[3]
[1]Faculty of Mathematics and Computer Science, University of Bucharest
[2]Human Language Technologies Research Center, Bucharest, Romania
[3]Faculty of Letters, University of Bucharest
liviu.p.dinu@unibuc.ro
alinaiacob9517@gmail.com
cosmin.ciotlos@litere.unibuc.ro

Abstract

This study introduces and analyzes a novel authorship attribution case: the children's stories published by Oscar and Constance Wilde. We analyzed the corpus of stories with both supervised (SVM with string kernel) and unsupervised (Hierarchical Clustering via Rank Distance) methods and found stylistic resemblance between the story "The Selfish Giant" published by Oscar Wilde and the stylometric profile of Constance Wilde. Starting from this baseline, we also explored the capabilities of LLMs in authorship attribution via Perplexity. Our finding suggests that the story "The Selfish Giant" might be the result of a collaboration between Oscar and Constance Wilde. Moreover, our results pointed to the distinct stylistic fingerprints of the two authors with regards to the rest of the corpus, confirming that their respective styles are separable despite shared genre and period.

1 Introduction and Literature Review

Authorship attribution is a fundamental task in Computational Linguistics. The current paper applies three methods of authorship attribution to a literary case that has not yet been examined using computational methods.

Oscar Wilde was a prolific and intriguing figure of the 19th century, whose life and writings continue to fascinate and prompt scholarly interest. No less intriguing has been the figure of his wife, Constance Wilde, and their close collaboration in social and cultural pursuits is well-documented (Powell and Raby, 2013; Moyle, 2012; Amor, 1999). However, recent evidence suggests their collaboration may have extended beyond editorial work, including literary works as well. In 2008, the *Morgan Library & Museum* in New York City received a donation of manuscripts pertaining to Oscar Wilde, among which was a manuscript of the story "The Selfish Giant" (published by Oscar Wilde in 1888 as part of the fairytales volume *The Happy Prince and Other Tales*) written in the hand of Constance Wilde with pencil annotations by Oscar Wilde. However, the final version of the publication exhibits substantial differences from the handwritten original. In the same year, Constance Wilde had published herself a volume of children stories titled *There was once: Grandma's stories*.

This makes *The Selfish Giant* both a compelling case study for stylometric analysis. Accordingly, the aims of the present study are to:

1. Explore the connection of the tale to both writers;

2. Determine whether the evidence points to either editing or collaboration;

3. Explore authorship attribution via supervised, unsupervised and LLM-based methods.

Since the problem at hand concerns a novel and previously unexplored authorship case, we employed authorship attribution models that have shown robust results in prior studies. Moreover, Constance Wilde's literary publications are limited to a short volume of fairy tales. Thus, the data availability constraint represents a main challenge in the current study. Consequently, we first selected two approaches described by Dinu et al. (2008): Classification based on Support Vector Machines (SVM) with a string kernel, and Clustering via Rank Distance - which have yielded positive results in authorship attribution tasks. This selection of complementary models allows us to cover both supervised and unsupervised modeling paradigms.

Furthermore, LLMs have increasingly been studied for their ability to capture complex linguistic and stylistic patterns, showing promise in the field of authorship attribution (Huang et al., 2024; Habib et al., 2025; Huang et al., 2025b,a). In this study,

[*]These authors contributed equally to this work.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 381–389
July 4, 2026 ©2026 Association for Computational Linguistics

we aim to further explore their potential and evaluate their performance on small, author-specific corpora by drawing from the methodology of Huang et al. (2025b) and exploring authorship attribution via LLM Perplexity.

At the core of stylometry lies the interplay between style, genre and stylistic fingerprint. While *style* is concerned with conscious aesthetic choices made by writers (Sotirova, 2015) and *genre* emerges from structural and rhetorical conventions of different types of texts (Biber and Conrad, 2019), it is the *stylistic fingerprint* of a writer that computational attribution techniques are most often concerned with - capturing the so-called "human stylome" (Van Halteren et al., 2005). So far, computational stylometric methods have been employed in the cases of a great number of authorship uncertainties, in both literary and non-literay fields (Mosteller and Wallace, 1964; Craig and Kinney, 2009; Labbé and Labbé, 2001; Juola, 2015; Dinu et al., 2012a; Tuzzi and Cortelazzo, 2018).

Our study brings a contribution to the field by presenting a novel case of authorship attribution in which questions of collaborative writing and editorial influence can be explored in a setting of little data availability. We investigate the potential involvement of Constance Wilde alongside Oscar Wilde, examining the manuscript evidence in relation to the final published version. In addition, we introduce a dataset of short fairy tales, enabling reproducible computational analysis and offering a resource for future studies in authorship attribution. This approach not only sheds light on a previously unexplored literary case but also demonstrates the applicability of stylometric techniques. Finally, we explore whether LLMs can yield superior results to traditional feature-based methods.

In order to achieve this, the rest of the article is structured as follows: Section 2 describes the dataset; Section 3 presents the methodology; Section 4 discusses the results; Section 5 draws the conclusions and Section 6 discusses the Ethical Considerations.

2 Data

The dataset used in this study consists of short fairy tales authored by Oscar Wilde and Constance Wilde. Given the fact that Constance Wilde only authored one volume of tales, we limited the Oscar Wilde dataset to a similar size and genre. Oscar Wilde's published collection, *The Happy Prince and Other Tales* (1888), contains 5 stories including "The Selfish Giant". In addition, we include the manuscript of "The Selfish Giant" in the handwriting of Constance Wilde with pencil annotations by Oscar Wilde, held at the *Morgan Library & Museum*, New York City. Constance Wilde's collection, *There Was Once: Grandma's Stories* (1888), is also included. We excluded the tales in verses *Little Bo Beep*, *Old Mother Hubbard*, *Babes in the Wood* and *The Three Little Kittens*, leaving five tales from her collection.

Table 1: Corpus summary with word counts

Author	Text	Words
Oscar Wilde	The Happy Prince	3473
Oscar Wilde	Nightingale and Rose	2328
Oscar Wilde	The Selfish Giant (published version)	1658
Oscar Wilde	The Selfish Giant (manuscript)	1652
Oscar Wilde	The Devoted Friend	4300
Oscar Wilde	The Remarkable Rocket	4384
Constance Wilde	Cinderella	1181
Constance Wilde	Jack the Giant Killer	1830
Constance Wilde	Little Red Riding Hood	837
Constance Wilde	Puss in Boots	860
Constance Wilde	The Three Bears	759

Table 1 shows a complete description of the corpus. All texts were processed to remove non-story content such as introductions, illustrations, and page numbers. The dataset is available for research purposes, and all texts are in the public domain. Oscar Wilde's stories were obtained from Project Gutenberg.

Finally, we computed the Jaccard Similarity in order to assess the rough similarity between the two versions of "The Selfish Giant":

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are the sets of unique words in each text. We obtained a score of 0.89, indicating a high degree of overlap between the manuscript and the published version.

3 Methods

As mentioned above, we aim to investigate the authorship question using three distinct methods: Classification based on Support Vector Machines (SVM) with a string kernel, Clustering via Rank Distance, and Authorship Attribution based on LLM Perplexity.

3.1 Classification Methodology

We followed the methodology described by Dinu et al. (2008), modeling a 2-class classification problem using SVM with a string kernel.

String kernels are particularly suited to authorship attribution problems because they treat texts as sequences of characters rather than words, meaning no preprocessing is needed other than space normalization. The similarity between two texts is measured by counting shared substrings of length p (p-spectrum kernel). By choosing a small value for p (usually 2-5), the kernel will mostly capture a mixture a suffixes, function words and punctuation - all strong indicators of stylome. To capture stylistic patterns at multiple scales, we used a blended spectrum kernel, which sums similarities over all substring lengths from 1 to p:

$$k_1^p(s,t) = \sum_{q=1}^{p} k_q(s,t)$$

To account for differences in text lengths, the kernel was normalized:

$$\hat{k_1^p}(s,t) = \frac{k_1^p(s,t)}{\sqrt{k_1^p(s,s) \cdot k_1^p(t,t)}}$$

Given the limited size of the dataset, a nested Leave One-Out Cross Validation (LOOCV) scheme was adopted. In the outer loop, each fairy tale was iteratively used as the test instance while the remaining texts were used for training. Within each training set, an internal leave-one-out procedure was performed to estimate the model's stability and generalization capacity. The internal accuracy reported corresponds to this inner validation step, while the final prediction for each tale corresponds to the outer test fold. We also created two datasets, to account for the two versions of "The Selfish Giant": one including the published version, the other the manuscript version.

The two key hyperparameters of the model were chosen empirically. The substring length p=4 was selected to balance sensitivity to local stylistic patterns and computational feasibility. Similarly, ν=0.05 was chosen empirically after preliminary experiments indicated that this value achieved stable performance while limiting margin violations to at most 5% of the training samples. Both parameters were thus fixed for the final cross-validation and leave-one-out experiments to ensure consistency across all classification runs.

3.2 Clustering Methodology

Hierarchical Agglomerative Clustering is a bottom-up approach that iteratively merges clusters (initially represented by each individual data point) by calculating the distance between every pair of clusters. The final resulting dendrogram shows the hierarchy of clusters and the relationships between them. Thus, when working with this distance-based clustering technique, the two main aspects to be considered are the selected features and the distance measure.

a	been	had	its	one	that	was
all	but	has	may	only	the	were
also	by	have	more	or	their	what
an	can	her	must	our	then	when
and	do	his	my	shall	there	which
any	down	if	no	should	things	who
are	even	in	not	so	this	will
as	every	into	now	some	to	with
at	for	is	of	such	up	would
be	from	it	on	than	upon	your

Table 2: List of functional words used in the clustering experiments

Since our aim is to capture the stylistic fingerprints of the two authors, the selected features need to unambiguous, quantifiable and used in an unconscious way.

Function words, first introduces by Mosteller and Wallace (1964), have been proposed as linguistic features that fit the criteria. It has subsequently been proven that they are indeed used unconsciously and carry part of the stylistic fingerprint of individuals (Chung and Pennebaker, 2011). We therefore based our experiments on the list of functional words introduced by Mosteller and Wallace (1964), which can be seen in Table 2.

Previous studies in authorship attribution have shown that Rank Distance is a metric that yields consistent results in clustering together works pertaining to an author (Dinu et al., 2008; Popescu and Dinu, 2008; Dinu et al., 2012b). Rank distance (Dinu, 2003) is an ordinal metric related to Spearman's footrule. To calculate the Rank Distance, the following steps were followed:

1. Calculate the raw frequencies of the selected functional words in each text

2. Replace the raw frequencies with the rank each functional word occupies in each text. The word with the highest frequency will be assigned the highest rank, while the words that are present on the function words list but absent in the text will be assigned the rank 0.

Tied objects are assigned the average rank of the positions they share.

3. Compute the pairwise rank distance between each two texts according to L1 norm, where $\sigma(i)$ represents the rank of the object:

$$D(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|$$

4. Obtain a square similarity matrix, which serves as input for the Hierarchical Clustering algorithm.

For our experiments we used Hierarchical Agglomerative Clustering with average linkage.

3.3 LLM Perplexity Methodology

While traditional authorship attribution methods are based on hand-crafted features, attribution via LLMs leverages deep contextual representations learned from vast corpora of text. Attribution via LLMs is still a relatively new area of research, and since the current study concerns an authorship case where the ground truth is not known, we opted for a simple approach to examine whether LLM-based analysis can provide additional insights. We followed the approach of Huang et al. (2025b), which proved competitive with state-of-the-art methods in authorship attribution. In this framework, a separate language model is fine-tuned on the texts of each candidate author. The resulting model is intended to capture the stylistic and lexical patterns characteristic of a specific writer, including preferences in vocabulary, phrasing, and contextual word usage. The fine-tuned models are then used to evaluate how well each one predicts an unknown text by computing its perplexity. For a causal language model m and a text tokenized as a sequence $T = \{x_1, x_2, \ldots, x_t\}$, perplexity is defined as:

$$\text{PPL}(m, T) = \exp\left(-\frac{1}{t}\sum_{i=1}^{t} \log p_m(x_i \mid x_{<i})\right) \quad (1)$$

where $p_m(x_i \mid x_{<i})$ denotes the probability assigned by the model m to the token x_i given all preceding tokens $x_{<i}$, and t denotes the length of the token sequence T. Intuitively, perplexity measures how "surprised" a model is by a text: a model that has been fine-tuned on an author's writing will assign higher probabilities to new texts written in a similar style. The candidate author whose model yields the lowest perplexity is considered the most likely author of the questioned document.

In contrast to Huang et al. (2025b), who perform full fine-tuning of GPT-2, we fine-tune the GPT-2 model via LoRA (Hu et al., 2022), a parameter-efficient method that updates only a small set of low-rank adapter matrices while keeping the base model weights frozen. This reduces computational cost while preserving the base model's language understanding.

We trained the Oscar model on a corpus of 8 short stories and a novel authored by Oscar Wilde, and the Constance model on three fairy tales attributed to her, leaving one for evaluation. Below are the configurations used for training the two models, chosen empirically:

Parameter	Value
LoRA rank (r)	16
LoRA alpha	32
Target modules	`c_attn`, `c_proj`
LoRA dropout	0.05
Bias	none
Task type	CAUSAL_LM
Chunk size	128 tokens
Gradient accumulation steps	2
Number of epochs	20
Learning rate	1×10^{-4}
Optimizer	AdamW (PyTorch)

Table 3: Training configuration for LoRA fine-tuning of the Constance-finetuned model

Parameter	Value
LoRA rank (r)	16
LoRA alpha	32
Target modules	`c_attn`, `c_proj`
LoRA dropout	0.05
Bias	none
Task type	CAUSAL_LM
Chunk size	512 tokens
Tokenizer pad token	EOS token
Gradient accumulation steps	2
Number of epochs	20
Learning rate	3×10^{-4}
Optimizer	AdamW (PyTorch)

Table 4: LoRA fine-tuning and training configuration for the Oscar-finetuned model

4 Results

The following section presents the results and analysis of our experiments.

4.1 Classification Results

We ran the classification experiments on two datasets: the first one including the manuscript version of the story "The Selfish Giant" and the second one including the published version of the text.

The results with the first dataset are presented in Table 5, and the second set of results is presented in Table 6. The results are highly similar: when the tale "The Selfish Giant" is included in the inner LOOCV training loop, the training accuracy of the model is low: and average of 0.55 for the first dataset and 0.58 for the second one. The presence of the story "The Selfish Giant" labeled as pertaining to Oscar Wilde in the training set obfuscates the model's ability to learn meaningful patterns and differentiate between the two authors, especially with the training set being small (just 9 entries). However, when the story "The Selfish Giant" is held out as a test document, the model reaches perfect accuracy and high confidence in both datasets (highlighted entry in Tables 5 and 6).

This fact that including the story "The Selfish Giant" in the training set degrades both training accuracy and confidence, inhibiting the model's ability to learn meaningful patterns, stands as evidence for the story's pronounced stylistic dissimilarity with the rest of the Oscar Wilde texts.

We reiterate the fact that the changes made to the manuscript version were most likely carried out by Oscar Wilde (based on the fact that the pencil notes on the manuscript are in Oscar Wilde's handwriting). The fact that both versions of the text are attributed to Constance Wilde with relatively high confidence (second highest in the first dataset, third highest in the second) points to the fact that Constance Wilde's input regarding "The Selfish Giant" goes deeper than mere editing. Moreover, Oscar Wilde's changes do not influence or improve the model's ability to differentiate between the two authors.

The results of this set of experiments bring initial insights into the case of "The Selfish Giant", showing evidence towards the hypothesis of at least a tight collaboration between Constance and Oscar Wilde in writing this tale.

4.2 Clustering Results

We ran the Clustering experiments on four datasets:

1. Full Corpus, including both versions of "The Selfish Giant" (Figure 1a)
2. A corpus including only the published version of "The Selfish Giant" (Figure 1b)
3. A corpus including only the manuscript version of "The Selfish Giant" (Figure 1c)
4. A corpus excluding all versions of "The Selfish Giant" (Figure 1d)

For ease of interpretation, in all of the plots the texts published by Constance Wilde are written in blue and those published by Oscar Wilde in red.

As is evident from the first three figures, reducing the texts to only functional words renders the two versions of "The Selfish Giant" are virtually indistinguishable. The two hierarchies in Figures 1b and 1c are identical: there is a clear separation between the cluster of texts published by Constance Wilde and the cluster of texts published by Oscar Wilde, with the exception of "The Selfish Giant" which is well integrated among Constance Wilde's writings.

Moreover, when we ran the Clustering experiment without including "The Selfish Giant" (Figure 1d), the algorithm produced a hierarchy that perfectly matches our dataset, the writings of the authors producing two clearly separated clusters. Since the clustering experiments are based on functional words only (the words that are unconsciously used by individuals), this result suggests that, if any literary collaboration occurred between the two spouses, it did not extend to a depth that would cause their writings to become stylistically "contaminated" by one another. In other words, this result proves that, despite the similarity of genre and the fact that the stories were published and probably written in the same year in a period when the two were still living together, both of them produced original and stylistically independent works.

The only exception is, then, "The Selfish Giant". When included in the dataset, both versions of the text perfectly integrate within Constance Wilde's cluster. Furthermore, despite the fact that the changes made to the manuscript might seem, on close reading, "significant" (according to Moyle (2012)), when the text is analyzed only based on those features that occur independently of the author's design the two versions of the text become so close that they form their own subcluster within that of Constance Wilde (Figure 1a)..

