

CODI 2025

**The 6th Workshop on Computational Approaches to
Discourse, Context and Document-Level Inferences (CODI
2025)**

Proceedings of the Workshop

November 9, 2025

The CODI organizers gratefully acknowledge the support from the following sponsors.

Sponsor



©2025 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 979-8-89176-343-2

Preface

Welcome to the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI-2025), co-organized with the 8th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC) and co-located with EMNLP 2025 in Suzhou, China!

CODI provides a venue to bring together researchers working on all aspects of discourse in Computational Linguistics, NLP and AI. Our aim is to foster an active and collaborative community around discourse processing by offering a platform to present and exchange theories, algorithms, software, datasets, and tools, as well as to discuss emerging challenges, new ideas, and future directions in the field.

The workshop consists of invited talks, contributed papers, extended abstracts, and EMNLP Findings presentations. We received paper submissions that span a wide range of topics, addressing issues related to discourse representation and parsing, connectives, entity tracking, discourse benchmarks, ambiguity, computational discourse and stance analysis, dialogue, applications, and more. As the workshop is hybrid this year, papers are presented live either in person or remotely and discussed during live Q&A sessions. We received 28 submissions, and accepted 14 regular long papers, 5 regular short papers, as well as 3 non-archival communications (Findings and extended abstracts). We accepted a total of 22 articles among the 28 regular submissions, 15 of which were presented in person and 6 are presented orally. We also organized two poster sessions this year, in order to encourage discussions.

This year the workshop hosted the fourth edition of the DISRPT shared task on Discourse Relation Parsing and Treebanking (DISRPT 2025), whose five submissions form part of a separate shared task proceedings volume. The workshop also included a discussion on future shared tasks, and sessions on coreference and anaphora resolution organized by CRAC, for which papers are also published in a separate proceedings volume.

We thank our invited speakers, **Nancy Chen** (Multimodal Generative AI Group Leader at the Institute for Infocomm Research and AI for Education Head at A*STAR) and **Tanya Goyal** (Assistant Professor in the Department of Computer Science at Cornell University). They helped us to prepare an excellent and well-rounded workshop program. We would also like to thank the EMNLP 2025 workshop chairs, Sunipa Dev, Maja Popović and Eleftherios Avramidis, who organized the workshops program.

Finally, we thank our sponsor HITS, Heidelberg Institute for Theoretical Studies <https://www.h-its.org>.

The CODI Organizers,

Chloé Braud, Christian Hardmeier, Chuyuan Li, Junyi Jessy Li, Sharid Loáiciga, Michael Strube and Amir Zeldes

Program Committee

Program Committee

Giuseppe Carenini, university of british columbia
Jackie Chi Kit Cheung, Mila / McGill University
Elisa Ferracane, Abridge AI, Inc.
Mark Finlayson, FIU
Zhengxian Gong, Computer science and technology school, soochow university
Jie He, University of Edinburgh
Veronique Hoste, LT3, Ghent University
Cassandra L. Jacobs, University at Buffalo
Yangfeng Ji, University of Virginia
Murathan Kurfali, RISE Research Institutes of Sweden
Sobha Lalitha Devi, AU-KBC Research Centre, Anna University
Wanqiu Long, The University of Edinburgh
Anna Nedoluzhko, Charles University in Prague
Jingcheng Niu, University of Toronto
Gerald Penn, University of Toronto
Hannah Rohde, University of Edinburgh
Ahmed Ruby, Uppsala university
Hannah Seemann, Ruhr-University
Bonnie Webber, University of Edinburgh
Suncheng Xiang, Shanghai Jiao Tong University
Deniz Zeyrek, Middle East Technical University
Zheng Zhao, University of Edinburgh

Table of Contents

<i>Long Context Benchmark for the Russian Language</i> Igor Churin, Murat Apishev, Maria Tikhonova, Denis Shevelev, Aydar Bulatov, Yuri Kuratov, Sergei Averkiev and Alena Fenogenova	1
<i>Unpacking Ambiguity: The Interaction of Polysemous Discourse Markers and Non-DM Signals</i> Jingni Wu and Amir Zeldes	14
<i>Enhancing the Automatic Classification of Metadiscourse in Low-Proficiency Learners' Spoken and Written English Texts Using XLNet</i> Wenwen Guan, Marijn Alta and Jelke Bloem	27
<i>Entity Tracking in Small Language Models: An Attention-Based Study of Parameter-Efficient Fine- Tuning</i> Sungho Jeon and Michael Strube	42
<i>Stance Detection on Nigerian 2023 Election Tweets Using BERT: A Low-Resource Transformer-Based Approach</i> Mahmoud Ahmad and Habeebah Kakudi	54
<i>Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production</i> Debasmita Bhattacharya, Anxin Yi, Siying Ding and Julia Hirschberg	64
<i>Otherwisein Context: Exploring Discourse Functions with Language Models</i> Guifu Liu, Bonnie Webber and Hannah Rohde	81
<i>On the Role of Context for Discourse Relation Classification in Scientific Writing</i> Stephen Wan, Wei Liu and Michael Strube	96
<i>Zero-Shot Belief: A Hard Problem for LLMs</i> John Murzaku and Owen Rambow	107
<i>Probing the Limits of Multilingual Language Understanding: Low-Resource Language Proverbs as LLM Benchmark for AI Wisdom</i> Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak and Usman Naseem	120
<i>Measuring Sexism in US Elections: A Comparative Analysis of X Discourse from 2020 to 2024</i> Anna Fuchs, Elisa Noltenius, Caroline Weinzierl, Bolei Ma and Anna-Carolina Haensch	130
<i>Discourse Relation Recognition with Language Models Under Different Data Availability</i> Shuhaib Mehri, Chuyuan Li and Giuseppe Carenini	148
<i>EmbiText: Embracing Ambiguity by Annotation, Recognition and Generation of Pronominal Reference with Event-Entity Ambiguity</i> Anna Sheikh and Christian Hardmeier	157
<i>Human and LLM-based Assessment of Teaching Acts in Expert-led Explanatory Dialogues</i> Aliko Anagnostopoulou, Nils Feldhus, Yi-Sheng Hsu, Milad Alshomary, Henning Wachsmuth and Daniel Sonntag	166
<i>Where Frameworks (Dis)agree: A Study of Discourse Segmentation</i> Maciej Ogrodniczuk, Anna Latusek, Karolina Saputa, Alina Wróblewska, Daniel Ziembicki, Bar- tosz Żuk, Martyna Lewandowska, Adam Okraśiński, Paulina Rosalska, Anna Śliwicka, Aleksandra Tomaszewska and Sebastian Żurowski	182

<i>Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser</i>	
Elena Chistova.....	197
<i>Corpus-Oriented Stance Target Extraction</i>	
Benjamin Steel and Derek Ruths.....	209
<i>Information-Theoretic and Prompt-Based Evaluation of Discourse Connective Edits in Instructional Text Revisions</i>	
Berfin Aktas and Michael Roth.....	228

Program

Sunday, November 9, 2025

- 09:00 - 09:10 *Opening Remarks*
- 09:10 - 10:00 *Invited Talk - Nancy F. Chen: From Speech to Sense: The Art of Listening in Artificial Intelligence*
- 10:00 - 10:15 *CRAC 2024 Shared Task on Multilingual Coreference Resolution*
- 10:15 - 10:30 *Discourse Relation Parsing and Treebanking (DISRPT) Shared Task*
- 10:30 - 11:00 *Coffee Break*
- 11:00 - 12:00 *Poster Session 1*
- Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production*
Debasmita Bhattacharya, Anxin Yi, Siying Ding and Julia Hirschberg
- Discourse Relation Recognition with Language Models Under Different Data Availability*
Shuhaib Mehri, Chuyuan Li and Giuseppe Carenini
- Where Frameworks (Dis)agree: A Study of Discourse Segmentation*
Maciej Ogrodniczuk, Anna Latusek, Karolina Saputa, Alina Wróblewska, Daniel Ziembicki, Bartosz Żuk, Martyna Lewandowska, Adam Okrański, Paulina Rosalska, Anna Śliwicka, Aleksandra Tomaszewska and Sebastian Żurowski
- Information-Theoretic and Prompt-Based Evaluation of Discourse Connective Edits in Instructional Text Revisions*
Berfin Aktas and Michael Roth
- Joint Modeling of Entities and Discourse Relations for Coherence Assessment*
Wei Liu and Michael Strube
- 12:00 - 13:30 *Lunch*
- 13:30 - 14:20 *Invited Talk - Tanya Goyal: Climbing the Right Hill: On Benchmarking Progress in Long-Form Text Processing*
- 14:20 - 15:30 *Poster Session 2*