The results of this set of experiments reinforce the findings of the classification analysis: there is a clear separation between the stylometric profiles of

Table 5: SVM classification results with each fairy tale iteratively used as the test text. Dataset including the published version of "The Selfish Giant"

Test Text	True Author	Predicted Author	Conf.	Int. Acc.
The Happy Prince (Oscar)	Oscar	Oscar	58.71	0.78
Nightingale and Rose (Oscar)	Oscar	Oscar	57.62	0.44
The Selfish Giant published (Oscar)	Oscar	Constance	76.78	1.00
The Devoted Friend (Oscar)	Oscar	Oscar	54.09	0.78
The Remarkable Rocket (Oscar)	Oscar	Oscar	60.92	0.44
Cinderella (Constance)	Constance	Oscar	50.00	0.44
Jack the Giant Killer (Constance)	Constance	Oscar	54.62	0.56
Little Red Riding Hood (Constance)	Constance	Constance	78.67	0.44
Puss in Boots (Constance)	Constance	Constance	58.81	0.56
The Three Bears (Constance)	Constance	Oscar	53.63	0.56

Table 6: SVM classification results with each fairy tale iteratively used as the test text. Dataset including the manuscript version of "The Selfish Giant"

Test Text	True Author	Predicted Author	Conf.	Int. Acc.
The Happy Prince (Oscar)	Oscar	Oscar	60.58	0.78
Nightingale and Rose (Oscar)	Oscar	Oscar	63.70	0.56
The Selfish Giant manuscript (Oscar)	Oscar	Constance	77.22	1.00
The Devoted Friend (Oscar)	Oscar	Oscar	58.22	0.67
The Remarkable Rocket (Oscar)	Oscar	Oscar	79.71	0.44
Cinderella (Constance)	Constance	Oscar	50.00	0.56
Jack the Giant Killer (Constance)	Constance	Oscar	61.11	0.56
Little Red Riding Hood (Constance)	Constance	Constance	90.08	0.56
Puss in Boots (Constance)	Constance	Constance	54.69	0.56
The Three Bears (Constance)	Constance	Oscar	54.05	0.56

the two authors, with "The Selfish Giant" emerging as the only outlier. The strong stylistic resemblance between this story and Constance Wilde's writings suggests that it may have been the product of a collaboration, or even authored by her alone.

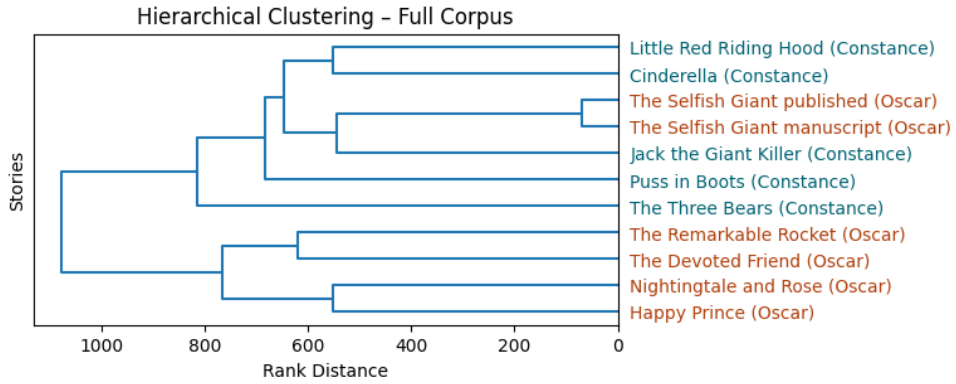
4.3 LLM Perplexity Results

The results of the two authorial language models are presented in Tables 7 and 8. A positive difference indicates that the fine-tuned model finds the text more predictable than the base model, suggesting stylistic similarity to the training author, while a negative difference suggests the opposite.

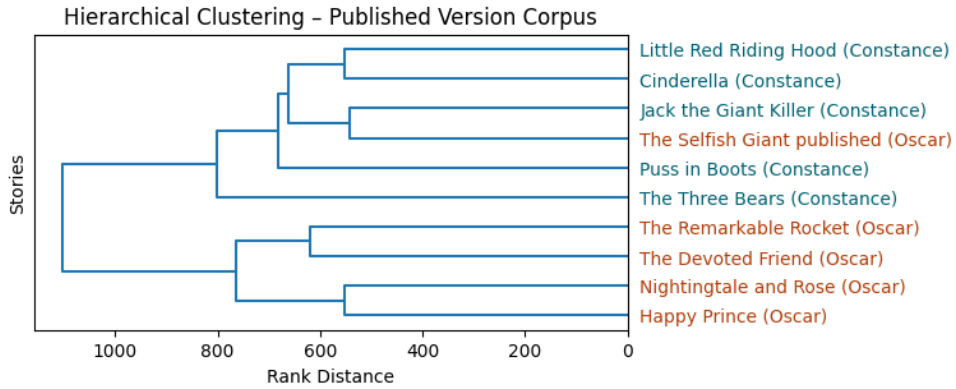
The Constance-finetuned model assigns the largest positive difference to Cinderella by a substantial margin (9.45), far exceeding all other texts. Oscar Wilde's fairy tales all receive negative values (ranging from -1.54 to -2.55), indicating that the model does not recognise them as stylistically similar to Constance's writing. Both version of *The Selfish Giant* receive small positive values (0.53 and 0.10 respectively), suggesting that after fine-tuning, the model assess them as being slightly closer to Constance's style.

The Oscar-finetuned model shows a complementary pattern. Oscar Wilde's four fairy tales receive the highest positive differences, confirming that the model successfully captured his stylistic fingerprint. All of Constance's fariytale receive negative perplexity scores, suggesting the model distinguishes between the style of the two authors. Interestingly, the published version of *The Selfish Giant* receives a small positive score, while the manuscript version receives a negative score, suggesting the model finds the latter more aligned with Constance's style.

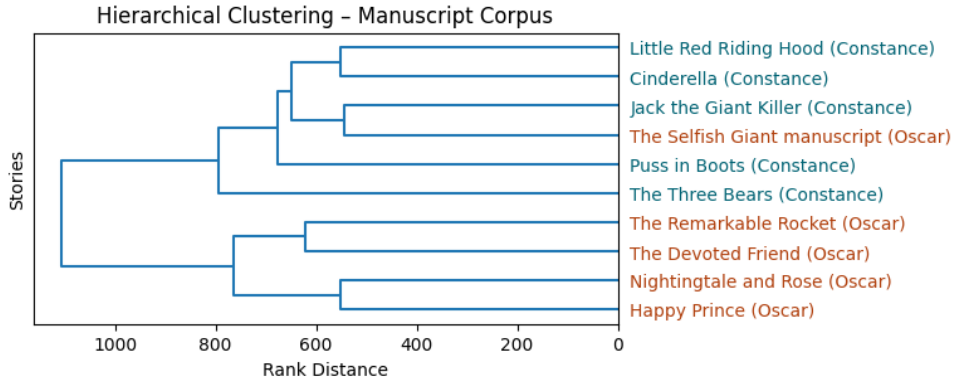
Both models indicate towards a shared stylistic profile between the story *The Selfish Giant* and Constance's writings. There is a compelling difference in the case of the manuscript under Oscar-finetuned model, where the manuscript is clustered closely with Constance Wilde, while the published version is closer to the Oscar texts. We also note that, due to data availability, the Oscar model was fine-tuned on a greater amount on text than the Constance model. However, our results prove that fine-tuning on even a small amount of data is sufficient for



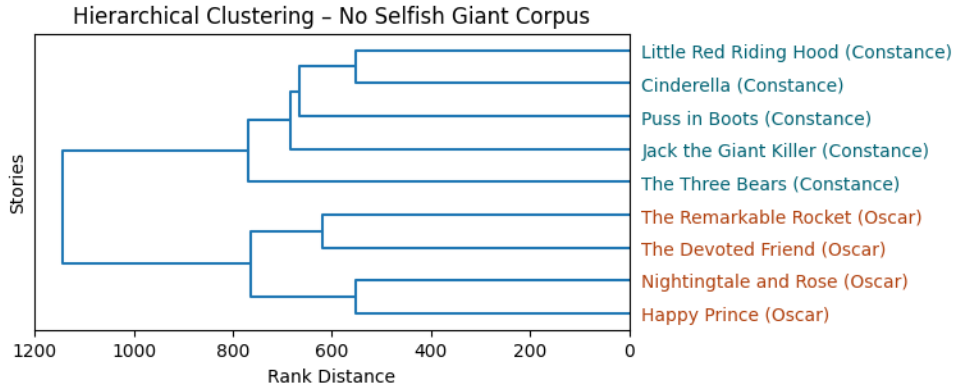
(a) Full Corpus



(b) Published Version Corpus



(c) Manuscript Version Corpus



(d) No “Selfish Giant” Corpus

Figure 1: Clustering Results for different versions of the corpus.

Table 7: Perplexity results for the Constance-finetuned model

Text	Base	Fine	Diff.
Cinderella (Constance Wilde)	50.57	41.12	9.45
The Selfish Giant	21.99	21.46	0.53
The Selfish Giant (manuscript)	18.58	18.48	0.10
The Star-Child (Oscar Wilde)	19.56	21.10	-1.54
The Fisherman and his Soul (Oscar Wilde)	20.97	22.92	-1.95
The Birthday of the Infanta (Oscar Wilde)	35.76	38.05	-2.29
The Young King (Oscar Wilde)	25.36	27.74	-2.37
A House of Pomegranates (Oscar Wilde)	29.79	32.34	-2.55

Table 8: Perplexity results for the Oscar-finetuned model

Text	Base	Fine	Diff.
The Birthday of the Infanta (Oscar Wilde)	35.76	30.89	4.87
A House of Pomegranates (Oscar Wilde)	29.79	26.14	3.65
The Fisherman and his Soul (Oscar Wilde)	20.97	18.80	2.17
The Star-Child (Oscar Wilde)	19.56	17.69	1.87
The Selfish Giant	21.99	20.27	1.72
Cinderella (Constance)	50.57	50.63	-0.06
The Selfish Giant (manuscript)	18.58	18.64	-0.06
Little Red Riding Hood (Constance)	25.08	25.27	-0.19
The Three Bears (Constance)	12.41	13.51	-1.10
Puss in Boots (Constance)	21.70	22.98	-1.29
Jack the Giant Killer (Constance)	21.60	24.06	-2.46

capturing nuances in stylistic profiles.

5 Conclusion

We introduced and analyzed a novel authorship attribution case - the children's stories authored by Constance and Oscar Wilde, which, to the best of our knowledge, have not been included in a computational authorship analysis up to date.

We analyzed the corpus using supervised (SVM classification with a string kernel) unsupervised (Hierarchical Clustering) and LLM-based methods (LLM-perplexity of fine-tuned models). Despite the small amount of available data, all three independent methods converged towards a shared conclusion: Constance Wilde might have authored or participated in the writing of one of the stories published in Oscar Wilde's collection.

Moreover, our results show that Constance Wilde was, at the time, a writer in her own right, with a style distinguishable from that of her husband.

Further studies could include collaborative authorship analyses through paragraph-level exploration in order to expand the collaboration hypothesis and lexical stylometric methods to provide insights into the similarities of "The Selfish Giant" and the rest of the texts authored by the two writers.

6 Ethical Considerations

This study makes use of literary works that are no longer under copyright protection. The research does not involve human participants, experiments, or any data that would raise ethical or privacy concerns.

7 Acknowledgements

This research was supported by the Ministry of Education and Research, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023- 1701, within PNCDI IV.

References

Anne Clark Amor. 1999. Constantly undervalued: A centenary appreciation of constance wilde. *The Wildean*, (14):8–25.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Cindy Chung and James Pennebaker. 2011. The psychological functions of function words. In *Social communication*, pages 343–359. Psychology Press.

Hugh Craig and Arthur F Kinney. 2009. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press.

Anca Dinu, Liviu P Dinu, Alina Resceanu, and Ion Resceanu. 2012a. Some issues on the authorship identification in the apostles' epistles. In *LRE-Rel Workshop on Language Resources and Evaluation for*

Religious Texts, co-located with LREC 2012, Istanbul, pages 18–24.

Liviu P Dinu. 2003. On the classification and aggregation of hierarchies with ifferent constitutive elements. *Fundamenta Informaticae*, 55(1):39–50.

Liviu P Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012b. Pastiche detection based on stopword rankings. exposing impersonators of a romanian writer. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 72–77.

Liviu Petrisor Dinu, Marius Popescu, and Anca Dinu. 2008. Authorship identification of romanian texts with controversial paternity. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Nudrat Habib, Tosin Adewumi, Marcus Liwicki, and Elisa Barney. 2025. Trends and challenges in authorship analysis: A review of ml, dl, and llm approaches. *arXiv preprint arXiv:2505.15422*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025a. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Weihang Huang, Akira Murakami, and Jack Grieve. 2025b. Attributing authorship via the perplexity of authorial language models. *PloS one*, 20(7):e0327081.

Patrick Juola. 2015. The rowling case: a proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(suppl_1):i100–i113.

Cyril Labbé and Dominique Labbé. 2001. Intertextual distance and authorship attribution corneille and molière. *Journal of Quantitative Linguistics*, 8(3):213–231.

Frederick Mosteller and David Wallace. 1964. *Inference and disputed authorship: The Federalist*. Reading, MA: Addison-Wesley.

Franny Moyle. 2012. *Constance: The Tragic and Scandalous Life of Mrs. Oscar Wilde*. Open Road Media.

Marius Popescu and Liviu P Dinu. 2008. Rank distance as a stylistic similarity. In *Coling 2008: Companion volume: Posters*, pages 91–94.

Kerry Powell and Peter Raby. 2013. *Oscar Wilde in Context*. Cambridge University Press.

Violeta Sotirova. 2015. *The Bloomsbury Companion to Stylistics*, 1st edition. Bloomsbury Companions. Bloomsbury Academic, London.

Arjuna Tuzzi and Michele A Cortelazzo. 2018. *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*. Padova University Press.

Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Evaluating Open-Source LLMs for Text Summarization and Named Entity Recognition in Apartheid Witness Reports

Pauline Kister
Technical University of Munich
pauline.kister@tum.de

Miriam Schirmer
Northwestern University
miriam.schirmer@northwestern.edu

Abstract

This work investigates the extent to which open-source Large Language Models (LLMs) can improve accessibility of unstructured historical documents by performing abstractive summarization and fine-grained Named Entity Recognition (NER) for role classification and violation types. We evaluate open-source LLMs in zero-shot settings and apply these tasks to witness testimonies collected by the South African Truth and Reconciliation Commission (TRC), which archived a large body of text documenting human rights violations during apartheid. Despite their historical significance, these texts are difficult to access due to their length, lack of standardized structure, and the absence of systematic indexing. Open-source LLMs show strong performance in summarization, with most models surpassing non-LLM baselines (maximum BERTScore 0.77), while NER performance remains limited (maximum F1-score 0.61). Results suggest a trade-off in which stylistic fluency is prioritized over factual precision. A two-stage pipeline, summarization followed by NER on LLM summaries, leads to measurable improvements.

1 Introduction

The preservation of historical records is essential for understanding the past and informing future generations. Natural Language Processing (NLP) tasks such as summarization and Named Entity Recognition (NER) can support this goal by making large, unstructured archives more accessible and searchable, particularly through transparent and reproducible open-source approaches.

In South Africa, the Truth and Reconciliation Commission (TRC)[1] produced an extensive collection of testimonies that document the experiences of victims and witnesses of human rights violations during apartheid. Prior research on witness testimonies from atrocity tribunals highlights the historical and legal value of such material (Keydar, 2020; Schirmer et al., 2022, 2025), suggesting that the TRC collection constitutes a valuable resource. However, it is difficult to access for a wider audience due to its volume, unstructured format, and the linguistic complexity of first-person testimonies. NLP can make such collections more accessible through automatic extraction and summarization of key information. Conventional NLP techniques frequently encounter difficulties with culture- and domain-specific vocabulary, irregular formatting, and historical contexts (Blouin et al., 2021). Although primarily designed for text generation, Large Language Models (LLMs) have demonstrated high performance on a wide range of NLP tasks, including information extraction (Hilgert et al., 2024; Kostina et al., 2025). Open-source LLMs of increasing quality can now be applied transparently and reproducibly, without task-specific training.

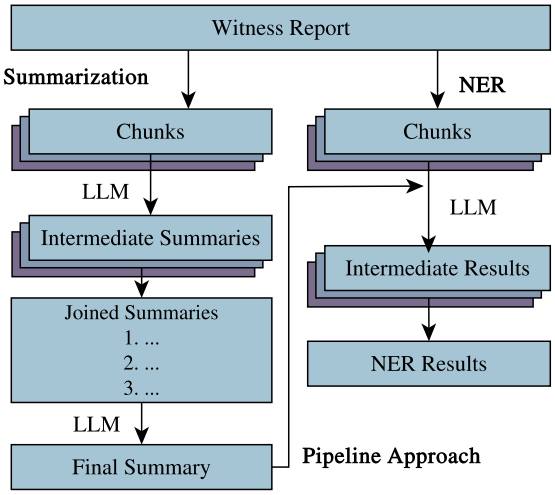


Figure 1: Study Overview.

Building on this, we examine whether open-source LLMs can reliably extract key information

[1] https://www.justice.gov.za/trc/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 390–410
July 4, 2026 ©2026 Association for Computational Linguistics

from complex, unstructured texts, focusing on two tasks in particular: abstractive summarization and NER of victim and witness names, as well as the identification of violations discussed in the respective hearing. While the extraction tasks technically combine aspects of NER, classification, and information extraction, we refer to them as NER tasks throughout this work for simplicity. This is consistent with established work on fine-grained NER, where entity types extend beyond standard categories such as PERSON or LOCATION to include domain-specific roles (Choi et al., 2018; Zhang et al., 2020a).

LLMs are particularly promising for these tasks, since most existing NER models are limited to standard entity types and require task-specific fine-tuning. In contrast, our task demands the identification of fine-grained, context-dependent roles, such as distinguishing between victims and witnesses, without additional training. This paper's contributions are as follows:

1. We show that **open-source LLMs are highly effective at abstractive summarization** of long and complex historical documents (BERTScores of up to 0.77), being competitive with commercial models, while achieving moderate performance on entity detection (F1-scores up to 0.61).

2. We present performance trends with a **comparative analysis of different open-source models**;

3. introduce a **pipeline approach** that first summarizes the text and then applies NER, resulting in significantly improved results; and

4. provide a **curated dataset** of 200 witness reports corresponding to over 1,000 pages of court documents, including human- and LLM-generated summaries and structured metadata such as victim and witness names and the violations discussed.

We use free and open-source language models, with a focus on small systems, aiming to reduce computational costs and improve usability in research contexts with limited resources. All data and code are available online.[2]

[2] https://github.com/ge56dit/llm-apartheid-summarization-ner

2 LLMs for Summarization and NER

For abstractive summarization, LLMs have been established as the state-of-the-art method, consistently outperforming previous approaches. In contrast, LLMs are rarely the method of choice for NER due to their focus on text generation tasks. Despite the challenges for LLMs in NER, their flexibility and high context-understanding make them a promising option when dealing with fine-grained NER and classification on complex texts.

Summarization for Long Inputs. With the introduction of LLMs, the capabilities of abstractive summarization methods have improved significantly. Models such as GPT-3 (Brown et al., 2020) have produced coherent and factual summaries, being fine-tuned to fit a specific task (Alexandr et al., 2021; Rallapalli et al., 2025) or used in zero- or few-shot settings (Leiva-Araos et al., 2025; Liu and Healey, 2023; Pokale et al., 2023).

One of the biggest challenges in summarization is how to deal with long inputs. A document is usually considered long if it exceeds the maximal input length of state-of-the-art models, causing biases and lower performance (Ravaut et al., 2024; Hosseini et al., 2024). The most recent models allow for very long to theoretically unlimited inputs; however, hardware limitations still require strategies for long inputs. Recent work suggests solutions, such as chunking the text and merging the results, intelligently truncating parts of the text that are unlikely to influence downstream tasks, or preprocessing with extractive summarization methods (Chang et al., 2024; Hosseini et al., 2024).

Summarization for Historical Documents. When working with historical texts, language models often encounter problems such as unknown vocabulary or scanning errors (Lyu et al., 2021). Task-independent word embeddings trained on historical documents like HistBERT (Wenjun Qiu and Xu, 2022) can partially address these problems. In addition, several methods have been developed specifically to summarize historical texts, e.g., using temporal information to cluster sentences and produce extractive summaries (Gung and Kalita, 2012).

For in-context learning with LLMs specifically for historical documents, research is more scarce. Murugaraj et al. (2025) conclude that fine-tuned neural models still outperform general-purpose LLMs in this domain and that the historical source

of their input texts contributes to the lower performance of LLMs, suggesting that LLMs struggle with the challenges of historical texts. In contrast, Zhang et al. (Zhang et al., 2024a) report that GPT3.5 achieved better results than other models fine-tuned for summarization of historical texts, but is still struggling with hallucinations.

NER for Historical Documents. Just as for summarization, NER in historical texts poses distinct challenges due to archaic orthography, digitization errors, and the scarcity of annotated data. Historical texts often contain OCR-induced errors, which degrade NER performance (Hamdi et al., 2019). However, careful evaluation and targeted error-correction methods can substantially mitigate these effects (Boros et al., 2020; Trias et al., 2021; Won et al., 2018). NER systems for historical documents primarily use deep learning, often combined with pretrained transformer-based embeddings (Aguilar et al., 2017; Ehrmann et al., 2016, 2023; Lample et al., 2016; Won et al., 2018). The use of transformer-based models led to improved results, especially in domains where limited data is available, such as historical texts (Schweter and Baiter, 2019) or in low-resource languages such as those of South Africa (Hanslo, 2022).

While LLMs excel at tasks such as summarization, their performance for NER has been ambiguous (Wang et al., 2023). In the historical domain, the quality of results for NER is even lower due to errors and domain-specificity (Tudor et al., 2025; González-Gallardo et al., 2024; Zhang and Colavizza, 2025). Methods to enhance the performance of LLMs include improving few-shot learning (Ashok and Lipton, 2023; Jiang et al., 2024b), cascading multiple LLMs (Luo et al., 2025), or self-correcting methods (Polak and Morgan, 2024; Wang et al., 2023; Xiao et al., 2024).

A Pipeline Approach to Fine-Grained NER Classification in TRC Texts. While most NER systems focus on high-level categories (e.g., person, location), our approach requires fine-grained classification to identify victims and witnesses. Existing NER models like TexSmart (Zhang et al., 2020a) can distinguish between more fine-grained labels such as towns and countries or identify victims in crime reports (Choi et al., 2018; Schirmer et al., 2024a). However, most fine-grained systems lack domain-specific categories, and pretraining constraints make it difficult to add new labels without retraining.

To address these challenges, we employ a pipeline approach to improve the results by first performing summarization, then NER. While it is known that splitting a task into multiple steps can improve an LLM's ability to follow instructions (Xie et al., 2024; Khot et al., 2023), this specific approach has rarely been investigated in detail. Nevertheless, NER on summarized texts has been performed successfully in the medical domain (Sasikala et al., 2024). There is also evidence that incremental summarization can help smaller language models perform sentiment classification, another task that is usually unrelated to text generation (Ma et al., 2018).

3 Dataset Collection and Annotation

We compile a new dataset from publicly available documents from the Truth and Reconciliation Commission (TRC). The TRC was established in 1996, a few years after the end of apartheid in South Africa, to investigate politically motivated human rights violations, provide reparation and rehabilitation, and grant amnesty (Truth and Comission, 1998). It collected more than 1300 witness testimonies about human rights violations, of which 200 were used in this study. Witness testimonies provide a direct and less formal retelling of events. They are mostly unscripted, offering emotionally charged but very detailed narratives. To ensure witnesses could testify in the language of their choice, translators were provided for 12 different languages (Truth and Comission, 1998). Most transcripts are available in English, usually without an indication of whether they were translated. This does not reduce the dataset's quality, as the translations were produced by professional court translators and preserve not only meaning but also the style, including errors and choice of wording (Gilbert and Heydon, 2021). Similar corpora have been compiled from genocide-related court proceedings (Schirmer et al., 2023, 2024b).

In contrast to the training of a neural network specialized in witness reports and summarization or NER, the LLM approach has the advantage of not requiring any training data and therefore no annotations. However, to validate model performance, reliable reference data is necessary.

To extend and diversify the validation corpus, three trained experts annotated a subset of 200 documents (ranging from $\sim$ 2 to 110 pages). Experts were part of the research team, including one

postdoc, one master's student, and one research assistant. All of the annotators followed detailed annotation guidelines (see A.1). To ensure sufficient annotator agreement, all annotators first annotated 10 identical documents. Inter-annotator agreement was measured using task-appropriate metrics: BERTScore for summarization, where free-text outputs cannot be compared categorically, and F1-score for the extraction tasks. Since we use the same metrics for model evaluation, these values indicate how much of the observed error can be attributed to the subjectivity of the task instead of a failure of the model. The agreement reached a good BERTScore of 0.75 for summaries, a very good F1-score of 0.94 for personal names, and an F1-score of 0.62 for type of crime labeling. The crime category is often named implicitly in the witness reports, and category boundaries may overlap, which makes the task highly subjective and constrains the achievable agreement. Since human summarization is also known to be a subjective task (Liu et al., 2023), we consider these results to be acceptable. In a group meeting, we discussed differences in labeling and adapted the guidelines accordingly. Each annotator then continued to annotate randomly chosen documents. The final dataset used for this study contained 200 transcripts, including over 1,000 pages of witness recounts describing human rights violations during apartheid, collected by the TRC and recorded between 1996 and 1997.

4 Methods

The models we compared were TinyLlama (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024a), OpenHermes (Teknium, 2023), Gemma3 (Team et al., 2025), and Phi-4 (Abdin et al., 2024). They are free to use and open-source, which ensures transparency and reproducibility. In addition, we included GPT-4o-mini (OpenAI, 2024) as a comparison to commercial models. To explore the influence of instruction fine-tuning on the model's responses, we also applied a second version of Mistral, uMistral, which receives unstructured prompts (default system prompt; instructions and input text in a single message) (see Table 1 for a model overview). Excluding the baseline models, we used the same models for both tasks. Hyperparameter tuning was conducted for all models with a manually supported grid search, mostly with parameters controlling the abstraction level (e.g., temperature, beam search parameters) and penalties for repetition and output length (A.2.4, Table 4).

For all prompted models, the prompt template was identical and systematically engineered (see A.2 for the final templates). Due to the limited context length of the models, reports sometimes had to be split into chunks at the sentence level before they could be passed to the model. To avoid excessive splitting and a loss of context, the prompt template had to be as short as possible, which is why we only consider zero-shot settings. Each text chunk was wrapped in this template and then formatted as expected by the instruction tuning of each model: One message for the system prompt, one for the instructions, and one for the input text. uMistral only received a single message consisting of the input text and instructions, disabling some aspects of the instruction tuning.

4.1 Summarization

This work focuses on abstractive summarization because LLMs are more suited to rephrase and generalize the original text than to determine the importance of individual sentences. We compared the performance of the LLMs against several smaller baseline models fine-tuned for summarization: BART (Lewis et al., 2019), T5 (Raffel et al., 2019), and Pegasus (Zhang et al., 2019). We chose general-purpose models rather than models specialized for historical texts, because available encoder-only models like HistBERT (Wenjun Qiu and Xu, 2022) are not comparable to fully trained encoder-decoder summarization models.

Experimental Setup. The models were prompted to write a short summary that describes the course of events in each text chunk. For the full prompt, see A.2.1. The resulting summaries were joined together. Experiments showed that clearly separating and enumerating the summaries is more effective than naive concatenation. The joined summaries were then passed back to the model to create a single, coherent text. If the joined summary was longer than the model's input, it had to be split again, and the process was repeated recursively.

We conducted additional experiments to assess how context window size and recursive summarization affect summary quality by uniformly limiting input length across models and by truncating intermediate summaries. These test how reduced con-

Table 1: Overview of the models.

Name	Version	Params	Context	Key Feature
T5	long-t5-tglobal-base	220M	4,096	Summarization-only, long context
BART	bart-large-cnn	403M	1,024	Summarization-only, BERT-based
Pegasus	pegasus-cnn-dailymail	568M	1,024	Summarization-only
TinyLlama	Chat-v1.0	1.1B	2,048	Highly compact LLM
Mistral	Instruct-v0.3	7.3B	32,768	Mistral with prompt formatting
uMistral	Instruct-v0.3	7.3B	32,768	Mistral without prompt formatting
Mixtral	Instruct-v0.1	8x7B	32,768	Mixture of experts
OpenHermes	2.5-Mistral-7B	7.3B	32,768	Strongly instruction-tuned Mistral
Gemma3	gemma-3-4b-it	4B	128,000	Good long-context abilities
Phi-4	Phi-4	14B	32,768	Good factuality and reasoning
GPT	GPT-4o-mini	unknown	128,000	Commercial model

text and recursive merging trade off against error propagation (see A.2.5).

Summarization Evaluation. The most straightforward way to evaluate a summary is to compare it to a human-written reference summary treated as ground truth. We rely on two widely adopted reference-based metrics: ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b). **ROUGE** is the average of three different metrics: the overlap of unigrams between the summary and reference, the overlap of bigrams, and the length of the longest common subsequence. **BERTScore** is the cosine similarity between the BERT embeddings of the model output and the reference. It compares texts on a semantic level and is less sensitive to paraphrasing than ROUGE.

Since reference summaries are not always available, we additionally use reference-free metrics. The **Unique Bigram Ratio** measures lexical diversity, the **Summarization Ratio** captures compression rate, and **Semantic Similarity** computes the BERT-based similarity between generated text and the source text to assess faithfulness, with high values potentially indicating extractiveness.

4.2 Named Entity Recognition

In the case of the TRC dataset, an important piece of information to extract is personal names, as they enable searching for documents related to a specific case. Similarly important is information about the type of crime that was committed (e.g., "shooting", "torture").

Experimental Setup. The NER process was similar to the process for summarization. The original text was split into chunks of maximum size for the model's context window, taking into account sentence boundaries. For personal names, models were prompted to extract the names of victims and witnesses from the text, while receiving definitions of who is considered to be a victim or witness (see the full prompt at A.2.2). Important improvements during prompt engineering included the phrase "most important names" to reduce false positives, and the instruction of what to do when the victim's or witness's names are not mentioned to avoid hallucinations. For violation classification, the model received short definitions for possible violation classes and was prompted to cite one or more of these labels.

Since conventional NER models cannot distinguish between different crime types and roles, they were unable to solve the exact task we propose. We still include a SpaCy (Honnibal et al., 2020) baseline for the victim and witness name recognition task to serve as a reference point for general-purpose person name recognition rather than role-specific entity extraction.

We conducted additional experiments to control for the effects of entity combination strategies and input length by comparing fuzzy matching and LLM-based aggregation and by reducing the maximal input to 1,024 tokens for all models (see A.2.6). In addition, we improved the results with a pipeline approach, where we applied NER on the reference summaries as well as on the model-generated summaries to assess whether summarization simplifies NER for long TRC documents.

NER Evaluation. For extracting personal names, the model was instructed to answer in JSON format. The first JSON object in the text was considered the model's answer; any text around it was ignored. For

the crime classification, the model was instructed to cite keywords from a given list (see the prompt A.2.3 for a definitive list). Only classes from the list were considered in the evaluation. We extracted lists of victims, witnesses, and classes from the model's answers and calculated precision, recall, and the micro F1-score. The token set ratio allowed for soft matching for personal names.

We additionally conducted a qualitative manual error analysis for summarization and NER in parallel. Starting with a subset of the evaluated documents, we compared the model outputs to both the reference annotations and the original source texts. This allowed the identification of potential explanations for model failures and document-specific challenges. Emerging patterns were then validated by searching for more instances in which the model showed similar behavior. In parallel, we identified outliers in the evaluation metrics, such as unusually long summaries or instances with many false positives. These cases were examined manually to determine the cause of the errors. Through this process, several patterns were identified that can give insights into the shortcomings of the models and the difficulties of the dataset.

5 Results

For summarization, GPT, Mixtral, and both Mistral models perform best with BERTScores of around 0.7 to 0.8. In contrast, the LLMs' NER performance is moderate. GPT, uMistral, and Mixtral are the only models that reach an F1-score close to 0.6. Most models yield results with high recall, but low precision. The pipeline approach improves these results.

5.1 Summarization

Across all models, performance varied notably in both ROUGE and BERTScore evaluations (see Tables 2 and A.3.1, Figure 4 and 5). In ROUGE, the Mistral models achieved the highest average score (both 0.26), followed closely by Mixtral and GPT (0.25). OpenHermes and Gemma3 still surpass the best baseline model, BART, which reached a ROUGE score of 0.19. Lower results were obtained by TinyLlama (0.17), Pegasus (0.14), and T5 (0.12). BERTScore rankings followed a similar trend. GPT led with a score of 0.78, closely followed by Mistral (0.77), uMistral (0.77), and Mixtral (0.75). BART (0.64) scored in the mid range, surpassing Phi-4 (0.63) and OpenHermes(0.64), while the other baseline models and TinyLlama scored significantly lower.

The unique bigram ratio indicates that BART, OpenHermes, and GPT models have greater lexical diversity and less repetition. The low score of T5 (0.64) hints at much repetition in the summaries. The variations in the summarization ratio indicate that models with lower scores, such as BART and Pegasus, may not have been able to cover all important facts. In comparison, models with a higher score, like Phi4 and GPT, may include too much unnecessary detail, although a high score does not say much about the factuality. Models like uMistral and GPT, which have a high similarity to the source text in combination with a good BERTScore, produce highly factual summaries. T5's high similarity and low BERTScore point towards highly extractive summaries that directly cite the original text.

Overall, LLMs outperformed other architectures across surface-level and semantic metrics. The exception is TinyLlama, which does not surpass the best baseline with either score. The highest-ranking non-LLM model (BART) is outperformed by the highest-ranking LLM (GPT) by 0.13 in BERTScore, and the highest-ranking open-source model (uMistral) by 0.11.

Most models were robust to input and output length constraints, with only minor BERTScore changes, including slight drops for most models and improved results for OpenHermes (A.3.1, Figure 6).

5.2 Named Entity Recognition

NER results revealed substantial differences between the models (Table 3; A.3.2, Figure 7). uMistral achieved the best results (F1-score 0.60), followed by GPT (0.59), Mixtral (0.58), and Mistral (0.51). All of these models achieved a much higher recall than precision, indicating that the models included false names in the list. Gemma3, Phi-4, and OpenHermes scored lower, but with a better precision and recall balance.