Sunday, November 9, 2025 (continued)

Long Context Benchmark for the Russian Language

Igor Churin, Murat Apishev, Maria Tikhonova, Denis Shevelev, Aydar Bulatov, Yuri Kuratov, Sergei Averkiev and Alena Fenogenova

Enhancing the Automatic Classification of Metadiscourse in Low-Proficiency Learners' Spoken and Written English Texts Using XLNet

Wenwen Guan, Marijn Alta and Jelke Bloem

Stance Detection on Nigerian 2023 Election Tweets Using BERT: A Low-Resource Transformer-Based Approach

Mahmoud Ahmad and Habeebah Kakudi

Otherwisein Context: Exploring Discourse Functions with Language Models

Guifu Liu, Bonnie Webber and Hannah Rohde

On the Role of Context for Discourse Relation Classification in Scientific Writing

Stephen Wan, Wei Liu and Michael Strube

Automated Conspiracy Narrative Detection Across Social Media Platforms

Calvin Yixiang Cheng and Mohsen Mosleh

Zero-Shot Belief: A Hard Problem for LLMs

John Murzaku and Owen Rambow

Probing the Limits of Multilingual Language Understanding: Low-Resource Language Proverbs as LLM Benchmark for AI Wisdom

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak and Usman Naseem

Measuring Sexism in US Elections: A Comparative Analysis of X Discourse from 2020 to 2024

Anna Fuchs, Elisa Noltenius, Caroline Weinzierl, Bolei Ma and Anna-Carolina Haensch

EmbiText: Embracing Ambiguity by Annotation, Recognition and Generation of Pronominal Reference with Event-Entity Ambiguity

Amna Sheikh and Christian Hardmeier

Human and LLM-based Assessment of Teaching Acts in Expert-led Explanatory Dialogues

Aliki Anagnostopoulou, Nils Feldhus, Yi-Sheng Hsu, Milad Alshomary, Henning Wachsmuth and Daniel Sonntag

Sunday, November 9, 2025 (continued)

Multi-token Mask-filling and Implicit Discourse Relations

Meinan Liu, Yunfang Dong, Xixian Liao and Bonnie Webber

Consistent Discourse-level Temporal Relation Extraction Using Large Language Models

Yi Fan and Michael Strube

15:30 - 16:00 *Coffee Break*

16:00 - 17:30 *Session - Oral presentations*

Unpacking Ambiguity: The Interaction of Polysemous Discourse Markers and Non-DM Signals

Jingni Wu and Amir Zeldes

Entity Tracking in Small Language Models: An Attention-Based Study of Parameter-Efficient Fine-Tuning

Sungho Jeon and Michael Strube

Corpus-Oriented Stance Target Extraction

Benjamin Steel and Derek Ruths

Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser

Elena Chistova

17:30 - 17:45 *Closing Remarks*

Long Context Benchmark for the Russian Language

Igor Churin¹, Murat Apishev^{1,2}, Maria Tikhonova^{1,3}, Denis Shevelev¹,
Aydar Bulatov^{4,5}, Yuri Kuratov^{4,5}, Sergej Averkiev¹, Alena Fenogenova^{1,3}

¹SberAI, ²Yandex, ³HSE University, ⁴AIRI, ⁵MIPT

Correspondence: alenush93@gmail.com

Abstract

Recent progress in Natural Language Processing (NLP) has driven the creation of Large Language Models (LLMs) capable of tackling a vast range of tasks. A critical property of these models is their ability to handle large documents and process long token sequences, which has fostered the need for a robust evaluation methodology for long-text scenarios. To meet this requirement in the context of the Russian language, we present our benchmark consisting of 18 datasets designed to assess LLM performance in tasks such as information retrieval, knowledge extraction, machine reading, question answering, and reasoning. These datasets are categorized into four levels of complexity, enabling model evaluation across context lengths up to 128k tokens. To facilitate further research, we provide open-source datasets, a codebase, and a public leaderboard associated with the benchmark.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive abilities in many NLP applications. Interacting with people through free-form text instructions, they serve as versatile tools for multiple scenarios, transforming the landscape of AI systems. One direction where LLM usage is developing rapidly includes tasks requiring long text processing, such as information retrieval (IR) and summarization, where their applications alleviate the handling of long texts for humans.

However, until recently, most LLMs had difficulties handling long sequences of tokens and were only able to work with a limited context length of several thousand tokens. In recent years, new methods have enabled the models to increase their context significantly, empowering them to solve a new variety of tasks. This, in turn, and the community’s demand for automatic systems solving such tasks at a good level has created a need for a

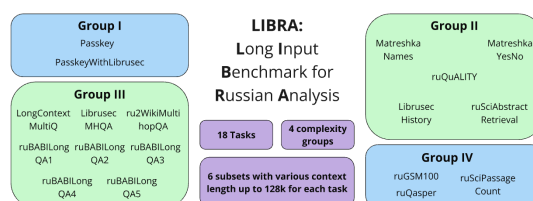


Figure 1: **The LIBRA benchmark** is a set of 18 long-context tasks ranging in length from 4k to 128k tokens, grouped into four categories based on the complexity of required skills.

thorough evaluation of LLM long context understanding.

To address this demand in English, several long context understanding benchmarks have been created recently with LongBench (Bai et al., 2023)¹ and L-Eval (An et al., 2023)² heading the list. However, the Russian language, at this point, lacks a fair instrument for transparent evaluation of long context understanding.

Our work addresses this problem and presents a new benchmark, which we call **Long Input Benchmark for Russian Analysis**, or **LIBRA**, for the evaluation of LLM long context understanding abilities in Russian (see Figure 1) including such aspects as IR, machine reading, question answering (QA), and reasoning. The contribution of our work can be summarized as follows:

- we present a methodology for the evaluation of long-context abilities of LLMs for the Russian language;
- we publicly release a set of 18 datasets of various skills and complexities in Russian, which form the LIBRA benchmark;

¹<https://huggingface.co/datasets/THUDM/LongBench>

²<https://huggingface.co/datasets/L4NLP/LEval>

- we release a codebase ³, a public leaderboard ⁴ and a set of baseline solutions.

2 Related Work

Long Context Large Language Models. One of the crucial tasks in the development of LLMs is to increase the length of the context that the model can understand. This problem has two key points: the complexity of calculations for long sequences and the ability of the model to extract important data in a long context. The solution of the first problem can be attributed to research on the effective processing of the self-attention as in Longformer (Beltagy et al., 2020), LongNet (Ding et al., 2023) and FlashAttention (Dao et al., 2022; Dao, 2023), using caches for previously calculated outputs such as Transformer-XL (Dai et al., 2019), Unlimiformer (Bertsch et al., 2024) and LongLLaMA (Tworkowski et al., 2024) or replacing it with another mechanism with more effective inference as in RetNet (Sun et al., 2023) and Mamba (Gu and Dao, 2023). The solution to the second problem is to improve positional encoding techniques such as ALiBi (Press et al., 2021) and RoPE-based approaches (Sun et al., 2022; Peng et al., 2023).

Long Context Benchmarks. Until recently, most LMs had relatively small context lengths limited by a few thousand tokens. Thus, standard Natural Language Understanding (NLU) benchmarks (Wang et al., 2018, 2019; Shavrina et al., 2020) contained tasks within this size. Even today, benchmarks created recently, such as HELM (Bommasani et al., 2023), MT-Bench (Zheng et al., 2023), and Russian-oriented benchmark MERA (Fenogenova et al., 2024) follow this pattern, limiting their tasks by relatively small context window size to simplify the evaluation procedure and reducing its cost.