The SpaCy baseline could detect personal names in the text, but was by design not able to distinguish between victims and witnesses or to classify the crime type. By extracting all names found in the text, it reached an F1-score of 0.19 in the victim and witness name detection. This supports our findings that non-LLM models struggle to solve the NER-based role classification task.

The better-performing models reached higher

Table 2: Results of the main summarization experiment. Mean and standard deviation over all witness reports. Values closest to the reference are marked in blue.

Model	BERTScore	ROUGE	Bigram ratio	Sum. ratio	Sim. to source
BART	0.6567± 0.131	0.1878± 0.086	0.9436± 0.045	0.0241± 0.018	0.4686± 0.086
T5	0.4186± 0.129	0.1219± 0.074	0.6492± 0.299	0.0938± 0.067	0.6210± 0.158
Pegasus	0.5574± 0.171	0.1386± 0.064	0.8581± 0.128	0.0162± 0.014	0.4427± 0.098
TinyLlama	0.5218± 0.114	0.1651± 0.096	0.8596± 0.044	0.1109± 0.098	0.5506± 0.091
Mistral	0.7708± 0.084	0.2621± 0.119	0.8268± 0.066	0.0987± 0.058	0.6386± 0.088
uMistral	0.7673± 0.077	0.2640± 0.115	0.8074± 0.078	0.1045± 0.061	0.6439± 0.085
Mixtral	0.7504± 0.153	0.2503± 0.115	0.8379± 0.071	0.0934± 0.069	0.6001± 0.142
OpenHermes	0.6374± 0.219	0.2068± 0.106	0.9130± 0.080	0.0623± 0.043	0.5435± 0.115
Gemma3	0.7292± 0.092	0.2197± 0.110	0.7993± 0.086	0.1335± 0.064	0.6819± 0.078
Phi-4	0.6304± 0.143	0.1812± 0.097	0.8192± 0.063	0.1672± 0.106	0.6688± 0.092
GPT-4	0.7832± 0.061	0.2480± 0.121	0.8832± 0.042	0.1124± 0.065	0.6483± 0.074
Reference	1.0	1.0	0.8946± 0.056	0.0556± 0.033	0.5348± 0.077

Table 3: NER results across models with precision, recall, and micro F1-score (main experiment). Best scores are marked in blue.

Model	Precision	Recall	F1-score
TinyLlama	0.2097	0.0132	0.0248
Mistral	0.4256	0.6316	0.5086
uMistral	0.5137	0.7399	0.6064
Mixtral	0.4906	0.7136	0.5814
OpenHermes	0.3828	0.1933	0.2569
Gemma3	0.3713	0.4028	0.3864
Phi-4	0.4778	0.2945	0.3644
GPT-4o-mini	0.5057	0.7126	0.5916

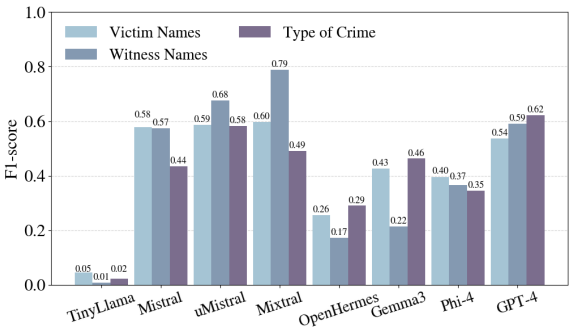


Figure 2: NER micro f1-scores for victim names, witness names, and violations.

scores at predicting witness names than victim names, indicating that this task is less complex (Figure 2). For the witness names, the highest F1-score of 0.79 is achieved by Mixtral, the model with the highest difference between victim and witness names. In contrast, the lower-performing models like OpenHermes and Gemma3 reached better results for the victim name. The crime classification scores are generally lower, with Gemma3 achieving surprisingly good results in contrast to a lower-performing Mistral.

Allowing the LLM to combine the results for each chunk independently through a final model call, thereby avoiding the possible errors introduced by fuzzy matching, results in slight improvement for the better-performing models, while models like OpenHermes and Gemma3 struggled with the combination (A.3.2, Figure 8).

5.3 The Pipeline Approach

Applying the **pipeline approach** that uses summaries as input for NER instead of the source text improved the results for 5 out of 8 models, with no changes for the other 3 models (see Figure 3). The results generally showed significant improvement over the main experiment results for the lower-performing models, and modest improvements for the higher-performing models, with a maximum improvement of 0.23 in F1-score for OpenHermes. NER on the human-written reference summary increased the F1-score for all models, especially the lower-performing models, sometimes by up to 0.4 in F1-score. These results represent an upper ceiling for the potential improvement that the pipeline approach could achieve on good summaries.

6 Discussion

The goal of this study was to evaluate open-source LLM performance for summarization and NER. The contrasting results – good scores for summarization and moderate scores for NER – show that, despite the LLMs' ability to solve a variety of

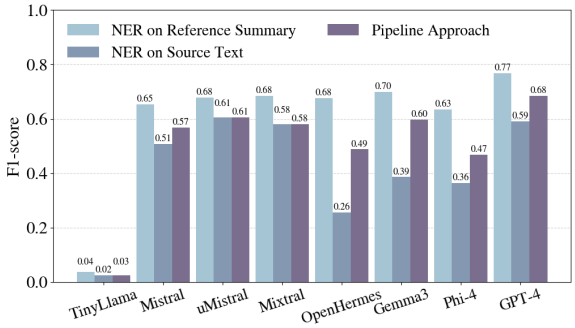


Figure 3: NER F1-scores on the reference summary, on the source text, and on the model-generated summary (pipeline approach).

tasks quickly without fine-tuning, they are mainly reliable for text generation tasks. Their greatest strength, their ability to produce coherent texts, is essential for summarization, but of less benefit for NER.

6.1 Summarization

Compared to NER and to the non-LLM models, summarization with LLMs yielded moderate to strong results, with the LLMs outperforming earlier neural approaches specifically fine-tuned for summarization. The difference between LLM and non-LLM models demonstrates that LLMs are a better choice for this task. The scale of their training data provides them with broader world knowledge and superior paraphrasing capabilities, a clear advantage, especially when considering that, due to their specificity and text format, the TRC reports require a high amount of background knowledge and abstraction.

Error Analysis. In general, the baseline models seem to struggle with overly extractive summaries, while the LLMs more often abstract too much, thereby generalizing to the point where the meaning is lost. Across models, distinct qualitative patterns emerge: TinyLlama sometimes replaced the content almost entirely with generic, hallucinated legal conclusions, while larger LLMs such as Mistral and especially Mixtral generated coherent and largely accurate narratives that recovered key details, but still occasionally included irrelevant or overly generic statements.

The following example is a summary excerpt by Mistral relating to the case of Peter Mabilo, a young man who was shot by the police (see A.4 for the full reference summary and more model summaries). Red coloring indicates hallucinations, green highlights relevant factual information, and blue text is not relevant for this summary:

> [...] Ms Mabozo, Luthli's grandmother, provided additional details about the shooting itself. She stated that after the police took Luthli's father away, they were told to take the children outside. Luthli was inside the house when the police shot him. The family found the house damaged and full of bullet holes [...]. In conclusion, the Commission expressed its sympathy for the family and vowed to investigate the circumstances surrounding Luthli's death [...].
> – Mistral, BERTScore 0.71, ROUGE 0.25

The last, irrelevant paragraph can be seen as an example of a generic statement typical for highly instruction-tuned models. Most models also struggle with reconstructing the course of events in temporal order, a known problem in NLP (Qiu et al., 2023; Barale et al., 2025)

6.2 Named Entity Recognition

NER results ranged from poor to moderate. Again, GPT, uMistral, and Mixtral achieved the best results. The quality of the NER results is hard to judge, since there is no existing baseline for this very specific task. The task differs from standard NER benchmarks in that it requires contextual role assignment rather than entity type detection. Other works that apply NER with LLMs can achieve better results with shorter texts, few-shot settings, and more extensive prompt engineering (Isaradech et al., 2024); however, it is also widely reported that LLMs struggle with NER (Lu et al., 2025). The results shown here are therefore to be expected when considering the difficulty of the task.

Interestingly, the highly instruction-aligned models performed worse than more general models, such as Mistral and Mixtral. While this may seem counterintuitive, those networks may be too specialized for such an unusual task. This effect can also be observed in the significant difference in performance with and without a correctly structured prompt for Mistral: The unstructured input and default system prompt partially disable the chat fine-tuning, resulting in outputs that are clearer and more concise. Therefore, uMistral, though being a relatively small open-source model, performs equally as good in NER as a commercial model like GPT-4o-mini.

Most models show higher recall than precision. This is also a result of the model's chat fine-tuning, which encourages it to provide very verbose and detailed answers (Zhang et al., 2024b). The models might therefore add false names to the result to appear more elaborate.

The mediocre performance of TinyLlama is likely due to its small size compared to the other networks. However, Phi-4, as the largest model, is outperformed by three different smaller models. This shows that as long as the models are sufficiently large to have decent language-understanding capabilities, the instruction training and specialization of the model seem to matter more than the number of parameters.

The success of the pipeline approach shows the effectiveness of splitting a task into multiple subtasks of less complexity. The summarization process cuts out parts of the text that are purely procedural and shifts the perspective and tone, which supports the entity recognition. Shorter inputs with fewer names simplify the victim and witness detection, while a direct and concrete retelling of the events of the crime helps with crime type classification. Observed patterns indicate that summarization not only shortens the input, but also changes the reports into a representation that is easier for the models to process.

Error Analysis. Typical mistakes made by most models mainly centered around challenges with naming conventions and the distinction between witnesses and victims. A recurring difficulty for all models was handling the multiple names by which a single person could be referred. For example, witnesses often used a victim's first name, while the commission referred to the same person by their family name. These connections were usually clear to human readers through context or by an early mention of the full name.

Cultural naming conventions added further complexity. Many people in the dataset followed the South African tradition of having two first names. The so-called home name often comes from Zulu or Xhosa tradition and is used by the close family, while the school name is used by the public (Suzman, 1994; Neethling, 2008). This led to false positives, when the model correctly classified both names as victims, but did not recognize that they belonged to the same person. For example, the name "Kwinda Daphne Tshinane" would be listed both as "Kwinda Tshinane" and "Daphne" in the model output. However, if both names were mentioned at least once in direct succession, the models would usually avoid this error.

For the NER task, hallucinations were less common: Many false positives were not full hallucinations, but actual names that appeared in the reports, such as family members or commissioners. This suggests that the models were generally able to identify names correctly, but struggled to classify them into the correct category.

7 Conclusion

This study showed that LLMs outperformed other models in summarization and were able to produce at least decent results in a fine-grained NER task that non-LLM models are not able to solve without further fine-tuning. The best performing open-source LLMs were competitive with GPT. However, the optimization of LLMs towards coherence and instruction following comes at the expense of accuracy. For summarization, this leads to overly long outputs with hallucinations and generic filler sentences. In NER, this leads models to avoid empty outputs and to label almost any detected name as either a victim or a witness, resulting in high recall but low precision.

Unlike general-purpose NLP tasks, mistakes in this context can have ethical and historical consequences. Omitting a victim's name or falsely attributing responsibility for a crime can alter the record and lead to wrongful accusations. LLM summaries may also be easier to read and more coherent than the witness reports, which might mislead readers due to higher credibility despite potential hallucinations. While LLMs offer new opportunities to make historical texts more accessible and easier to analyze, at the current state, the application of LLMs to sensitive historical datasets should always be accompanied by careful human oversight and transparency about model limitations. Pipeline approaches with models specialized in the task, combined with the flexibility of LLMs, may improve the results in future work.

8 Limitations

One essential limitation of these results, particularly for summarization, is the observation that numerical evaluation methods are not entirely reliable when assessing the quality of the results. While BERTScore and ROUGE score capture summary quality relatively well, they do not always align

with human judgment and are dependent on the quality of the reference summary. Even embedding-based BERTScore occasionally fails when it comes to more complex semantics (Zhang et al., 2020b; Fabbri et al., 2021). We therefore interpret the results primarily in terms of overall trends rather than absolute values and employ multiple evaluation metrics to mitigate metric-specific biases. Reference-free metrics, too, cannot reliably indicate the quality of a summary. However, when observed in combination with reference-based metrics, they can give insight into the model's strengths and weaknesses, such as the repetitiveness of T5 or the verbosity of Mixtral.

Although the ROUGE scores reported here may seem low, they are relatively good results considering the task. Paraphrasing is part of the process of abstractive summarization, and the wording of two summaries will inevitably differ. ROUGE-scores for this task are usually in a similar range to the results achieved here (Fabbri et al., 2021; Lam et al., 2022).

The raw witness transcripts published by the TRC were initially recorded on tape and later transcribed without modification. As a result, they preserve all the imperfections of natural speech, including stuttering, repetition, unfinished sentences, filler words, and word mix-ups. The emotional nature of the testimonies often intensified this. Additional errors arise from the recording process itself, including microphone malfunctions and losses of several seconds to minutes when tapes were flipped. The transcription process introduced additional errors, such as missing speaker labels, misspelled names, and arbitrary line breaks.

The cultural and historical background of the TRC-dataset was certainly a challenge for most models. Understanding these reports requires not only good reasoning abilities but also detailed knowledge about apartheid and South African history, which the models seem to lack. LLMs are known to show a bias against non-Western cultures (Tao et al., 2024); it is therefore likely that an underrepresentation of South African historical and cultural material in the training corpora of LLMs has contributed to the models' shortcomings. If LLMs are used in practice for tasks on cultural datasets, special training methods are needed to mitigate this bias (Gallegos et al., 2024; Li et al., 2024).

The results presented here open up possibilities for future studies and improvements. Firstly, establishing a strong non-LLM baseline on the TRC dataset would allow clearer benchmarking. This would require more human-annotated data and the training of specialized neural models. Secondly, given that further instruction tuning did not improve NER performance, it may be worthwhile to evaluate non-instruction-tuned LLMs. Another possible improvement could be achieved with more complex pipeline approaches, where LLMs identify potential entities and a smaller classifier assigns roles. Alternative formulations of the NER task, such as tagging within the text rather than direct extraction, may also prove fruitful, although this would require networks with very large context windows. Thirdly, expanding experiments to other datasets could help disentangle whether the main challenges stem from the dialogue format, South African cultural references, or text length. The use of different yet similar annotated datasets may also enable fine-tuned LLMs to learn domain-specific language.

9 Ethical Considerations

This research only relies on information that is publicly available on the TRC website. Witnesses' names and personal information are not disclosed beyond the information that has been published by the TRC. Testimonies from mass atrocity proceedings carry particular psychological weight for those who engage with them (Schirmer, 2024). Human annotators were informed of and aware of the potentially violent content before the annotation process, with the ability to decline annotation at any time. They were given the chance to discuss any distressing material encountered during annotation and provided with a guide designed to aid in identifying changes in cognition and minimizing emotional risks associated with the annotation process (Kennedy et al., 2022).

Beyond annotation, ethical considerations should also address the potential impact of hallucinations. Here, a fabricated name in a victim or witness list could wrongfully tie falsely link someone to a human rights violation, or a could erase a victim from the record. LLM-generated summaries should therefore always be read alongside the original testimony, not instead of it.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael

Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Nikolich Alexandr, Osliakova Irina, Kudinova Tatyana, Kappusheva Inessa, and Puchkova Arina. 2021. Fine-tuning gpt-3 for russian text summarization. In *Data Science and Intelligent Systems*, pages 748–757, Cham. Springer International Publishing.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition. *Preprint*, arXiv:2305.15444.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. Lextime: A benchmark for temporal ordering of legal events. *Preprint*, arXiv:2506.04041.

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India. NLP Association of India (NLPAI).

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *Preprint*, arXiv:1807.04905.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. page 97–107, Bochum, Germany. Bochumer Linguistische Arbeitsberichte.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *Preprint*, arXiv:2402.01981.

David Gilbert and Georgina Heydon. 2021. Translated transcripts from covert recordings used for evidence in court: Issues of reliability. *Frontiers in Communication*, 6:779227.

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. 2024. Leveraging open large language models for historical named entity recognition. In *Linking Theory and Practice of Digital Libraries*, pages 379–395, Cham. Springer Nature Switzerland.

James Gung and Jugal Kalita. 2012. Summarization of historical articles using temporal event clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 631–635.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334.

Ridewaan Hanslo. 2022. Deep learning transformer architecture for named entity recognition on low resourced languages: State of the art results. In *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, volume 30 of *FedCSIS 2022*, page 53–60. IEEE.

Lukas Hilgert, Danni Liu, and Jan Niehues. 2024. Evaluating and training long-context large language models for question answering on scientific papers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 220–236, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. 2024. Efficient solutions for an intriguing failure of llms: Long context window does not mean llms can analyze long sequences flawlessly. *Preprint*, arXiv:2408.01866.

Natthanaphop Isaradech, Andrea Riedel, Wachiranun Sirikul, Markus Kreuzthaler, and Stefan Schulz. 2024. Zero-and few-shot named entity recognition and text expansion in medication prescriptions using chatgpt. *arXiv preprint arXiv:2409.17683*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Guochao Jiang, Zepeng Ding, Yuchen Shi, and Deqing Yang. 2024b. P-icl: Point in-context learning for named entity recognition with large language models. *Preprint*, arXiv:2405.04960.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and 1 others. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Renana Keydar. 2020. Listening from afar: An algorithmic analysis of testimonies from the international criminal courts. *University of Illinois Journal of Law, Technology & Policy*, pages 55–84.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. *Preprint*, arXiv:2210.02406.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arXiv:2501.08457.

Khang Lam, Tuong Do, Nguyet-Hue Pham, and Jugal Kalita. 2022. *Vietnamese Text Summarization Based on Neural Network Models*, pages 85–96.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Andrés Leiva-Araos, Bady Gana, Héctor Allende-Cid, José García, and Manob Jyoti Saikia. 2025. Large scale summarization using ensemble prompts and in context learning approaches. *Scientific Reports*, 15(1):10259.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Park: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sengjie Liu and Christopher G. Healey. 2023. Abstractive summarization of large document collections using gpt. *Preprint*, arXiv:2310.05690.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *Preprint*, arXiv:2212.07981.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in Token-Level clinical named entity recognition. *AMIA Annu Symp Proc*, 2024:748–757.

Hanjun Luo, Yingbin Jin, Xinfeng Li, Xuecheng Liu, Ruizhe Chen, Tong Shang, Kun Wang, Qingsong Wen, and Zuozhu Liu. 2025. Dynamicner: A dynamic, multilingual, and fine-grained dataset for llm-based named entity recognition. *Preprint*, arXiv:2409.11022.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural OCR post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4251–4257. International Joint Conferences on Artificial Intelligence Organization.

Keerthana Murugaraj, Salima Lamsiyah, and Christoph Schommer. 2025. Abstractive summarization of historical documents: A new dataset and novel method using a domain-specific pretrained model. *IEEE Access*, 13:10918–10932.

Bertie Neethling. 2008. Xhosa first names: A dual identity in harmony or in conflict? *Names: A Journal of Onomastics*, 56(1):32–38.

OpenAI. 2024. Gpt-4o technical report. *arXiv preprint arXiv:2405.19102*.

Sairaj Pokale, Karan Taware, Gavin Fernandes, Sakshi Kangane, Parth Bhosale, and Laxmi Bewoor. 2023. Text summarization: Gpt perspective. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–7.

Maciej P. Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Are large language models temporally grounded? *Preprint*, arXiv:2311.08398.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Swati Rallapalli, Shannon Gallagher, Andrew O. Mellinger, Jasmine Ratchford, Anusha Sinha, Tyler Brooks, William R. Nichols, Nick Winski, and Bryan Brown. 2025. Fine-tuning llms for report summarization: Analysis on supervised and unsupervised data. *Preprint*, arXiv:2503.10676.

Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. *Preprint*, arXiv:2310.10570.

D Sasikala, R Sudarshan, and S. Sivasathya. 2024. Harnessing llms for medical insights:ner extraction from summarized medical text. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Miriam Schirmer, Christian Brechenmacher, Endrit Jashari, and Jürgen Pfeffer. 2024a. Gentrac: A tool for tracing trauma in genocide and mass atrocity court transcripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7666–7671.

Miriam Schirmer, Udo Kruschwitz, and Gregor Donabauer. 2022. A new dataset for topic-based paragraph classification in genocide-related court transcripts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4504–4512.

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, and David Jurgens. 2024b. The language of trauma: Modeling traumatic event descriptions across domains with explainable ai. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13224–13242.

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, and Jürgen Pfeffer. 2023. Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Miriam Schirmer, Jürgen Pfeffer, and Sven Hilbert. 2025. Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, 19(1):83–102.

Miriam Katharina Doris Schirmer. 2024. *Natural language processing for violence studies: Investigating trauma and online aggression*. Ph.D. thesis, Technische Universität München.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Susan M. Suzman. 1994. Names as pointers: Zulu personal naming practices. *Language in Society*, 23(2):253–272.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. 2021. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 172–179, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Truth and Reconciliation Comission. 1998. Truth and reconciliation commission of south africa report.

Crina Tudor, Beata Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering LLMs. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, Volume 5 - 2018.

Le Xiao, Yunfei Xu, and Jing Zhao. 2024. Llm-der:a named entity recognition method based on large language models for chinese coal chemical domain. *Preprint*, arXiv:2409.10077.

Yuanzhen Xie, Xinzhou Jin, Tao Xie, Matrixmxlin Matrixmxlin, Liang Chen, Chenyun Yu, Cheng Lei, Chengxiang Zhuo, Bo Hu, and Zang Li. 2024. Decomposition for enhancing attention: Improving LLM-based text-to-SQL through workflow paradigm. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10796–10816, Bangkok, Thailand. Association for Computational Linguistics.

Haisong Zhang, Lemao Liu, Haiyun Jiang, Yangming Li, Enbo Zhao, Kun Xu, Linfeng Song, Suncong Zheng, Botong Zhou, Jianchen Zhu, Xiao Feng, Tao Chen, Tao Yang, Dong Yu, Feng Zhang, Zhanhui Kang, and Shuming Shi. 2020a. Texsmart: A text understanding system for fine-grained ner and enhanced semantic analysis. *Preprint*, arXiv:2012.15639.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024a. Cross-lingual cross-temporal summarization: Dataset, models, evaluation. *Preprint*, arXiv:2306.12916.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model. *Preprint*, arXiv:2508.18090.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2024b. Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models. *Preprint*, arXiv:2411.07858.

A Appendix

A.1 Annotation

Expert annotators followed the following instructions (summarized, the original guidelines can be found in this paper's code repository):

Summaries should be short, contain only factual information, and cover all important facts that are mentioned. The focus lies on the course of events surrounding the human rights violation, rather than on procedural matters or possible outcomes of the hearing. It is assumed that the report is accurate and complete. The purpose of the summary is to make the reports easier to search and filter, for example, for a specific type of crime or situation. Therefore, key information, such as the type of crime, names of the victims and witnesses, place, and time, must be mentioned. Additionally, the TRC typically considers information about the victim's political activities, the perpetrator's name and motive, and the long-term consequences of the crime to be important.

Annotation for NER was a complex task, as the process of the TRC hearings was flexible and witnesses could change without proper announcement. Each witness could also talk about multiple incidents involving different victims or a single incident with multiple victims. For the purpose of annotation, any person who is testifying in front of the Commission is considered a witness, not including eye-witnesses that are mentioned, but do not appear in front of the Commission. A victim is any person mentioned in the text who has been harmed, mentally or physically, as long as their case falls roughly into the jurisdiction of the TRC. Because this annotation is intended for NER, only people whose names are known can be included, even though other unnamed victims might be mentioned in the text.

A.2 Methods

A.2.1 Summarization Prompt

```
You are an academic assistant who summarizes historical witness reports. Your summaries should be concise, factual, and neutral in tone.

Write a short summary of the given witness report. Describe the course of events of the alleged crime and mention the names of the victims and witnesses.
```

A.2.2 NER Prompt

```
You are an academic assistant analyzing historical witness reports about human rights violations.

The given text is a witness statement about human rights violations during the South African apartheid.

Your task is to extract and classify the most important names of people that you find in the text.
Victims: The person that the witness is talking about and who was harmed
Witnesses: The person who is telling their story in front of the commission

Give your answers in JSON formatting with the fields "victims" and "witnesses". If the name does not appear in the text, write "NULL".
```

A.2.3 Violation Classification Prompt

```
You are an academic assistant analyzing historical witness reports about human rights violations.

You are given a witness report. Your task is to identify which type of crime the witness is talking about.

Choose a small number of classes from the following list. Only choose the classes when you are sure that they apply to the witness report. Only use crimes that are defined in the list and keep the exact same names.

Possible crime types:
* **killing**: someone was killed (intentionally or unintentionally)
* **torture**: pain inflicted
```

```
by state officials for coercion, information gathering or discrimination
* **detention**: victim was detained unlawfully or under unreasonably poor conditions
* **serious injuries**: victim has lasting health problems from the incident
* **assassination**: targeted, intentional killing for political reasons (usually by special forces soldiers)
* **shooting**: the victim was shot and hurt (may be deadly or not)
* **disappearance**: the victim has disappeared and it is currently unknown if they are dead or alive
* **context statement**: no specific incident, only general context
* **assault**: someone was assaulted and injured
* **harassment**: only use this if the harassment is severe (unlawful house search, death threats etc.) or if it is the main point of the hearing
* **bombing**: there was a bombing
* **destruction of property**: any form of vandalism that is not arson
* **arson**: someone set something on fire (but not human burning)
* **human burning**: a human was set on fire (deadly or not)
* **blackmail**
* **theft**

Output format: Comma-separated list of crime types.
```

A.2.4 Hyperparameter Tuning for Summarization

.

The only two models that were sensitive to their hyperparameters were Pegasus and TinyLlama. For Pegasus, increasing the number of beams for beam search and restricting the early stopping conditions resulted in summaries that were clearly more abstractive, contained fewer hallucinations, and were closer in length to the reference summaries.

For TinyLlama, the default parameters generated a single sentence with almost no relevance, which was repeated multiple times. Improved beam search parameters, combined with strong penalties on repetition, forced the model to produce coherent summaries. A low temperature setting reduced hallucinations at the cost of abstraction, resulting in improved numerical scores on the validation set. The exact parameters can be found in table 4.

A.2.5 Additional Experiments for Summarization

Limited Input: The process of successively summarizing paragraphs of texts seems intuitive to humans. Still, for a neural model, these intermediate steps could pose a problem, as contextual information between the chunks is lost. Since the model does not retain memory of the whole text when creating the final summary, errors in the intermediate summaries are easily propagated. It therefore seems as if models with a long context window have a clear advantage over models with a shorter context window, as they do not have to split the text. To test this hypothesis, we conducted a second experiment in which the maximum input size was reduced to 1024 tokens for all models. In this experiment, larger models like Mistral could still benefit from their superior context understanding abilities, but had to deal with the potential drawbacks of splitting and rejoining the texts like the other models.

Limited Output: Another potential source of error in the summarization process is the possibility that the joined intermediate summaries are too long to be processed, causing the model to repeat the process recursively. This means that the model processes the summary more than once, which can have a good or bad impact on the summary quality – in the best case, the model successively removes less important information, in the worst case, errors and hallucinations are propagated and amplified each time the text is passed to the model. Limiting the length of the intermediate summaries would reduce the risk of a document being processed multiple times. While limiting the output works well for the smaller, summarization-tuned models, the LLMs lack an encoder and predict the summary token by token, which means they cannot be properly limited to a specific length. We resorted to

Table 4: Hyperparameters: T5 and BART use default parameters from the Huggingface summarization pipeline, Pegasus and TinyLlama were fine-tuned, and all other models use default parameters.

Parameter	T5	BART	Pegasus	TinyLlama	Others
early_stopping	False	True	True	True	False
num_beams	1	4	5	4	4
temperature	1.0	1.0	0.6	0.3	0.8
top_p	1.0	1.0	1.0	0.8	0.9
top_k	50	50	50	50	40
repetition_penalty	1.0	1.0	1.0	1.2	1.1
length_penalty	1.0	2.0	2.0	2.0	1.0
no_repeat_ngram_size	0	3	0	3	0

truncating the generated intermediate summaries to 100 tokens, which might cause some critical information to be lost. This third experiment will determine whether the loss of information outweighs the potential benefits from reduced recursion.

A.2.6 Additional Experiments for NER

LLM Combination: For NER, the combination of the results for each chunk is not as straightforward as for summarization. The naive approach of concatenating the lists for victims and witnesses, respectively, resulted in too many duplicates. Removing duplicates is complicated since the context is lost – for example, "Paul" and "Mr. Smith" may or may not refer to the same person. For most experiments, I removed duplicates by fuzzy matching with token set ratio, a similarity measure that is robust to missing words, such as middle names or initials. This method works for all cases in which the full name of a person was detected at least once; however, it is also error-prone, as for example "Paul Smith" would match "Mrs. Smith". In a separate experiment, the LLM was tasked with combining the results independently. While this relies on the LLM's understanding of the JSON format and is prone to hallucinations, it can capture some details that fuzzy matching cannot.

Pipeline Approach: The length of the TRC documents is one of the main challenges for the LLM. Most of the information given in the text is irrelevant for the task, and only very few tokens determine whether or not a word should be included in the name list. Summarizing the input before applying NER could improve the results, as it simplifies the task for the LLM. We tested this hypothesis using human-written reference summaries to determine if summarization in general improves the results. We also performed NER with the result summaries from the first task to assess whether an LLM can complete the task fully on its own. An overview of all experiments can be seen in Table 5.

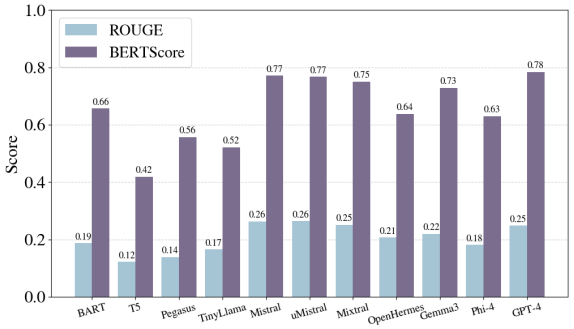


Figure 4: Summarization: Model Comparison with BERTScore and ROUGE-score.

A.3 Results

A.3.1 Additional Results for Summarization

Limited Input: When limiting the maximal input length to 1024, the setup for BART and Pegasus remained unchanged, as their context window is already only 1024 tokens long. For these models, the minimal changes between experiments are only due to the randomness of the model. For most of the other models, limiting the input did result in slightly lower summary quality, especially for Mistral and T5. For OpenHermes, however, limiting the input length improved the quality of the summary. Results like this can occur when a model's short context summarization abilities are significantly better than its long context summarization abilities, thereby overcoming the disadvantages introduced by chunk processing. Overall, the differences are small with a relatively high standard deviation. It can therefore be concluded that the advantages and disadvantages of chunk processing are balanced, confirming the effectiveness of this method in dealing with long inputs in summariza-

Table 5: Overview of different experiments.

Experiment	Task	Description
Main experiment	All	main results
Limited Input	All	input max. 1024 tokens, better model comparison
Limited Output	Sum.	intermediate output max. 100 tokens, reduced risk of recursion
LLM Combination	NER	JSON results combined by LLM
Reference Pipeline	NER	NER on reference summary
Pipeline Approach	NER	NER on LLM-generated summary

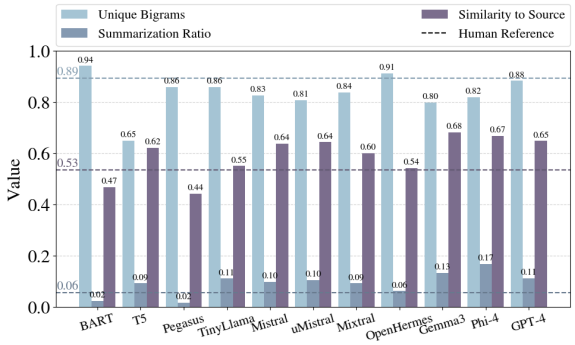


Figure 5: Summarization: Model Comparison with reference-free metrics.

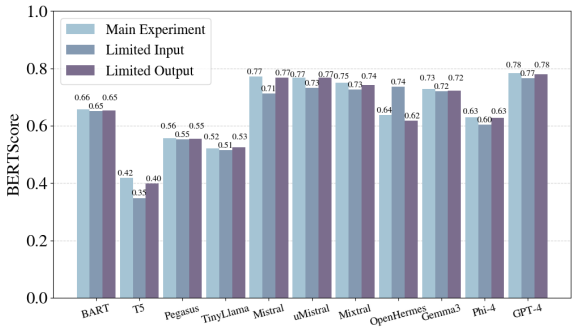


Figure 6: Summarization: BERTScore Results for the main experiment in comparison to the results with limited input and output.

tion.

Limited Output: When limiting the maximum output length of the intermediate summaries to 100 tokens, the models once again demonstrate surprising robustness. The performance decrease introduced by truncating intermediate summaries and potentially losing important information is minimal. The overall small changes in this experiment show that multiple passes over a long document are not necessarily a disadvantage, and that small truncations in the intermediate summaries do not influence the summary quality much.

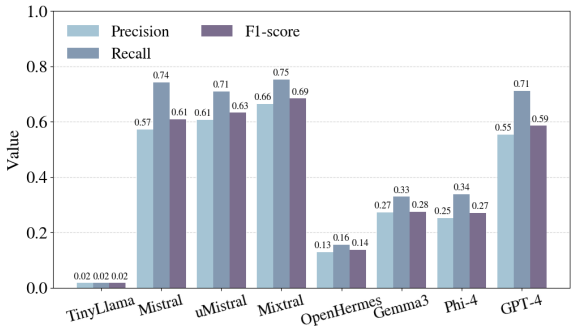


Figure 7: NER: Precision, recall and F1-score for each model, averaging over victims and witnesses.

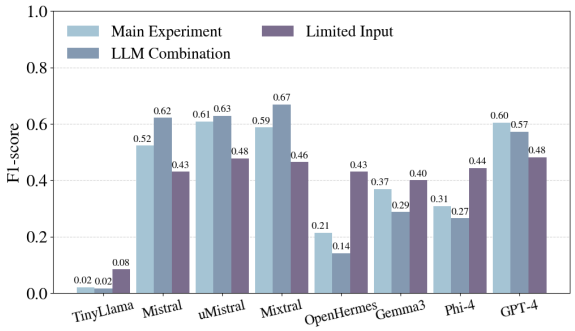


Figure 8: NER: Results for additional experiments with LLM Combination and Limited Input.

A.3.2 Additional Results for NER

Limited Input: For NER, the result quality is more dependent on the chunking process than for summarization. The Mistral models and Mixtral suffer from the limited input, while models that previously did not perform very well, like OpenHermes and Phi-4, improve. Even Gemma-3, with its long context, seems to improve when the information is presented in small chunks. For the better-performing models, the decrease in performance is probably due to the loss of context in between chunks. This problem is more prevalent for NER than for summarization, because for NER, there are very specific parts of the text that are relevant. If these parts are not contained in a chunk, the model

struggles with prediction and hallucinates. The robustness of summarization towards the chunking process is also one of the reasons for the success of the pipeline approach.