The pioneers of long context processing benchmarks have been ZeroSCROLLS (Shaham et al., 2023)⁵, designed to test zero-shot model capabilities for NLU over long texts; L-eval (An et al., 2023)⁶, focused on a standardized evaluation methodology for long context LMs addressing two key aspects: dataset construction and evaluation metrics; Loong (Wang et al., 2024b), which aligns

with realistic scenarios through extended multi-document QA; and LongBench (Bai et al., 2023), the bilingual multi-task benchmark for long context understanding, comprising 21 tasks in English and Chinese. Finally, Goldman et al. (2024) categorizes the existing long-context datasets and positions them with respect to their difficulty, which they define by the dispersion and the scope.

The concept of the *needle-in-a-haystack* (Kamradt, 2023) is frequently employed in long-context benchmarks, involving the insertion of sentence-level information at varying depths within a document to create tasks of differing complexity. In addition to categorizing tasks by type, many benchmarks classify tasks based on their complexity using various criteria. For instance, tasks may be grouped by the number of facts required for reasoning (Kuratov et al., 2024), the type of reasoning QA (e.g., single-hop vs. multi-hop, as shown by Wang et al. (2024a)), or the complexity and depth of the needle. In the latter case, deeper or more abstract needles challenge models more significantly, testing their ability to locate and reason over critical details in long documents (Karpinska et al., 2024).

However, the limitation of the benchmarks mentioned above is that they are mainly English-oriented (or Chinese). As for the Russian language, there is an urgent need for a reliable system able to evaluate LLM long context understanding abilities. To address this problem, we propose LIBRA, which brings a methodology and 18 tasks for a long context understanding evaluation in Russian.

3 LIBRA

3.1 Benchmark Overview

In this section, we introduce LIBRA (Long Input Benchmark for Russian Analysis), a new benchmark for long context understanding, which includes 18 tasks for LLM evaluation created specifically for Russian. LIBRA aims to evaluate a large scope of LLMs, including pretrain models and models with supervised finetuning (SFT) with any system prompt that can be picked up.

The main purpose of the benchmark is to create a reliable instrument for the long context understanding evaluation, enabling the study of the model’s ability to solve various tasks of different complexity with respect to the input context length. For this purpose, all tasks in the LIBRA benchmark are divided into 4 complexity groups, and the datasets have several subsets of various context lengths rang-

³<https://github.com/ai-forever/LIBRA>

⁴<https://huggingface.co/spaces/ai-forever/LIBRA-Leaderboard>

⁵<https://www.zero.scrolls-benchmark.com/>

⁶<https://huggingface.co/papers/2307.11088>

	Task Name	Data Origin	Skills	Metric	Dataset Size
I	Passkey	New	Reasoning	EM	1200
	PasskeyWithLibrusec	New	Reasoning	EM	1200
II	MatreshkaNames	New	Dialogue Context, Reasoning	EM	900
	MatreshkaYesNo	New	Dialogue Context, Reasoning	EM	1799
	LibrusecHistory	New	Reasoning	EM	128
	ruSciAbstractRetrieval	New	Reasoning	EM	1240
	ruQuALITY	Translated	Reasoning	EM	202
III	LongContextMultiQ	New	Reasoning	EM	1200
	LibrusecMHQA	New	Reasoning	EM	384
	ru2WikiMultihopQA	Translated	Reasoning	EM	300
	ruBABILongQA1	New	Reasoning	EM	600
	ruBABILongQA2	New	Reasoning	EM	600
	ruBABILongQA3	New	Reasoning	EM	600
	ruBABILongQA4	New	Reasoning	EM	600
	ruBABILongQA5	New	Reasoning	EM	600
IV	ruSciPassageCount	New	Reasoning	EM	600
	ruQasper	Translated	Reasoning	F1	203
	ruGSM100	Translated	Math, Logic	EM	100

Table 1: The LIBRA tasks outline. The numbers **I**, **II**, **III**, and **IV** in the left column indicate the complexity group of the tasks described in Subsection 3.2. The **Skills** column defines the skills to be tested on a specific task. **Data Origin** discloses the source of the dataset. The **Dataset Size** column shows the number of items in the whole dataset.

ing from 4k up to 128k tokens⁷. The latter makes it possible to explore the influence of the context length on the model results.

3.2 Complexity group description

We describe each complexity group of tasks using criteria inspired by other benchmarks that classify tasks by complexity. Specifically, we considered the depth of the needle, the complexity of reasoning, and the difficulty of the domain.

The first complexity group (I) consists of tasks that require finding a short text fragment in long textual paragraphs containing irrelevant information. This group includes Passkey and PasskeyWithLibrusec datasets.

The second complexity group (II) includes tasks that require answering the question based on a relevant context. The following types of tasks are related to this group: QA such as MatreshkaNames, MatreshkaYesNo, LibrusecHistory, ruSciAbstractRetrieval and multiple choice QA, such as ruQuALITY.

The natural development of tasks from the second class of complexity are tasks with questions, the answers to which are not explicitly contained in the text but require the analysis of fragments of input data and the generation of an answer based on it. Such tasks in our classification belong to **the third complexity group (III)** and represent a multi-hop

QA (MHQA) type. This group includes the following tasks: ruBABILongQA1, ruBABILongQA2, ruBABILongQA3, ruBABILongQA4, ruBABILongQA5, LongContextMultiQ, LibrusecMHQA and ru2WikiMultihopQA.

Finally, to **the fourth complexity group (IV)** belongs to the tasks that require understanding the whole context, solving mathematical problems, and QA tasks within complex domains. This group includes ruSciPassageCount, ruGSM100 and ruQasper datasets.

We do not include code generation and analysis tasks in LIBRA as most of the software code in the world is written in languages based on English.

3.3 Context Length Estimation

We divide all datasets into subsets of various context lengths. The latter, however, may vary across different models and tokenizers. In our work, we used the fertility of tokenizer to distribute samples across different context lengths, which indicates the average number of tokens in which one word is tokenized. Thus, the average length in tokens for the text can be approximated by the number of words multiplied by the fertility number.

For the fertility approximation, we calculate the average fertility of the classic LLM tokenizers, which are further evaluated as baselines (see Appendix C for model description) on a complete list of datasets, by computing it as the total number of tokens divided by the total number of words. The

⁷See explanation on token length calculation in Section 3.3

Model Name	Fertility
GLM4-9B-Chat	2.15
T-lite-instruct-0.1	2.34
Saiga-LLaMA-3-8B	2.40
LLaMA-3-8B	2.40
LLaMA-3-8B-Instruct	2.40
LLaMA-3.1-8B-Instruct	2.40
LLaMA-3.1-8B	2.40
Phi-3-mini-128k-instruct	2.74
LLaMA-2-7B-32K	2.83
LongAlpaca-7B	2.83
LongChat-7B-v1.5-32k	2.83
Mistral-7B-v0.1	3.08
Mistral-7B-v0.3	3.08
Mistral-7B-Instruct-v0.3	3.08
Mistral-Nemo-Instruct-2407	3.08
ChatGLM2-6B-32k	3.50

Table 2: The average fertility of tokenizers, where fertility is defined as the average number of tokens per word.

fertility of each model is shown in Table 2. The average fertility is 2.7. However, we decided to choose it with a margin so that the multilingual model with the highest fertility can be tested on the entire benchmark. As a result, we set the standard fertility to 3.

Finally, using the selected fertility value, we divided all datasets into subsets of various context lengths ranging from 4k to 128k tokens. Table 3 gives the resulting dataset sizes and average sample context lengths.