LLM Combination: For the better-performing models like Mistral and Mixtral, letting the model combine the intermediate results itself lead to improved results. Meanwhile, lower-performing models did not outperform the algorithmic approach of fuzzy matching.

A.4 Error Analysis

A.4.1 Summarization Reference

This is the reference summary referred to in Section 6.1.

> "Three witnesses testify about the death of Peter Mabilo, called Lathli or Luthli, who was shot and killed by police at his grandmother's house. The three witnesses are his mother, Mary Mabilo, his sister Florence Matlakala, and his grandmother Sheila Mabozo, who is also an eyewitness to the shooting. Peter Mabilo was a student and an active member of the ANC. He was arrested twice in 1985 and spent at first 25 days, then 8 months in prison. He was severely beaten by police before all charges were dropped, and he was released. After that, he fled the country without telling his family where he went. He returned in August 1987, when he lived with his grandmother and visited his mother, sister, and girlfriend for a day. The next day, police came to Mrs. Mabozo's house and ordered the family to come outside. Mr. Mabilo was hiding inside the house, but surrendered after the police threw teargas inside. He was then shot in the hallway of the house. Mrs. Mabozo recalls that the furniture was destroyed and that there were bullet holes in the walls. The house of Mr. Mabilo's mother and sister was searched afterwards, and the family was harassed by the police. Mr. Mabilo's funeral was interrupted by police, who only gave them 5 minutes to bury Mr. Mabilo. Mrs. Mabilo accuses a policeman named Terra Cacamba of killing her son."

A.4.2 Detailed Error Analysis for Summarization Example

In the following quotations of the model's summaries of this case, hallucinations are highlighted in red, highly irrelevant facts are highlighted in blue, and important facts are highlighted in green.

> 1.:ADV POTGIETER: Okay, then we'll stand down - stand your evidence down for the moment and I think we'll - will take the evidence of your daughter, of Florence and we could come back to you if there's something you want to add to what you've said..ADV POTGIETER: Okay, then we'll stand down - stand your evidence down for the moment and I think we'll - will take the evidence of your daughter, of Florence and we could come back to you if there's something you want to add to what you've said.
> – T5, BERTScore 0.1896, ROUGE 0.0777

This summary is highly extractive, as were most summaries that T5 created. This is likely due to the training on news articles, where not much rephrasing is necessary for summarization. In some cases, the extractive summaries contained important information, although in this example, it merely quotes a moment from the middle of the testimony where procedural matters are discussed.

> Mary Mabilo is the mother of Lathli Mabilo .<n>Lathli was shot dead in 1987 at the home of Ms Sheila Mabozo .<n>Ms Mabilo says five of the accused were arrested in 1985 and spent 25 days in jail .<n>Misty Mabozo was found dead at her home in Ipopeng .<n>She was allegedly murdered by Terra, who is on trial for her murder .<n>Mary Mabilo told the court how he was told Terra had killed his son .
> – Pegasus, BERTScore 0.6502, ROUGE 0.1825

Pegasus' summaries are much more abstractive. Each sentence can still be traced back to a sentence in the original text, but the model was able to abstract from the dialogue format of the original text by adding names and adapting the pronouns. The information in the text is mostly correct, but not always important and coherent. For example, the sentence about the five accused who were arrested

refers to the victim's first arrest before his exile, which is not clear from the summary.

In addition, Pegasus' summaries are usually very short, and while they include important details in the summary, they are often incomprehensible due to missing context or too little paraphrasing. In this case, the model also switched the victim's name to "Misty Mabilo", a name that appears in the original text only once in reference to one of the witnesses. This is an example of a minor error that can significantly misrepresent the witness report, rendering the summary inaccurate and useless. Most models included small but significant errors like this in their summaries.

> Mary Mabilo, Florence Matlakala and Sheila Mabozo give evidence. All three are from Galishewe Village in Mpumalanga province. Mary tells how her son, Lathli, was shot dead in 1987. Florence tells how five of her children were arrested and held for eight months. Sheila says she was blindfolded and had sharp instruments put into her ears.
> – BART, BERTScore 0.6508, ROUGE 0.1202

While BART sometimes has issues with over-extractiveness, too, in this example, it abstracts reasonably well. The information is mostly correct, but important details of the main incident are missing. In contrast to the reference summary, this summary follows the temporal events of the hearing (one witness after the other), instead of combining the testimonies into one description of the course of events. The inability to understand the order of events was a common factor among all models. In the case of the non-LLM models, this can be attributed to the chunking process, which makes it especially challenging to maintain context over long texts and hinders the model's ability to, for example, recognize that witnesses are describing the same incident. While this summary is objectively better than the Pegasus summary, it has the same BERTScore and a lower ROUGE score due to more abstraction.

> The Commission has concluded that there is insufficient evidence to charge the accused with the crime of murder. The evidence presented by the prosecution does not meet the standard of proof required for a conviction of murder, and there is no evidence of a motive or premeditation in relation to the crime. The investigation into the incident is ongoing, and the Commission will continue to monitor the progress of the investigation to determine whether any further charges should be brought against the accused. This entry was posted in Uncategorized and tagged murder, South Africa. Bookmark the permalink. Post a comment or leave a trackback: Trackback URL.
> – TinyLlama, BERTScore 0.4590, ROUGE 0.1342

TinyLlama occasionally provides relatively good summaries. They mostly suffered from the omission of important details, which were replaced by well-formulated, but generic and meaningless sentences. This summary is an extreme example, where the model's entire answer was comprised of hallucinated or meaningless sentences. No trace of the original text remains, except for the fact that there was a murder. This problem of over-abstraction into meaninglessness is the exact opposite of the issues that the encoder-decoder models struggled with, and a result of the architecture. Without a decoder, the LLM cannot read the complete text and then write a full summary; instead, it simply predicts the next most likely token. Without a very high number of parameters and sufficient training data and instruction fine-tuning, this results in many generic sentences.

> The witnesses, Mary Mabilo, Florence Matlakala, and Sheila Mabozo, testified about the shooting of Mary Mabilo's son, Lathli Mabilo, in 1987. Lathli had been missing for more than a year before his death. He was wanted by the police and was frequently harassed. He was killed at Sheila Mabozo's house in Ipopeng, South Africa. The police interrogated Mary Mabilo about the incident, and two women came to her house and took her furniture without her consent. Ms Matlakala testified that her brother, Lathli, returned home in 1987 but was killed on the same day. The police harassed the family before and during the funeral. Ms Mabozo, Lathli's grandmother, was interrogated by the police, blindfolded, and her ears were damaged with sharp objects. She was held in the Barkley cells

for two days. Luthli, who had been living with Ms Mabozo for a month, was shot and killed by the police outside their house. The police had previously shot tear gas into the house. Ms Mabilo, Luthli's mother, also testified and asked for Terra, a policeman, to be called to explain why he killed her son. Ms Mabilo is a poor woman who
– uMistral, BERTScore 0.7854, ROUGE 0.3012

This is a subjectively good summary of the incident. The information is correct. Small details that could be considered significant, such as the victim's previous arrests, are omitted. The order of events is accurate, but described in a somewhat confusing manner, and the structure of the original document (one witness after the other) is still recognizable.

[...] Ms Mabozo, Luthli's grandmother, provided additional details about the shooting itself. She stated that after the police took Luthli's father away, they were told to take the children outside. Luthli was inside the house when the police shot him. The family found the house damaged and full of bullet holes when they returned. [...]
In conclusion, the Commission expressed its sympathy for the family and vowed to investigate the circumstances surrounding Luthli's death [...].
– Mistral, BERTScore 0.7143, ROUGE 0.2470

In this example, the difference between Mistral and uMistral is evident to a human reader, which was not always the case. In contrast to Mistral, uMistral produced more hallucinations. The last, highly irrelevant paragraph can be seen as an example of a generic statement that was generated because the system prompt encourages the model to be polite and eloquent. Even though both models were run with the same hyperparameters, Mistral seems to abstract more.

The witness report contains statements from three witnesses, Mary Mabilo, Florence Matlakala, and Sheila Mabozo, regarding the shooting and death of Mary's son, Lathli Mabilo, in Kimberley, South Africa in 1987. Mary, Lathli's mother, states that Lathli was involved in anti-apartheid political activities and had been arrested multiple times. She describes how Lathli had been missing for a year and seven months before she saw him again for one day, and then he was killed. Mary suspected that "those who were after him" were responsible for his death. Florence, Lathli's sister, testified that Lathli had returned to Kimberley to deliver a message to someone and that she took him to meet his girlfriend before he planned to leave town again. She describes how the family was harassed by the police before and after Lathli's funeral, which was attended by his organization and family. Sheila, Lathli's grandmother, testifies that Lathli was shot outside her home while he had his hands up, and that the police shot him multiple times and then entered the house and continued to shoot him. The witnesses request that the Commission investigate the circumstances surrounding Lathli's death and the identity of the police officer "Terra" who they believe was involved in his killing.
– Mixtral, BERTScore 0.7092, ROUGE 0.2653

Even though this summary has a lower BERTScore than both Mistral models, it is objectively the best summary of the case. Although it remains structurally similar to the original document, the temporal order of events is now comprehensible. It is a long summary, and some of the details may not be crucial, but the information provided is accurate and covers nearly all the important facts. The lower scoring of this summary means that it is more different from the reference summary without necessarily being worse. This indicates that the evaluation metrics, although providing a good overview when averaging over multiple texts, cannot fully assess the quality of a single summary. Very interestingly, Mixtral tends to capture details that other models miss, in this case, for example, the legally important fact that the victim had his hands up when he was shot.

Perspectives – Interactive Document Clustering for Qualitative Data Analysis

Tim Fischer, Chris Biemann
Language Technology Group, Department of Informatics, University of Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

This paper introduces *Perspectives*, an interactive extension of a qualitative data analysis tool suite developed at our university, designed to empower Digital Humanities (DH) scholars to explore and organize large, unstructured document collections. *Perspectives* implements a flexible, aspect-focused document clustering pipeline with human-in-the-loop refinement capabilities. We showcase how this process can be initially steered by defining analytical lenses through document rewriting prompts and instruction-based embeddings, and further aligned with user intent through tools for refining clusters and mechanisms for fine-tuning the embedding model. The demonstration highlights a typical workflow, illustrating how DH researchers can leverage *Perspectives*'s interactive document map to uncover topics, sentiments, or other relevant categories, thereby gaining insights and preparing their data for subsequent in-depth analysis.

1 Introduction

The discourse analysis tool suite[1] (DATS), a platform developed at our university, supports Digital Humanities (DH) researchers in data analysis by providing access to computational methods for large-scale, multimodal data. Key functionalities include pre-processing, qualitative and quantitative analysis, and document retrieval (including keyword, metadata, and semantic search). However, DATS users often struggle to explore and organize extensive corpora. Keyword-based or semantic search functionalities are often insufficient and incompatible with a hermeneutic approach to conceptualization, and DATS lacks functionality for systematic structuring and a holistic overview. Organizing documents, however, is vital for qualitative analysis and advanced DATS features, such as timeline analysis.

[1] https://github.com/uhh-lt/dats

Document clustering offers a promising solution, particularly for discovering themes within a corpus, a key interest of DH scholars. Modern clustering-based topic modeling (Grootendorst, 2022; Angelov and Inkpen, 2024; Reuter et al., 2024), which leverages pre-trained language model embeddings, has gained considerable traction not only because it offers advantages in simplicity, modularity, and performance, but also because it presents interesting opportunities for steerability and interactivity.

Unsupervised clustering methods can present a *take it or leave it* scenario for non-experts if the outputs do not align with their intent, rendering fully automatic methods ineffective. For DH scholars, who often approach corpora with specific research questions in mind, the ability to interactively manipulate clustering is of great interest. Such challenges are addressed by the Interactive Topic Modeling (ITM) community, which aims to achieve a controllable, collaborative human-machine process.

To address this gap, we introduce *Perspectives*, an interactive document clustering extension for DATS. Inspired by ITM and built on clustering-based neural topic modeling, we aim to make document clustering more accessible for DH scholars. The system is centered around an interactive document map to support exploratory workflows and implements a clustering pipeline that can be steered and refined in multiple ways. While offering visualization and analysis tools, its primary goal within DATS is to help users organize their collections into tagged datasets, ready for use with other analytical tools. The contributions of this paper are:

1. We introduce a flexible, aspect-focused clustering pipeline that combines (a) LLM-driven document rewriting to emphasize user-defined aspects, (b) instruction-steered embeddings to generate aspect-oriented document representations, and (c) few-shot fine-tuning of the embedding model to further align the representations with user intent.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 411–422
July 4, 2026 ©2026 Association for Computational Linguistics

2. We evaluate the pipeline on multiple datasets, showing the effectiveness of aspect-oriented representations, document rewriting, and fine-tuning.

3. We present *Perspectives*, a human-in-the-loop document (HITL) clustering system that integrates the pipeline and provides an interactive 2D map for users to explore, validate, and directly manipulate clusters with established refinement operations.

2 Related Work

The task of identifying topics in texts has evolved from traditional probabilistic models, such as Latent Dirichlet Allocation (Blei et al., 2003). Recent approaches, termed clustering-based neural topic modeling or topic discovery (Wu et al., 2024), build on pre-trained language model embeddings. Top2Vec (Angelov, 2020) uses Doc2Vec (Le and Mikolov, 2014) embeddings, reduces their dimensionality with UMAP, clusters them using HDBSCAN, and identifies topics by finding words closest to cluster centroids. BERTopic (Grootendorst, 2022) follows a similar pipeline but employs a class-based TF-IDF mechanism to extract descriptive words for each identified cluster. Contextual Top2Vec (Angelov and Inkpen, 2024) refines this by representing documents with multiple segment vectors for fine-grained topic segmentation.

The topic modeling community has long recognized the need to make topic modeling more user-centric. Seed wordlists (Jagarlamudi et al., 2012), seed documents (Grootendorst, 2022), or seed sentences (Fang et al., 2023) can guide initial topic formation. Further, interactive topic modeling systems address alignment with user intent by providing operations such as merging, splitting, or deleting topics, as well as adding or removing documents/words from specific topics (Fang et al., 2023). Seelman et al. (2024) proposes to update topic models by embedding user-assigned word labels to topics. GPTopic (Reuter et al., 2024) leverages LLMs to build an accessible chat interface to explore and refine topics through natural language.

The field of interactive (semi-supervised) clustering also acknowledges the value of integrating expert knowledge. While prior methods, such as hand-picked seed points (Basu et al., 2002) or pairwise constraints (Basu et al., 2002), offered control, they required extensive user feedback. Viswanathan et al. (2024) demonstrated that LLMs can guide semi-supervised clustering with minimal feedback by generating and encoding task-specific keyphrases for document representation.

Several systems facilitate the interactive exploration of document collections or topics. Nomic Atlas[2] is a notable commercial platform that offers high-performance 2D scatter-plot visualizations of document embeddings, enabling users to explore automatically extracted hierarchical topic clusters and interact via filtering and a chat assistant. Related academic systems innovate with HITL focus, such as direct integration into the labeling workflow (Seelman et al., 2024), tracking model changes (Fang et al., 2023), or RAG-based question-answering (Reuter et al., 2024).

Effective visualization helps explore and understand clusters. Standard techniques, such as top words, word clouds, document-cluster distributions, and 2D embedding projections (Wu et al., 2024), are all available in our extension.

While inspired by existing platforms, such as Nomic Atlas, *Perspectives* distinguishes itself by offering a user-centric HITL workflow tailored for DH: Users define the analytical lens before clustering and iteratively refine the outcomes afterward, with visualizations serving as a central component of this interactive sense-making process.

3 Envisioned Workflow

This section illustrates an idealized workflow using *Perspectives*, following Alice, a DH researcher investigating public discourse on climate change in Germany. Her goal is to understand major topics and the general sentiment towards climate change.

Alice uploads her corpus to DATS and opens the new *Perspectives* extension. To discover topics, she provides a *topic discovery* instruction to guide document embeddings (e.g., "Identify the topic"). The system generates an interactive 2D document map and a dashboard that provides a high-level summary, including clusters with LLM-generated names, keywords, and example documents, their relative frequencies, and a similarity matrix showing inter-cluster relationships. Alice navigates the map (see Figure 1), panning, zooming, and inspecting the visually distinct clusters. She identifies broad topics but notices that a theme like "Climate Protests" is missing.

Alice uses selection tools 1) to investigate documents of specific clusters. Linked statistic and information panels 2) provide aggregated details, such as frequent keywords and named entities. She

[2] atlas.nomic.ai

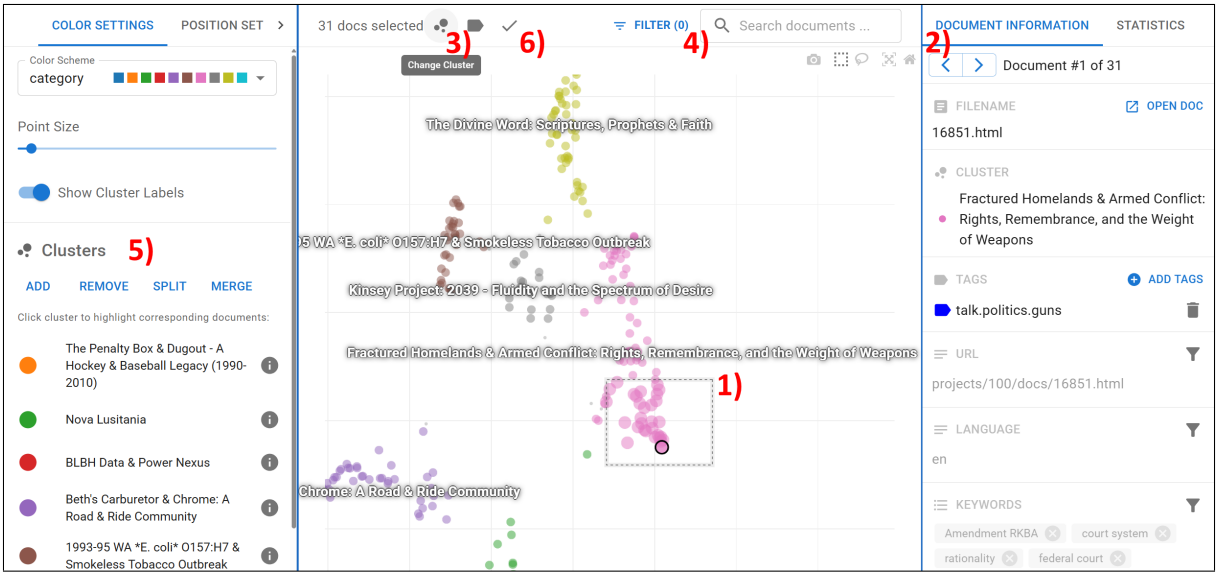


Figure 1: *Perspectives'* document map. Center: interactive scatter plot of documents colored by cluster. Hovering over a document previews its content. Left: settings & refinement operations, including the model fine-tuning operation. Right: statistics & information about selected documents. Top: toolbar with search & filtering. The UI is inspired by popular clustering interfaces such as Nomic Atlas, leveraging a familiar design to accelerate adoption.

spots misclassified articles, correcting them with the **Change Cluster** 3) function. To create her expected "Climate Protests" topic, she selects relevant documents (found via integrated search or filtering, 4) and uses the **Add cluster** function. If a topic is too broad, she uses the **Split Cluster** function to compute more granular sub-clusters. Conversely, she can **Merge Clusters** if several clusters represent a single cohesive theme 5).

With all topics identified, Alice improves their definitions by reviewing representative documents and marking them with the **Accept Cluster Assignment** function 6). Having provided minimal feedback, she triggers the **Refine Model** process, which fine-tunes the embedding model. The map reloads, showing more distinct and coherent clusters. Satisfied with her thematic map, Alice explores sentiment. She creates a new perspective, this time providing a *sentiment analysis* instruction and an optional document rewriting prompt (e.g., "Summarize sentiment towards climate change"). This generates a new map visualizing sentiment-based clusters, which she can explore and refine using the same interactive tools.

Finally, Alice exports her refined topics and sentiments from *Perspectives* as tags back into her main project, enabling further analysis tools within the DATS ecosystem.

4 Interactive Clustering

As illustrated by the workflow, *Perspectives* facilitates interactive document clustering, empowering users to guide analysis towards specific aspects and iteratively refine categorizations. This process begins with an aspect-focused clustering pipeline and continues through HITL refinement operations. The pipeline is illustrated in Figure 2.

We propose two strategies to define the analytical lens and initially steer the clustering process (green). First, we employ alternative document representations, in which an LLM rewrites documents based on a prompt, thereby modifying the content for clustering. This is a one-time process per perspective. Second, we use instruction-finetuned embedding models (Su et al., 2023) that generate aspect-oriented embeddings from an instruction, capturing specific semantic facets.

We continue with established steps (orange): reduce dimensionality with UMAP for improved clustering, then use HDBSCAN to group documents. Cluster representations include (1) keywords from c-TF-IDF, (2) LLM-generated names and descriptions seeded by those keywords, (3) the cluster centroid (mean of document embeddings), and (4) representative documents identified via cosine similarity to the centroid.

Next, document embeddings (reduced to 2D using UMAP) are presented as colored dots on a map. Here, users can iteratively post-process the clus-

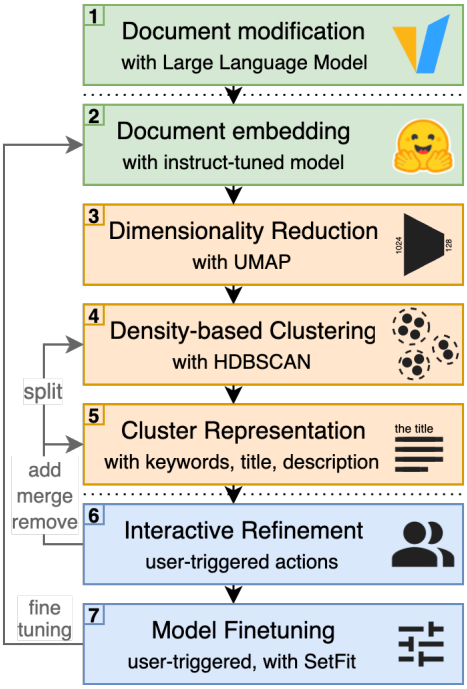


Figure 2: The proposed interactive clustering pipeline. The initial clustering process is guided by providing rewriting and embedding instructions (green) to focus the document representations on user-defined aspects. The established core pipeline (orange) identifies clusters and builds various textual representations. Users can post-process the clustering (blue) through refinement operations (grey), triggering some steps of the pipeline.

tering with several HITL refinement capabilities (blue; grey arrows). Whenever clusters are manipulated (i.e., documents are added or removed), their representations (keywords, name, description, centroid, and representative documents) are recomputed. The refinement operations are:

Change Cluster Documents are reassigned to an existing cluster or marked as an outlier.

Add Cluster (Documents) Selected documents form a new cluster.

Add Cluster (Text) Provided name and description are embedded to define the cluster centroid. Documents are automatically reassigned if their cosine similarity to the new centroid exceeds their similarity to their current cluster's centroid.

Merge Clusters Two clusters are combined into a single new cluster.

Remove Cluster The cluster is deleted. Its documents' embeddings are compared to the centroids of all remaining clusters using cosine similarity and assigned to the most similar one.

Split Cluster HDBSCAN is applied only to the documents within the cluster, yielding a set of more granular sub-clusters, and the original cluster is removed.

Additionally, we propose fine-tuning the model to further align the document representations with user intent. It is inspired by the findings of Thielmann et al. (2024). They demonstrated that with few (1-5) labeled documents per class, significant improvements in topic coherence can be achieved. Perspectives operationalizes this by allowing users to *accept cluster assignments*. These labeled examples are used to create a contrastive dataset to fine-tune the embedding model with the SetFit[3] library. This adjusts the embedding space, pulling similar documents closer and pushing dissimilar ones apart (details are explained in Section 5.4 and Appendix Section A.2). After fine-tuning, the entire dataset is re-embedded and re-clustered, resulting in an updated map that reflects the improved cluster structure.

HDBSCAN inherently identifies outlier documents, visualized as distinctively colored (small, grey) points on the map. They are considered during refinement operations. For instance, when a new cluster is added, outliers can be assigned to it if their similarity is sufficient.

5 Evaluation

We evaluate the effectiveness of our additions to the established clustering pipeline, which aims to be steerable towards certain aspects. We analyze initial guidance through (1) document rewrites and (2) instruction-tuned embeddings, as well as our proposed refinement operation of (3) model fine-tuning with few labeled data.

5.1 KNN Accuracy Metric

The goal of *Perspectives* is to assist users in obtaining a solid document categorization using the 2D document map. While topic modeling inspired our work, we refer to the extensive evaluation of Angelov and Inkpen (2024), which demonstrated the superiority of clustering-based topic modeling approaches over traditional and neural methods. Instead, we evaluate KNN accuracy, a standard metric for clustering interfaces. A good visualization should group documents with the same label and yield high classification accuracy within the visualization space (Pham and Le, 2021). A KNN classifier is trained to classify texts based on their 2D coordinates, and accuracy is reported. It mim-

[3] huggingface.co/docs/setfit

ics user behavior by deriving classes from dense clusters and can be seen as a proxy for end-to-end evaluation. Note that this metric operates directly on the two-dimensional embeddings and does not account for the HDBSCAN clustering.

5.2 Datasets

To evaluate the effectiveness of steering and guidance, we selected various datasets that cover text materials (news articles, books, social media, and songs) and tasks (topic discovery, stance detection, frame analysis, and bias detection) similar to real-world research projects conducted by our project partners who actively use DATS.

Amazon Product Reviews[4] consists of 2,500 sampled reviews with 5 *star* ratings and 15 *product* categories. Spotify Songtexts[5] comprises 5,000 sampled song lyrics with 10 *genres* and 5 *emotions*. 20 Newsgroups includes 3,600 sampled posts on 20 *topics*. GVFC (Liu et al., 2019) contains 1300 news article summaries with 9 *frames*. Blurbs (Remus et al., 2019) consists of 1,200 sampled German book *blurbs* with 8 main genres. Israel-Palestine (Ali et al., 2025) comprises 10,000 Reddit comments with 3 *stances*. News Bias[6] contains 8,500 sampled news articles with 3 political *biases*. For detailed descriptions, see Appendix Section A.1.

5.3 Models

For evaluation, we exclusively use open-licensed models that run locally. This approach ensures data privacy for our primary users (academic researchers) handling sensitive information. It also ensures the system is deployable in resource-constrained environments, accommodating users with limited computing resources. We use Gemma 3 (27B-instruct) (GemmaTeam, 2025) for LLM-driven document rewriting and multilingual-e5-large-instruct (500M) (Wang et al., 2024) for the instruction-tuned embedding model, selected due to its strong performance on the MTEB benchmark (Muennighoff et al., 2023). Further model details are provided in Appendix A.3.

5.4 Experiment Setup

First, we evaluate the impact of document rewriting and task-specific embeddings. We use straightforward prompts to generate summaries and keyphrases of the original documents (see Appendix Table 2). Then, we use short instructions to compute task-specific document embeddings (see Appendix Table 3). This results in 6 scenarios: original text (*text*), summary (*summ*), keyphrases (*keyp*), embedded with (+*inst*), and without instructions. Second, we evaluate the impact of fine-tuning the embedding model with few labeled examples (*2, 4, 8, 16-shot*), aiming to align the embedding space better with validated clusterings. We use SetFit to tune the embedding model for one epoch on increasing amounts of examples sampled randomly from the training data. The few labeled training examples are used to create a contrastive dataset with an equal number of positive and negative pairs. Instead of updating all model parameters, we train LORA adapters (Hu et al., 2022) with a cosine similarity loss to speed up training and reduce hardware requirements. The KNN accuracy metric is computed on document embeddings reduced to two dimensions with UMAP. Few-shot experiments are repeated 10 times with different training sets. The results are averaged across runs to mitigate fluctuations. All experiments were conducted on a single A100 GPU (80GB). All parameters, prompts, and training details are explained in Appendix A.2. The code to reproduce the results is available[7].

5.5 Results

Table 1 presents our evaluation results (see Appendix Figure 3 for a detailed breakdown. Consistent with previous work, we observe that embeddings generated with instructions (+*inst*) outperform those without in 8 out of 9 datasets we evaluated. More interestingly, we observe that document rewrites (keyphrases or summaries) prove beneficial in 6 out of 9 datasets: For example, *summary+inst* improves product categorization accuracy by 24.46 points, stance detection by 17.55 points, and bias detection by 5.33 points. For all other aspects, *text+inst* yields the best performance. Table 1 also highlights our best-performing few-shot results. These scores were achieved using the best-performing *zero-shot* configurations, for example *summary+inst* for product, frames, stance, and bias, *keyphrases+inst* for genre, and *keyphrases* for blurbs. While fine-tuning the embedding model with a few labeled examples consistently improved KNN accuracy, the gains were generally minor,

[4] kaggle.com/datasets/mexwell/amazon-reviews-multi
[5] kaggle.com/datasets/devdope/900k-spotify
[6] kaggle.com/datasets/articoder/news-bias-dataset
[7] https://github.com/uhh-lt/perspectives

	emotion	genre	product	stars	topic	frames	blurbs	stance	bias
text	45.10	27.20	36.72	47.54	71.17	59.78	73.84	51.19	47.20
+inst	**50.60**	27.50	55.76	**58.86**	**71.24**	65.00	73.25	48.21	48.51
keyphrases	49.15	34.85	54.64	46.34	65.70	60.34	**76.73**	61.63	50.76
+inst	49.51	**34.90**	60.64	49.68	65.40	58.31	76.66	63.14	50.69
summary	47.69	32.59	48.16	52.14	60.62	64.34	76.46	62.47	48.41
+inst	48.04	32.64	**61.18**	54.02	61.91	**65.91**	74.80	**68.74**	**52.53**
2-shot	50.70	36.63	62.98	59.71	71.36	65.10	76.79	70.17	52.72
4-shot	50.92	36.60	63.54	60.24	71.49	65.59	76.74	70.56	53.03
8-shot	51.64	36.66	63.62	60.09	71.58	66.92	76.92	71.06	53.09
16-shot	52.07	37.02	64.04	61.27	72.15	67.27	77.85	71.51	54.09

Table 1: Evaluation of document rewriting, instruction-based embeddings, and few-shot fine-tuning. Top: unsupervised (*zero-shot*) results. Bottom: few-shot results using the best configuration (e.g, *text+inst* for emotion & topic).

in the range of 2-3 points, when compared to the *zero-shot* setting. Thielmann et al. (2024), using 1-5 labeled examples per class, observed more significant improvements in topic coherence. Still, our results suggest that incorporating more labeled examples could further enhance performance.

Our findings guided the implementation of *Perspectives*: Task-specific instructions are mandatory for creating a perspective, as they consistently yield superior performance. The impact of document rewrites proved to be dataset-dependent. Therefore, providing a document-rewriting prompt is optional but recommended. To streamline the user experience, we integrated templates for common tasks, such as emotion, sentiment, and topic clustering.

6 Tool Suite Integration

The *Perspectives* frontend is built with React, uses Plotly.js[8] for efficient, interactive scatter plots, and MUI[9] for all other material design components. A FastAPI[10] (Python) backend provides the functionalities through the asynchronous task manager RQ[11]. We implement GPU-powered RQ workers, which ensure UI responsiveness during intensive computations, such as embedding, clustering, and model fine-tuning. We integrate our proposed clustering pipeline into DATS using the same parameters as in our experiments.

Fast response times are critical for interactive systems. We achieve high performance for our LLM-driven functionalities (like name generation and document rewriting) by combining vLLM and LiteLLM[12]. This setup leverages batch processing, streaming, and prefix caching for rapid inference. While the initial, optional document rewriting process may incur some latency depending on corpus size, it is an acceptable, one-time operation. Subsequent refinement operations are significantly faster, as their processing time scales with the number of clusters. Furthermore, the embedding model fine-tuning is also quick, requiring only one training epoch on a small set of labeled examples.

Using modern libraries such as Sentence Transformers and vLLM, switching models is straightforward. This ensures that DATS stays up to date with recent advances in NLP and provides users with state-of-the-art models.

7 Conclusion

This paper introduced *Perspectives*[13], an interactive extension to the open-source DATS platform designed to help DH scholars explore and organize large document collections. We demonstrated an aspect-focused clustering pipeline that leverages LLM-based document rewriting and instruction-steered embeddings. Central to *Perspectives* is HITL refinement, providing a suite of operations to manipulate clusters and fine-tune the embedding model with minimal input. This iterative approach, visualized on an interactive 2D map and supported by a dashboard, empowers users to create meaningful categorizations. By integrating these capabilities, *Perspectives* provides a powerful yet accessible tool that enables domain ex-

[8] https://plotly.com/
[9] https://mui.com/
[10] https://fastapi.tiangolo.com/
[11] https://python-rq.org/
[12] www.litellm.ai, docs.vllm.ai
[13] Demo: https://dats.ltdemos.informatik.uni-hamburg.de/

perts to leverage their knowledge effectively. It transforms document clustering into a collaborative sense-making activity, prepares data for further analysis within DATS, and provides a valuable tool for user-centered DH research. Future work will expand the UI's functionality, extend the scope of the extension to include multimodal data (images), and incorporate more labeled examples for few-shot scenarios.

Limitations

During initial testing of *Perspectives* with our project partners, a challenge in the model fine-tuning mechanism was identified: The **Refine Model** operation requires re-embedding and re-clustering of the entire dataset, resulting in the loss of spatial understanding and disruption of the user's mental map as the 2D projection recalculates. To mitigate this, we are currently developing a refinement history component. This feature, presented as an interactive timeline at the bottom of the map, stores and allows navigation between all previous map states (e.g., pre- and post-refinement views). We animate the transition between any two states, which helps users reconstruct their spatial understanding of the data. This feature also addresses a request from project partners, allowing them to revert undesirable refinement operations.

Our experimental evaluation focused on a single LLM for document rewriting and a single embedding model. Different choices of LLMs or embedding models could yield varying performance outcomes. Furthermore, the few-shot training approach is sensitive to the selection of training samples. To mitigate this high dependency, we conducted 10 runs with different sampled training sets. However, an optimal combination of training examples may not have been discovered. Our experiments also did not involve extensive prompt engineering for document rewriting, as we limited the prompts to summarization and keyphrase generation. Exploring other prompts could potentially lead to further improvements.