3.4 Datasets

This section describes the datasets and data collection process in detail. The benchmark datasets originate from the following sources: 1) entirely new datasets based on open data in Russian (14 datasets out of 18) and 2) translation of English datasets using Google translator API⁸ followed by manual verification and correction. We do not generate samples with LLMs and use annotators markup. This helps reduce bias from using models like GPT-4, which are also part of the assessment. However, it has some drawbacks, as full annotation can be costly and time-consuming in certain cases. The exact dataset format can be found in Appendix B.

Passkey The Passkey is a synthetic QA dataset based on the idea of the original passkey dataset from LongLLaMA’s GitHub repository⁹. The main idea of the task is to extract a relevant piece of code number from a long text fragment that was created by repeating short sentence template containing

noise. The model must find this code among the irrelevant information.

PasskeyWithLibrusec The PasskeyWithLibrusec is a more complicated version of Passkey QA dataset, in which we use randomly selected texts from the Librusec dataset¹⁰ as noise to make this dataset more difficult for LLMs.

MatreshkaNames This dataset is based on Russian names¹¹ and Matreshka¹². The Matreshka dataset comprises brief interactions involving “user” and “bot” roles, along with a brief description of the topic being discussed by each participant. To form longer contextual samples, we combined multiple interactions and replaced the names “user” and “bot” with the pull of names taken from the dataset of Russian names. Subsequently, we randomly selected a topic from the combined interactions and the name of the person discussing that topic. The dataset requires the model to identify the individual who discussed the selected topic.

MatreshkaYesNo The MatreshkaYesNo is a binary classification dataset based on Matreshka and Russian names sets. It is similar to the MatreshkaNames dataset but instead of predicting names, the model is supposed to indicate whether this topic was mentioned in the dialog. The dataset is balanced across Yes/No answers.

LibrusecHistory This dataset was created in QA format using Librusec. Each sample comprises a text paragraph and a corresponding question. To create tasks with different input lengths, we selected large texts from books in different domains and styles, divided them into fragments of several thousand tokens, and created the annotation (see Appendix A). These fragments became the dataset’s samples. Longer samples, with lengths up to 64,000 tokens, were created by supplementing these fragments with neighboring paragraphs from the original large text on both sides.

ruSciAbstractRetrieval The ruSciAbstractRetrieval is a QA dataset ideologically similar to the PassageRetrieval (Bai et al., 2023)¹³ dataset from LongBench, that aims to evaluate model’s reasoning skills. Each element of the dataset consists of a summary description of the topic and a set text para-

⁸<https://pypi.org/project/googletrans/>

⁹https://github.com/CStanKonrad/long_llama/blob/main/examples/passkey.py

¹⁰<https://huggingface.co/datasets/IlyaGusev/librusec>

¹¹<https://www.kaggle.com/datasets/rai220/russian-cyrillic-names-and-sex/data>

¹²<https://huggingface.co/datasets/zjkarina/matreshka>

¹³https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_retrieval_en

	Dataset Name	4k size / avg len	8k size / avg len	16k size / avg len	32k size / avg len	64k size / avg len	128k size / avg len
I	Passkey	200 / 2790	200 / 5450	200 / 10996	200 / 21730	200 / 43391	200 / 87974
	PasskeyWithLibrusec	200 / 2705	200 / 5563	200 / 10835	200 / 22215	200 / 44682	200 / 88189
II	MatreshkaNames	150 / 3190	150 / 6314	150 / 12128	150 / 24168	150 / 48184	150 / 96135
	MatreshkaYesNo	299 / 3200	300 / 6317	300 / 12134	300 / 24173	300 / 48189	300 / 96142
	LibrusecHistory	-	32 / 4515	32 / 9003	32 / 17976	32 / 35924	-
	ruSciAbstractRetrieval	210 / 3264	210 / 7260	210 / 15245	210 / 31231	200 / 63594	200 / 127777
	ruQuALITY	-	41 / 6380	161 / 12387	-	-	-
III	LongContextMultiQ	200 / 2940	200 / 6360	200 / 12240	200 / 26572	200 / 37482	200 / 68239
	LibrusecMHQA	-	384 / 4574	-	-	-	-
	ru2WikiMultihopQA	-	49 / 6378	128 / 11633	123 / 25523	-	-
	ruBABILongQA1	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA2	100 / 4002	100 / 8001	100 / 16002	100 / 32001	100 / 64002	100 / 128001
	ruBABILongQA3	100 / 4011	100 / 8010	100 / 16011	100 / 32010	100 / 64011	100 / 128010
	ruBABILongQA4	100 / 4014	100 / 8013	100 / 16014	100 / 32013	100 / 64014	100 / 128013
	ruBABILongQA5	100 / 4006	100 / 8005	100 / 16006	100 / 32005	100 / 64006	100 / 128005
IV	ruSciPassageCount	100 / 3528	100 / 7128	100 / 13616	100 / 27160	100 / 53108	100 / 105949
	ruQasper	-	48 / 5768	134 / 11071	21 / 25185	-	-
	ruGSM100	-	-	100 / 9083	-	-	-

Table 3: Sizes and average sample lengths for the task subsets of various context lengths. **Dataset Name** shows the name of the dataset. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** show the number of samples and average sample lengths in tokens for the corresponding context length.

graphs created from abstracts of scientific articles from ruSciBench¹⁴. The goal is to identify the paragraph where the specified topic is discussed. To create this dataset, we randomly choose some abstracts from ruSciBench and generate descriptions of their topics using human annotators to acquire targets.

ruQuALITY The ruQuALITY dataset was created as a translation of the original QuALITY¹⁵ from L-Eval, which consists of selected samples with a long context from the original multiple choice QA dataset called QuALITY (Pang et al., 2021). The model must find relevant information in the text and answer by choosing one of the four suggested options.

LongContextMultiQ The LongContextMultiQ is a multi-hop QA long context dataset for Russian that is based on data used for the MultiQ (Taktasheva et al., 2022)¹⁶ dataset creation. The original MultiQ dataset is created by multi-hop dataset generation based on Wikidata¹⁷ and Wikipedia, and consists of samples with different length. We selected 200 samples from these generated sources with a long context for each context length.

LibrusecMHQA This dataset was created in

¹⁴https://huggingface.co/datasets/mlsa-iai-msu-lab/ru_sci_bench

¹⁵<https://huggingface.co/datasets/L4NLP/LEval/viewer/quality>

¹⁶<https://huggingface.co/datasets/ai-forever/MERA/viewer/multiq>

¹⁷<https://www.wikidata.org/wiki/Wikidata:Introduction>

MHQA format, also using Librusec as a LibrusecHistory. The main difference between these datasets is that in the LibrusecMHQA dataset, the necessary information for the answer is distributed in several parts of the context, making the task more difficult and allowing us to evaluate the model’s reasoning skills better. The generation procedure for samples of different lengths remains the same.

ru2WikiMultihopQA The ru2WikiMultihopQA was created by translating the dataset 2WikiMultihopQA¹⁸ from LongBench, which consists of selected samples with a long context from the original multi-hop QA dataset 2WikiMultihopQA (Ho et al., 2020). This Wikipedia-based dataset tests reasoning skills by requiring a model to combine information from multiple texts to answer a question. The format of this dataset, which consists of up to 5-hop questions, makes it difficult for LLMs.

ruBABILong We created a new methodology based on the idea from Kuratov et al. (2024) to create the Russian Benchmark for Artificial Intelligence for Long (ruBABILong)-context evaluation. It contains five long-context QA reasoning tasks using facts hidden among distractor facts and long books. The **ruBABILongQA1** task requires answering a question about a person’s location using a single supporting fact. The **ruBABILongQA2** and **ruBABILongQA3** tasks introduce the challenge of differentiating subjects and objects, utilizing

¹⁸https://huggingface.co/datasets/THUDM/LongBench/viewer/2wikimqa_e

two and three supporting facts, respectively. The **ruBABILongQA4** task tackles spatial reasoning through two-argument relations, while the **ruBABILongQA5** task involves tracking multiple objects to solve the three-argument relation problem. Each task contains 100 samples, scaled to six sequence lengths from 4k to 128k. We constructed the task facts for Russian according to the methodology of the original data from the bAbI dataset (Weston et al., 2016); no translation was performed, and facts were created directly in Russian. The background texts were sampled from Russian Librusec books.

ruSciPassageCount The dataset ruSciPassageCount uses the basic idea of the original PassageCount¹⁹ dataset. This QA dataset requires the model to use the full context to solve the problem. To generate the data, we randomly select abstracts from ruSciBench, choose a number of repeats and an ID for the paragraph to repeat. Next, we add the remaining non-repeated paragraphs to the repeated paragraph until up to the desired context length. The resulting sequence of paragraphs is randomly shuffled. The ground truth for each sample is the number of unique paragraphs.

ruQasper The ruQasper was created by translating the Qasper²⁰ dataset from LongBench, which consists of selected samples with a long context from the original QA dataset over academic research papers (Dasigi et al., 2021). The goal of the task is to find the answer to the question in one of the parts of the article. The context for samples is drawn from scientific articles.

ruGSM100 The ruGSM100 dataset is a translation of gsm100²¹ one from L-Eval. It contains 100 math problems to be solved using Chain-of-Thought in a few-shot mode. This dataset aims to evaluate the model’s reasoning and logical skills in maths. The context for all tasks is a prompt of 16 examples with problem descriptions and answers.

Datasets BABILong, MHQA, and Passkey serve as examples of needle-in-a-haystack tasks.

3.5 Submission

To make a submission to the leaderboard, users first create a configuration file by adapting con-

¹⁹https://huggingface.co/datasets/THUDM/LongBench/viewer/passage_count

²⁰https://huggingface.co/datasets/THUDM/LongBench/viewer/qasper_e

²¹<https://huggingface.co/datasets/L4NLP/LEval/viewer/gsm100>

figs/template.ini (e.g., llama_3.1.ini) from the project’s repo to specify the model parameters. Once the config is ready, they generate predictions and run the evaluation script from the repository. Both predictions and evaluation results are saved locally. Finally, users submit their results by creating a pull request to the repository. Upon approval, the model name and its evaluation are integrated into the system, with results made available on the leaderboard.

4 Experimental setup

We evaluate 17 popular LLMs that feature long context capability, including GPT-4o²² (see Appendix C for the baseline details).

In order not to go beyond the context window we use zero-shot evaluation for all tasks, except for ruGSM100 in which the few-shot examples provided as a part of long context input. When the input length of the sample surpasses the maximum model context length, we truncate the input sequence from the right. For reproducibility, the baselines were evaluated with greedy decoding (temperature = 1.0, num_beams = 1, do_sample = False). We select the best result for each model from the two supported formats: with/without the chat template.

In addition, for each task, we fixed a natural language prompt unified for all the models (see Appendix B for the exact prompt formulation). The prompts were estimated from an empirical analysis of the tasks through a series of experiments. However, it should be noted that the benchmark methodology does not rigidly fix the prompts. Users can use their own prompts for evaluation. The choice of effective prompts requires additional research, which we leave for future work. We run all the experiments on two NVIDIA A100 GPU.

5 Results

The baseline results with respect to context length are given in Table 7 and with respect to tasks are in Tables 4, 5, and 6. Model-wise detailed results are provided in the benchmark repository. Analyzing baseline performance, we can draw the following conclusions.

Group I The tasks from this group are relatively simple, and most models pass them well within

²²GPT-4o was included via API access as the state-of-the-art model representing the upper bound for long-context capabilities.

Model Name	Passkey	MatreshkaYesNo	MatreshkaNames	PasskeyWithLibrusec	LibrusecHistory	ruGSM100	ruSciPassageCount	ru2WikiMultiHopQA
Complexity group	I	II	II	I	II	IV	IV	III
GPT-4o	100.0	79.9	58.7	100.0	99.2	84.0	37.2	58.5
GLM4-9B-Chat	100.0	68.0	47.3	100.0	82.0	8.0	7.5	48.8
LLaMA-3.1-8B-Instruct	100.0	69.5	39.8	100.0	64.8	23.0	5.6	27.8
LLaMA-3.1-8B	100.0	39.9	22.4	100.0	95.3	20.0	4.1	33.4
Mistral-Nemo-Instruct-2407	97.8	53.2	32.2	99.4	53.1	0.0	12.8	27.9
Mistral-7B-Instruct-v0.3	66.7	35.3	16.3	66.6	50.8	11.0	8.2	43.2
Phi-3-mini-128k-instruct	84.7	70.7	18.8	85.5	41.4	24.0	6.2	18.9
Mistral-7B-v0.3	66.7	32.0	10.0	66.7	68.0	9.0	0.0	41.0
LLaMA-2-7B-32K	66.7	33.4	3.4	65.5	40.6	7.0	4.7	37.2
LongChat-7B-v1.5-32k	66.5	33.4	5.9	66.0	26.6	5.0	4.8	35.2
LLaMA-3-8B-Instruct	33.3	27.3	16.6	33.3	22.7	0.0	6.5	17.7
T-lite-instruct-0.1	33.3	25.7	14.0	33.3	22.7	0.0	5.1	12.9
Saiga-LLaMA-3-8B	33.3	28.0	15.6	33.2	24.2	0.0	3.8	17.7
LLaMA-3-8B	33.3	20.2	10.0	33.3	22.7	0.0	3.3	18.4
Mistral-7B-v0.1	35.0	16.8	8.1	38.3	23.4	13.0	1.3	23.0
ChatGLM2-6B-32k	63.7	33.4	1.3	65.0	8.6	5.0	3.7	17.5
LongAlpaca	42.4	30.5	0.4	40.6	13.3	2.0	3.8	30.3

Table 4: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The best score is put in bold, the second best is underlined.

Model Name	LongContextMultiQ	ruSciAbstractRetrieval	LibrusecMHQA	ruBABILongQA1	ruBABILongQA2	ruBABILongQA3
Complexity group	III	II	III	III	III	III
GPT-4o	<u>7.8</u>	78.0	52.9	77.3	53.3	<u>27.2</u>
GLM4-9B-Chat	<u>7.8</u>	77.8	44.5	54.1	29.8	22.3
LLaMA-3.1-8B-Instruct	7.9	<u>77.4</u>	31.8	55.8	24.0	23.9
LLaMA-3.1-8B	6.0	73.6	45.3	<u>53.9</u>	<u>25.4</u>	29.6
Mistral-Nemo-Instruct-2407	5.2	65.3	29.9	54.7	17.3	16.0
Mistral-7B-Instruct-v0.3	4.8	43.6	33.6	14.3	2.8	6.0
Phi-3-mini-128k-instruct	5.2	29.3	13.8	30.5	8.8	9.0
Mistral-7B-v0.3	5.2	30.5	39.1	37.3	16.7	15.7
LLaMA-2-7B-32K	7.9	39.1	27.6	40.3	16.6	16.3
LongChat-7B-v1.5-32k	3.2	41.1	24.7	17.5	7.2	4.0
LLaMA-3-8B-Instruct	4.9	31.4	46.1	23.7	4.1	4.5
T-lite-instruct-0.1	5.2	31.2	<u>48.4</u>	21.7	14.5	8.2
Saiga-LLaMA-3-8B	4.8	31.7	45.1	25.4	4.4	6.1
LLaMA-3-8B	7.0	30.9	41.4	20.8	7.7	9.1
Mistral-7B-v0.1	4.4	28.5	34.1	21.0	7.7	9.0
ChatGLM2-6B-32k	1.2	13.6	6.8	12.2	1.5	2.5
LongAlpaca	0.8	23.5	7.8	3.8	0.3	3.5

Table 5: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The best score is bold, the second best is underlined.