References

Hasin Jawad Ali, Ajwad Abrar, S. M. Hozaifa Hossain, and M. Firoz Mridha. 2025. Social media polarization during conflict: Insights from an ideological stance dataset on Israel-Palestine Reddit comments.

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics.

Dimo Angelov and Diana Inkpen. 2024. Topic Modeling: Contextual Token Embeddings Are All You Need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, FL, USA. ACL.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2002. Semi-supervised Clustering by Seeding. In *Proceedings of the Nineteenth International Conference on Machine Learning*, page 27–34, San Francisco, CA, USA. Morgan Kaufmann Publishers.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. ACL.

Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. A User-Centered, Interactive, Human-in-the-Loop Topic Modelling System. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522, Dubrovnik, Croatia. ACL.

GemmaTeam. 2025. Gemma 3 Technical Report.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with c-TF-IDF.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, Online.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France. ACL.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, page 1188–1196, Beijing, China. JMLR.org.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. ACL.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. ACL.

Dang Pham and Tuan M. V. Le. 2021. Neural Topic Models for Hierarchical Topic Detection and Visualization. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 35–51, Cham. Springer International Publishing.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. GermEval 2019 Task 1: Hierarchical Classification of Blurbs. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Arik Reuter, Anton Thielmann, Christoph Weisser, Sebastian Fischer, and Benjamin Säfken. 2024. GPTopic: Dynamic and Interactive Topic Representations.

Kyle Seelman, Mozhi Zhang, and Jordan Boyd-Graber. 2024. Labeled Interactive Topic Models.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In *Findings of the Association for Computational Linguistics*, pages 1102–1121, Toronto, Canada. ACL.

Anton F. Thielmann, Christoph Weisser, and Benjamin Säfken. 2024. Human in the Loop: How to Effectively Create Coherent Topics by Manually Labeling Only a Few Documents per Class. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 8395–8405, Torino, Italia. ELRA and ICCL.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large Language Models Enable Few-Shot Clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333. Place: Cambridge, MA Publisher: MIT Press.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

A Appendix

A.1 Dataset Details

Amazon Product Reviews Customer reviews organized by product category and star rating (proxy for sentiment). We use English split and consider reviews of the 15 most frequent categories. We sample 500 reviews per star for testing.

Spotify Songtexts Song lyrics categorized by *emotion* and *genre*. We consider the top 10 genres and top 5 emotions for English song texts. We sample 500 song texts per genre for testing.

20 Newsgroups A topic modeling dataset comprising 18000 newsgroups posts on 20 topics. We use 20% of the articles for testing.

GVFC The Gun Violence Frame Corpus comprises news article headlines, annotated into 9 different frames. We use an extension, which includes three-sentence summaries of 1300 articles.

German Blurbs This dataset consists of 20k blurbs categorized into 8 main genres, each with its own fine-grained sub-genres. We sample 1200 blurbs for testing. We consider the main genre.

Israel-Palestine A stance detection dataset of 10k Reddit comments related to the Israel-Palestine conflict, categorized as Israel, Palestine, Neutral.

News Bias A dataset 8.5k news events retrieved from AllSides. It provides three views (left, center, right) for each news event. We sample 10k documents from the top 8 topics for testing.

A.2 Experiment Parameters & Libraries

Document Modification We use the instruction-tuned Gemma 3 (27B) model. The vLLM and LiteLLM libraries handle this process for batch inference, with the *max tokens to generate* parameter set to 2048; other sampling parameters, such as *temperature* and *top p*, remain at their default values. Prompts used for document modification are listed in Table 3.

Document Embedding We use the Sentence Transformers library with *batch size* set to 32 and *normalize embeddings* set to true to generate aspect-oriented embeddings with the instruction-tuned *multilingual-e5-large-instruct* (500M) embedding model. Instructions used for document embedding are listed in Table 2

Dimensionality Reduction A GPU-accelerated variant of UMAP from the cuML library reduces the dimensionality of resulting embeddings. The *distance metric* is cosine similarity, *N neighbors* is 15, and *min distance* is 0.0. For evaluation, *N components* is 2. For production, it is 128.

KNN Accuracy Metric The k-nearest neighbors accuracy metric is computed using the sklearn library via 5-fold cross-validation, with the reported accuracies averaged across folds.

Density-based Clustering We use HDBSCAN for density-based clustering. The *min samples* is set to 40, *distance metric* is set to euclidean.

Cluster Representation c-TF-IDF extracts 50 keywords, Gemma 3 generates title & description using constrained generation to ease parsing.

Interactive Refinement The split operation uses the same HDBSCAN parameters, other operations have no parameters.

Model Fine-tuning We use the SetFit library to fine-tune the embedding model in two stages.

For the first stage (one epoch), ten distinct few-shot training sets are created per dataset and shot count (2, 4, 8, 16) by randomly sampling *num shots* examples per class. They are constructed as contrastive datasets of document pairs, where a pair is deemed positive if both belong to the same class and negative otherwise. Positive pairs are oversampled to balance the dataset, which enables substantial training even with limited labeled samples. We use rank-stabilized LoRA adapters with a *rank* of 24, an *alpha* of 8, and a *dropout* of 0.1, drastically reducing the number of trainable parameters. The training objective is to minimize cosine similarity loss.

The second stage (16 epochs) involves training a differentiable classification head on the updated embedding model. We set the *end-to-end training* parameter to true to further train the embedding model. Since our focus is solely on the document embeddings, the classification head is discarded after training.

A.3 Model Details

We utilized two primary models for the evaluation:

Gemma 3 (27B-instruct) for Document Rewriting This is a *lightweight*, instruction-tuned LLM from Google. It was trained on 14T tokens, features a large 128k context window, and supports over 140 languages.

multilingual-e5-large-instruct (500M) for Document Embedding This instruction-tuned embedding model was selected as one of the best-performing models on the clustering subtask of the MTEB benchmark (Muennighoff et al., 2023). It uses mean pooling as its pooling strategy. It is initialized from XLM Roberta (Conneau et al., 2020), continually trained on a mixture of multilingual datasets, and supports 100 languages.

Dataset	Instruction
emotion	Identify the main emotion expressed in the given (summary of a | keyphrases of a) song text
genre	Identify the main genre of the given (summary of a | keyphrases of a) song text
stars	Identify the sentiment of the given (summary of an | keyphrases of an) Amazon review
product	Identify the category of an Amazon product based on the review (summary | keyphrases)
topic	Identify the topic or theme of the given (summary of a | keyphrases of a) news article
frame	Identify the framing of the given (summary of a | keyphrases of a) news article
blurb	Identifiziere das Genre, dass durch die (Zusammenfassung | Schlüsselbegriffe) des Klappentextes beschrieben wird.
stance	Identify the stance towards the Israel-Palestine conflict of the given (summary | keyphrases) of a news article
bias	Identify the political leaning of the given (summary of a | keyphrases of a) news article

Table 2: Instructions provided to the embedding model. We adopted them to resemble those of the original authors.

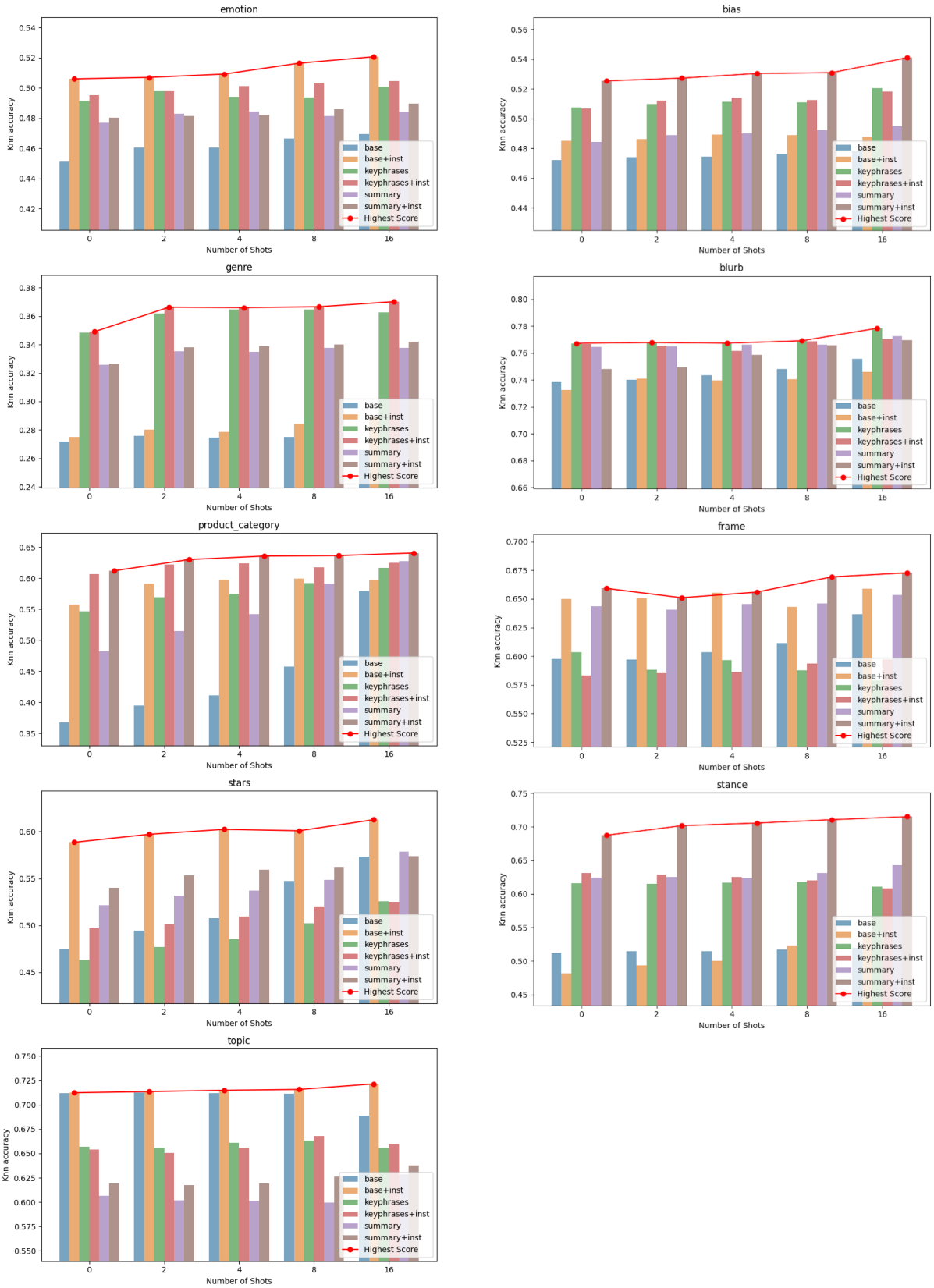


Figure 3: Detailed evaluation with increasing number of labeled examples per class.

Dataset	Prompt
emotion-summary	Write a concise summary (maximum 5 sentences) that focuses on the emotional tone of the following song lyrics. Analyze the lyrics to determine the main emotion being conveyed and describe how it is expressed. Conclude with an emotion categorization:
emotion-keyphrases	Generate keyphrases (max 5 phrases) that describe the emotional tone of the following song lyrics. Focus on phrases that reflect the emotional tone, mood, or feelings expressed in the lyrics.
genre-summary	Write a concise summary (maximum 5 sentences) that focuses on the genre of the following song lyrics. Analyze the lyrics to determine the musical genre, referencing stylistic elements, themes, or influences. Conclude with a genre categorization:
genre-keyphrases	Generate keyphrases (max 5 phrases) that describe the musical genre of the following song lyrics. Focus on phrases that reflect the genre's characteristics, style, or influences.
stars-summary	Write a concise summary (maximum 5 sentences) that focuses on the sentiment of the following Amazon review. Analyze the review to determine the main sentiment being conveyed and describe how it is expressed. Conclude with a sentiment categorization:
stars-keyphrases	Generate keyphrases (max 5 phrases) that describe the sentiment of the following Amazon review. Focus on phrases that reflect the sentiment, mood, or feelings expressed in the review.
product-summary	Write a concise summary (maximum 5 sentences) that focuses on the product categorization of the following Amazon review. Analyze the review to determine the discussed product's categorization, referencing its features, type, or purpose. Conclude with a categorization:
product-keyphrases	Generate keyphrases (max 5 phrases) that describe the product of the following Amazon review. Focus on phrases that reflect the product's type or category.
topic-summary	Write a concise summary (maximum 5 sentences) that focuses on the topic or theme of the following news article. Analyze the article to determine the main topic or theme being discussed. Conclude with a topic categorization:
topic-keyphrases	Generate keyphrases (max 5 phrases) that describe the topic of the following news article. Focus on phrases that reflect the main topic being discussed.
blurbs-summary	Schreibe eine prägnante Zusammenfassung (maximal 5 Sätze), die sich auf das Genre des folgenden Klappentextes konzentriert. Analysiere den Text, um das Genre zu bestimmen. Schließe mit einer allgemeinen Genre-Kategorisierung ab.
blurbs-keyphrases	Generiere Schlüsselbegriffe (max 5 Begriffe), die das Genre des folgenden Klappentextes beschreiben. Fokusiere dich auf Begriffe, die das allgemeine Genre widerspiegeln.
stance-summary	Write a concise summary (maximum 5 sentences) that focuses on the stance of the following Reddit post. Analyze the post to determine whether it is Pro-Israel, Pro-Palestine, or Neutral. Conclude with a stance categorization:
stance-keyphrases	Generate keyphrases (max 5 phrases) that describe the stance of the following Reddit post. Focus on phrases that reflect the stance being discussed.
bias-summary	Write a concise summary (maximum 5 sentences) that focuses on the political leaning of the following news article. Analyze the article to determine the main political framing. Conclude with a political categorization:
bias-keyphrases	Generate keyphrases (max 5 phrases) that describe the political leaning of the following news article. Focus on phrases that reflect the political framing.

Table 3: Prompts used to rewrite the original documents. We use the same prompts for topics and frames.

Author Index

Almogi, Orna, 150
Amano, Kyoko, 41
Arias, Esteban Garces, 276, 297
Aßenmacher, Matthias, 276

Bajetta, Nicola, 150
Bar, Kfir, 150
Barré, Jean, 255
Basta, Christine, 50
Biemann, Chris, 411
Bizzoni, Yuri, 190
Breton, Alice, 69
Brody, Justin, 320
Burgers, Gert-Jan, 83

Cabral, Cassandra, 167
Cafiero, Florian, 346
Chae, Songeun, 24
Chatterjee, Ahan, 276
Choden, Sonam, 150
Choi, Jiwoo, 24
Choi, Yeeun, 13
Ciotlos, Cosmin, 381
Cohen, Shay, 150
Consolo, Giuseppe, 236

Dinu, Liviu P, 381
Drori, Omri, 150

El-Wazzi, Rami, 202

Fine, Shai, 150
Fischer, Tim, 411
Fraser, Alexander, 106
Fundal, Halfdan Nordahl, 190

Ganascia, Jean-Gabriel, 69
Gao, Fangzhou, 320
Ginter, Filip, 214
Gordon, Jonathan, 226
Griebel, Sarah, 178
Grimes, Tyler, 242
Guhr, Svenja, 368
Guo, Haoze, 1

Harb, Abdelrahman, 50
Henriksson, Erik, 138
Horin, Goody Ben, 150

Iacob, Alina, 381
Icard, Benjamin, 69

Jahan, Labiba, 264
Jamtsho, Sonam, 150
Jeong, Cheonkam, 13
Jung, Donghoon, 24
Jung, Seohyon, 24

Kanner, Antti, 138
Keramati, Ali, 325
Khan, Ishmam, 355
Khan, Md Tarique Hasan, 264
Kister, Pauline, 390
Kovács, Tamás, 236
Krause, Lea, 83
Kyogoku, Yuki, 41

Lacsam, Moneah Shaila, 167
Laippala, Veronika, 138
Li, Feiyue, 58
Lionnet-Rollin, Alexandre, 346
Lundberg, Tuomas, 138
LYU, ZEYU, 132

Mahadevan, Ananth, 214
Mathioudakis, Michael, 214
Mehrotra, Sharad, 325
Miyagawa, So, 41

Ocal, Mustafa, 264
Okabe, Shu, 106

Ponay, Charmaine Salvador, 167
Puttick, Alexandre, 202

Rabinovitz, Gal, 150
Raihan, Nishat, 96
Razzaghi, Alireza, 138
Reichbauer, Sebastian, 106

Sainero, Lila, 69
Schirmer, Miriam, 390
Schöffel, Matthias, 276, 297
Shmidman, Avi, 121
Shmidman, Cheyn Shmuel, 121
Shtrosberg, Ofir, 150

Sundén, Rebecca, 150
Swearingen, Jadon, 264

Tabuzo, Abdiel Clarence, 167
Tal, Rachel, 121
Thogarrati, Sindhuja, 355
Tiwari, Tanishk, 314
Tolonen, Mikko, 214
Toutou, Ammar, 50
Tsukagoshi, Yuzuki, 41

Underwood, Ted, 178

Velazco, Vladimir Gray, 167
Vogeler, Georg, 236

Wang, Zhichao, 132
Wangchuk, Dorji, 150
Warschauer, Mark, 325
Washington, Marshall, 242
Wei, Ziqi, 1
Wiedner, Marinus, 276
Wimalasuriya, Ravini, 83
Wu, Yu, 214

Yang, Jingyi, 150
Yeshpanov, Rustem, 31

Zampieri, Marcos, 96
Zhang, Linyue, 58
Zhang, Shuo, 355
Zhou, Shiyuan, 325
Zve, Evangelia, 69