Model Name	ruBABILongQA4	ruBABILongQA5	ruQuALITY	ruQasper	Overall
Complexity group	III	III	II	IV	Overall
GPT-4o	66.0	84.7	89.5	<u>23.0</u>	65.4
GLM4-9B-Chat	<u>52.8</u>	<u>70.3</u>	<u>74.1</u>	5.0	<u>50.0</u>
LLaMA-3.1-8B-Instruct	14.0	59.2	42.1	6.5	43.0
LLaMA-3.1-8B	52.1	67.9	12.0	4.3	43.6
Mistral-Nemo-Instruct-2407	12.4	45.9	67.0	24.5	39.7
Mistral-7B-Instruct-v0.3	27.6	37.6	30.6	5.4	28.0
Phi-3-mini-128k-instruct	1.0	44.1	38.8	3.5	29.7
Mistral-7B-v0.3	23.6	47.1	15.2	5.8	29.4
LLaMA-2-7B-32K	16.7	43.0	15.5	4.7	27.0
LongChat-7B-v1.5-32k	12.7	33.3	23.1	5.0	23.1
LLaMA-3-8B-Instruct	19.6	25.3	34.6	2.2	19.6
T-lite-instruct-0.1	22.3	24.4	11.0	2.7	18.7
Saiga-LLaMA-3-8B	20.3	25.2	17.9	2.5	18.8
LLaMA-3-8B	19.1	22.6	8.5	2.2	17.3
Mistral-7B-v0.1	12.4	23.2	17.3	2.5	17.7
ChatGLM2-6B-32k	0.6	8.8	49.2	2.6	16.5
LongAlpaca	0.2	29.4	44.0	2.0	15.5

Table 6: The table presents the evaluation results. **Model Name** shows the name of the model. **Complexity group** indicates the complexity groups into which the tasks were divided in Table 1. The score for each task is averaged by the context length. The **Overall** score is obtained by averaging the results over each task. The best score is put in bold, the second best is underlined.

Model Name	4k	8k	16k	32k	64k	128k
GPT-4o	76.0	70.6	67.6	61.2	55.5	53.6
GLM4-9B-Chat	<u>61.9</u>	<u>57.6</u>	<u>52.1</u>	<u>49.6</u>	<u>49.1</u>	<u>43.8</u>
LLaMA-3.1-8B-Instruct	56.1	48.5	44.7	43.8	44.5	39.1
LLaMA-3.1-8B	56.6	51.1	45.4	45.2	48.2	34.1
Mistral-Nemo-Instruct-2407	56.1	49.8	43.2	39.5	34.3	26.3
Mistral-7B-Instruct-v0.3	47.6	43.8	37.1	32.3	-	-
Phi-3-mini-128k-instruct	18.5	36.2	34.5	33.1	34.3	28.6
Mistral-7B-v0.3	50.5	45.2	39.8	36.6	-	-
LLaMA-2-7B-32K	47.0	44.6	37.3	34.7	-	-
LongChat-7B-v1.5-32k	41.5	37.3	31.7	26.9	-	-
LLaMA-3-8B-Instruct	58.0	56.1	-	-	-	-
T-lite-instruct-0.1	61.4	53.2	-	-	-	-
Saiga-LLaMA-3-8B	59.3	53.9	-	-	-	-
LLaMA-3-8B	56.1	49.5	-	-	-	-
Mistral-7B-v0.1	51.0	44.9	-	-	-	-
ChatGLM2-6B-32k	30.5	25.9	23.6	16.2	-	-
LongAlpaca	28.1	24.4	19.9	15.5	-	-

Table 7: The evaluation scores of various models across different context lengths. The columns **4k**, **8k**, **16k**, **32k**, **64k**, **128k** present evaluation scores averaged over all tasks. The best score is put in bold, the second best is underlined.

their maximum input length.

Group II MatreshkaYesNo, turns out to be the most straightforward task in the group, which all models cope with naturally. The ruQuALITY task is of medium complexity; several models achieved good scores on them. The classic QA task LibrusecHistory is effectively handled by modern models. The most complex task in this group is MatreshkaNames. For it several models (e.g., ChatGLM2-6B-32k, LLaMA-2-7B-32K) show low results for any input length.

Group III For tasks from ruBABI Long, an increase in context leads to worse results. ruBABI LongQA2 and ruBABI LongQA3 turn out to be significantly more complex than others, which coincides with BABI Long results from Kuratov et al. (2024). The length of the context plays a significant role; as it grows, the quality immediately begins to decline for all but the strongest models.

LibrusecMHQA turns out to be a complex dataset; the maximum quality of the models for solving this problem is only 52.9.

Group IV ruSciPassageCount is the most difficult task created from scratch, which all models except GPT-4o handle poorly; the result’s sensitivity to the context’s size is high. Most models fail to cope with ruQasper for complex tasks and domains and with mathematical problems from ruGSM100.

Overall, GPT-4o stands out among others, significantly exceeding its closest competitor GLM4-9B-Chat. SFT models generally perform better than the pretrained ones. In most cases, an increase in the input length negatively affects the model results on the task. In general, the results indicate that our prior division of tasks into groups

is highly correlated with their complexity.

We also compared model results in English and Russian for the 4 translated datasets. The analysis and the detailed comparison can be found in our repository due to the page limit.

6 Conclusion

The rapid development of LLM has posed new challenges in evaluating their ability to process long texts. To address this problem, we have introduced LIBRA. This benchmark evaluates LLM long context understanding abilities through 18 long-context textual tasks and enables model evaluation across various context lengths ranging from 4k to 128k.

Our contribution encompasses a benchmark methodology with open-sourced datasets of different lengths and domains, a codebase for model evaluation, and baseline solution scoring. The datasets are published under the MIT license, and the leaderboard is available on HuggingFace ²³.

Limitations

Data Representation. The texts included in the benchmark are gathered from specific domains, which might not cover the full range of Russian language usage. As a result, models may excel in benchmark tasks but struggle with texts outside these domains, limiting their generalization ability. Several datasets were created using automatic translation followed by manual adaptation. This approach was mainly chosen due to the high cost of manual data creation.

Methodology limitations. When creating the datasets, we hypothesized that synthetic augmentation of the context length of the datasets, such as LibrusecHistory, would not affect the results. Our experiments show that these tasks are pretty challenging for many models. We made this methodological assumption due to the limitations of human data annotation; it is difficult for people to read large texts and concentrate enough to create questions and search for information within them. This data creation method may result in task errors, particularly when a newly extended text fragment contains conflicting information that could impact the answer. However, we found this approach acceptable due to the increased speed and cost-effectiveness.

²³<https://huggingface.co/spaces/ai-forever/LIBRA-Leaderboard>

Long context. The benchmark focuses on evaluating long contexts, but the definition of “long context” can differ based on the application and the model. The chosen context lengths may not be ideal for all usage scenarios, and models could exhibit varying performance. In this paper, we have measured the average fertility of baseline model tokenizers on a full list of datasets from our benchmark to sample different contexts and analyzed the models’ results on our datasets across various context lengths. LMs with more parameters may inherently perform better, but this does not necessarily reflect improvements in long context understanding.

Additionally, in the present work we focused exclusively on evaluating performance with respect to context length, without considering the relative position of important information within the context. Future work should include performance evaluation on needle-in-a-haystack tasks with respect to the position of the needle along with an in-depth error analysis.

Data leakage is a critical concern for modern benchmarks because current models are trained on a significant amount of text from the Internet. Long context benchmarks are particularly risky, as their texts are based on web sources and books. This could potentially lead to data leakage. However, creating original long texts from scratch not found on the web is exceptionally costly. As a result, we use open sources to develop the benchmark, acknowledging the potential risks. Nevertheless, we firmly believe this will make a valuable contribution to the Russian community, as no long context datasets are currently available.

Ethical Considerations. The data used in the benchmark was created from open data sources. When annotating the data, we obtained transparent permission from all users and made efforts to maintain the confidentiality and anonymity of participants. As the benchmark develops, ongoing efforts are required to identify and minimize biases in the benchmark datasets and evaluation metrics. The benchmark does not currently contain the datasets covering the ethical or AI safety skill evaluation, but this is a space for future work.

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation

for long context language models. *arXiv preprint arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2024. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. **Holistic Evaluation of Language Models**. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. **MERA: A comprehensive LLM evaluation in Russian**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

- Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *arXiv preprint arXiv:2407.00402*.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.
- Gregory Kamradt. 2023. [Needle in a haystack - pressure testing llms](#).
- Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*.
- Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.
- Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and 1 others. 2021. Quality: Question answering with long input texts, yes! *arXiv preprint arXiv:2112.08608*.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.
- Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.
- Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.
- Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.
- Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*.
- Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, and 1 others. 2022. Tape: Assessing few-shot russian language understanding. *arXiv preprint arXiv:2210.12813*.
- Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *arXiv preprint arXiv:2403.12766*.
- Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.
- Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. [Towards ai-complete question answering: A set of prerequisite toy tasks](#). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track*.

A Data Annotation Details

The datasets LibrusecHistory, LibrusecMHQA, and ruSciAbstractRetrieval were created via the crowd-sourced platform.

In the LibrusecHistory, annotators were instructed to read a lengthy text and generate four questions based on the text and answer them. Guidelines were provided regarding the type of questions to ask: 1) questions should be answerable using information present in the text 2) the questions must not be about widely known information but should be related to the text 3) questions can cover various aspects such as character actions, appearance, thoughts, events, and scene descriptions 4) logical deductions are not required to answer the questions 5) Each question should have a single, clear, unambiguous answer from the text.

The design of the dataset LibrusecMHQA project follows a similar structure to LibrusecHistory, but the question criteria were more complex. In this dataset, the questions were answered by expert editors rather than through crowd-sourcing. The main distinction in the criteria for annotators is the multi-hop questions, where simply reading the sentence containing the answer is insufficient. Instead, reading at least a paragraph of 2-5 sentences, or the entire relevant fragment, is necessary to gather information and generate a complete answer.

The ruSciAbstractRetrieval was collected by crowd-sourced annotators. These annotators were asked to read a long text annotation and briefly describe the contents. The criteria for the description were as follows: 1) The description must start with the word “Describes”. 2) It must be a single sentence, which can be complex. 3) The description should not exceed 30 words, including conjunctions, particles, and prepositions. 4) It should include the main general ideas identified in the abstract but should not include details.

Training examples were available for all projects. The contributions of human annotators are amassed and stored in a manner that ensures anonymity. The average hourly compensation exceeds the minimum wage per hour in Russia. Each annotator is informed about topics that may be sensitive in the data, such as politics, societal minorities, and religion. Table 8 summarizes general details concerning the creation of the datasets via crowd-source on ABC²⁴ data labeling platform.

B Dataset Examples

This section provides examples of the task format for the benchmark datasets. The exact prompts for

²⁴<https://elementary.activebc.ru>

the benchmark are not fixed. Here, we provide prompts used in our experiments²⁵.

Passkey: *You are provided with a long text that contains the access key. Just remember the access key.*

Context: {context}

You only need to specify the access key in the response.

Question: {input}

Answer:

PasskeyWithLibrusec: *You are provided with a long text that contains the access key. Just remember the access key.*

Context: {context}

You only need to specify the access key in the response.

Question: {input}

Answer:

MatreshkaNames: *You are provided with several dialogues. Remember the names of the people and the topics they talked about.*

Context: {context}

In the answer, specify only the name of the interlocutor who spoke on the topic from the next question.

Question: {input}

Answer:

MatreshkaYesNo: *You are provided with several dialogues. Remember the names of the topics that the interlocutors talked about.*

Context: {context}

In the answer, you only need to specify 'Yes' if there was such a topic and 'No' if there was no such topic in the dialogues.

Question: {input}

Answer:

LibrusecHistory: *You are given a long text in which you need to find the answer to the question.*

Context: {context}

Find the answer in the text to the following question.

Question: {input}

Answer:

ruSciAbstractRetrieval: *Below are a few paragraphs. Determine which paragraph the short description corresponds to.*

Context: {context}

Determine which paragraph the short description corresponds to. The response must contain the paragraph number.

Question: {input}

Answer:

ruQuALITY: *You are given a long text in which you need to find the answer to the question.*

Context: {context}

You will be given several answers to the question in the text; choose only one correct one.

Question: {input}

Answer:

LongContextMultiQ: *You are given a long text where you need to find the answer to the question.*

Context: {context}

Find the answer in the text to the following question.

Question: {input}

Answer:

LibrusecMHQA: *You are given a long text where you need to find the answer.*

Context: {context}

Find the answer in the text to the following question.

²⁵All examples are presented in English for transparency and are given for illustrative purposes only to clarify the idea of a given task. The examples are not necessarily a direct translation of specific examples from the dataset. The exact prompts in their original formulation in Russian can be found in our repository <https://github.com/ai-forever/LIBRA>.

Task Name	Total	Pay Rate	Example Number	Overlap
LibrusecHistory	84\$	6.25\$/hr	32	1
LibrusecMHQA	458\$	6.25\$/hr	40	3
ruSciAbstractRetrieval	290\$	6.25\$/hr	100	3

Table 8: The details of datasets collection. **Total** is the budget spent to annotate the tasks employed for metric evaluation. **Pay Rate** is the hourly rate computed as a simple average of pay rates based on time spent annotating one row and the reward for this row. **Example Number** refers to the total number of samples processed while collecting or verifying the dataset. **Overlap** is the median number of votes per dataset sample averaged across all annotation tasks for the same dataset (if more than 1 task is provided).

Model Name	Model Type	Parameters	Max Length	Model HuggingFace link
GPT-4o	SFT	-	128k	-
GLM4-9B-Chat	SFT	9B	128k	THUDM/glm-4-9b-chat
LLaMA-3.1-8B-Instruct	SFT	8B	128k	meta-llama/Meta-Llama-3.1-8B-Instruct
LLaMA-3.1-8B	Pretrain	8B	128k	meta-llama/Meta-Llama-3.1-8B
Mistral-Nemo-Instruct-2407	SFT	12B	128k	mistralai/Mistral-Nemo-Instruct-2407
Phi-3-mini-128k-instruct	SFT	3.8B	128k	microsoft/Phi-3-mini-128k-instruct
Mistral-7B-Instruct-v0.3	SFT	7B	32k	mistralai/Mistral-7B-Instruct-v0.3
Mistral-7B-v0.3	Pretrain	7B	32k	mistralai/Mistral-7B-v0.3
LLaMA-2-7B-32K	Pretrain	7B	32k	togethercomputer/LLaMA-2-7B-32K
LongChat-7B-v1.5-32k	SFT	7B	32k	lmsys/longchat-7b-v1.5-32k
ChatGLM2-6B-32k	SFT	6B	32k	THUDM/chatglm2-6b-32k
LongAlpaca-7B	Pretrain	7B	32k	Yukang/LongAlpaca-7B
LLaMA-3-8B-Instruct	SFT	8B	8k	meta-llama/Meta-Llama-3-8B-Instruct
T-lite-instruct-0.1	SFT	8B	8k	AnatoliiPotapov/T-lite-instruct-0.1
Saiga-LLaMA-3-8B	SFT	8B	8k	IlyaGusev/saiga_llama3_8b
LLaMA-3-8B	Pretrain	8B	8k	meta-llama/Meta-Llama-3-8B
Mistral-7B-v0.1	Pretrain	7B	8k	mistralai/Mistral-7B-v0.1

Table 9: The models evaluated as baselines. **Model Type** shows whether the model is a pretrain or an SFT. **Parameters** indicate the number of model parameters in Billions. **Max Context Length** shows maximal context lengths in tokens. **Model HuggingFace Link** provides the model link on HuggingFace Hub for the open-source models.

Question: {input}

Answer:

ru2WikiMultihopQA: *The answer to the question is based on the above excerpts.*

Context: {context}

Answer the question briefly, based on the above excerpts.

Question: {input}

Answer:

ruBABILongQA1: *I'm giving you a context with facts about the location of different people. You need to answer the question based only on information obtained from the facts. If the person was in different places, use the last location to answer the question.*

Context: {context}

Answer the question as briefly as possible.

Question: {input}

Answer:

ruBABILongQA2: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first place and moves to the second place, the item remains in the first place.*

Context: {context}

Answer the question as briefly as possible.

Question: {input}

Answer:

ruBABILongQA3: *I'm giving you a context with facts about*

the location and actions of different people. You need to answer the question based only on factual information. If a person took an item in one place and went to another, that item is also in the second place. If a person leaves an item in the first place and moves to the second place, the item remains in the first place.

Context: {context}

Answer the question as briefly as possible.

Question: {input}

Answer:

ruBABILongQA4: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*

Context: {context}

Answer the question as briefly as possible.

Question: {input}

Answer:

ruBABILongQA5: *I'm giving you a context with facts about the location and actions of different people. You need to answer the question based only on factual information.*

Context: {context}

Answer the question as briefly as possible.

Question: {input}

Answer:

ruSciPassageCount: *Below are a few paragraphs. Read them and determine the number of unique paragraphs.*

Context: {context}

Determine the number of unique paragraphs. The answer must contain only one number.

Question: {input}

Answer:

ruQasper: *You are provided with a scientific article and a question.*

Context: {context}

Answer the question as briefly as possible, using a single phrase or sentence if possible. Don't give any explanations.

Question: {input}

Answer:

ruGSM100: *Examples of mathematical problems are given below. Think step by step and answer the question.*

Context: {context}

Think step by step and answer the question.

Question: {input}

Answer:

C Detailed Model Information

We evaluate 17 popular LLMs, including GPT-4o²⁶. All models except for GPT-4o are open-source. The baseline models and their specifics are presented in Table 9.

²⁶<https://chatgpt.com/>

Unpacking Ambiguity: The Interaction of Polysemous Discourse Markers and Non-DM Signals

Jingni Wu

Georgetown University
jw2175@georgetown.edu

Amir Zeldes

Georgetown University
amir.zeldes@georgetown.edu

Abstract

Discourse markers (DMs) like ‘but’ or ‘then’ are crucial for creating coherence in discourse, yet they are often replaced by or co-occur with non-DMs (‘in the morning’ can mean the same as ‘then’), and both can be ambiguous (‘since’ can refer to time or cause). The interaction mechanism between such signals remains unclear but pivotal for their disambiguation. In this paper we investigate the relationship between DM polysemy and co-occurrence of non-DM signals in English, as well as the influence of genre on these patterns. Using the framework of eRST, we propose a graded definition of DM polysemy, and conduct correlation and regression analyses to examine whether polysemous DMs are accompanied by more numerous and diverse non-DM signals. Our findings reveal that while polysemous DMs do co-occur with more diverse non-DMs, the total number of co-occurring signals does not necessarily increase. Moreover, genre plays a significant role in shaping DM-signal interactions.

1 Introduction

Identifying and understanding discourse relations is fundamental to discourse comprehension. Discourse markers (DM) such as ‘and’, ‘because’, and ‘however’ have been widely recognized as the most typical indicator of coherence relations and are also referred to as discourse connectives or cue phrases (Forbes-Riley et al., 2006). Early research focused on DMs as the sole device indicating relations, and their presence is often used to distinguish explicit and implicit relations (Webber and Joshi, 1998; Robaldo et al., 2008). In applied Natural Language Processing (NLP) they also remain the focus of research on automatic detection of discourse relation signaling, as evidenced by the series of DISRPT (Discourse Relation Parsing and Treebanking, see Braud et al. 2024) shared tasks including DM detection as a track. Since such markers come from a closed list, systems can target only these words or

phrases (Yu et al., 2019), then focus on disambiguation, with recent system scores achieving over 93% F1-scores for English (Liu et al., 2023).

However, more recent studies have shown that DMs account for only a small fraction of discourse relations, which can be signaled by *reference* (e.g. anaphora to indicate ELABORATION¹), *semantic* (antonymy to indicate CONTRAST), *lexical* (‘the next day’ can indicate temporal SEQUENCE like the DM ‘then’), *morphological* (past followed by present tense can also indicate SEQUENCE) and *graphical* cues (e.g. a question mark signaling a QUESTION relation). In this paper we follow the taxonomy of non-DM signal types proposed by Zeldes et al. (2025), which distinguishes eight major classes with a total of 45 subtypes, illustrated in Table 1. Such non-DM signals can be crucial for disambiguating otherwise ambiguous DMs, such as ‘since’, which can signal both CAUSE and temporal CIRCUMSTANCE relations. Taken together, DMs and such similar non-DM devices are referred to collectively as discourse relation *signals* (Das and Taboada, 2018a,b; Zeldes et al., 2025).

Despite extensive research on DMs and other signals individually, far less attention has been given to their interaction. Prior studies have examined the distribution of DM-signal co-occurrence and explored potential motivations from corpus-based (Das and Taboada, 2019; Crible, 2020) and experimental perspectives (Crible and Demberg, 2020; Grisot and Blochowiak, 2017). These studies have revealed that DM-signal co-occurrence is influenced by cognitive constraints and information density, and that several factors, such as the ambiguity of DMs (Crible, 2020), the semantics of discourse relations (Das and Taboada, 2019; Crible and Demberg, 2020), and genres (Crible, 2020), af-

¹Here and below we will assume discourse relation labels commonly used in Rhetorical Structure Theory (Mann and Thompson, 1988). Our definition of what constitutes anaphoric reference aligns with (Zeldes, 2022).

signal type	subtypes	example
dm	but, then, on the other hand...	[They wanted to] [but couldn't] <adversative-contrast>
graphical	colon, dash, semicolon layout items in sequence parentheses, quotation marks question mark	[Let me tell you a story :] <organization-preparation> [Introduction] <organization-heading> I. wash [2. cut] <joint-list> it rained [(and snowed a bit)] <elaboration-additional> [Did you?] <topic-question> No.
lexical	alternate expression indicative word/phrase	He agreed. [That is he said yes] <restatement-repetition> They planned a party! [That's nice/Can't wait!] <evaluation-comment>
morphological	mood tense	Go with them [I think you should] <explanation-motivation> I started an hour ago, [now I'm resting] <joint-sequence>
numerical	same count	[Two reasons.] <organization-preparation> First..
reference	comparative demonstrative / personal propositional	[I don't want it] <adversative-antithesis> I want another one . They met Kim . [This person / she was..] <elaboration-additional> They met Kim . [This encounter was..] <elaboration-additional>
semantic	antonymy attribution source lexical chain meronymy negation repetition/synonymy	Beer is cheap , [wine is expensive] <adversative-contrast> [Kim said] <attribution-positive> they would it was funny [so they laughed] <causal-result> The house was big, [the door two meters tall] <elaboration-additional> Kim danced , [Yun didn't dance] <adversative-contrast> They met Dr. Kim . [Dr. Kim/The surgeon was..] <elaboration-additional>
syntactic	infinitival/relative clause interrupted matrix clause modified head nominal modifier parallel syntactic construction past/present participial clause reported speech subject auxiliary inversion	a plan [to win] <purpose-attribute> [I meant -] <organization-phatic> I mean, a plan [to win] <purpose-attribute> articles [explaining chess] <elaboration-attribute> it's all tasty [it's all pretty] <joint-list> Kim appeared [dressed in black] <elaboration-attribute> [Kim said] <attribution-positive> that they would I would have [had I known] <contingency-condition>

Table 1: Signal types and subtypes, with examples highlighting in red the signal tokens which indicate the relation of the unit in square brackets.

fect the likelihood of co-occurrence. However, the specific mechanisms governing DM-signal interactions remain unclear. In particular, little is known about which conditions favor such co-occurrences, how different signals contribute to disambiguation and the resulting effect, what happens when conflicting signals appear, and how these patterns vary across discourse relations and genres.

While previous studies have confirmed that polysemous DMs co-occur with additional signals, there has been little systematic analysis of how different types and combinations of non-DM signals help resolve ambiguity. This study seeks to bridge this gap by analyzing the distribution, number, type, and co-occurrence patterns of signals with polysemous DMs across genres. We focus on the following research questions:

1. Are polysemous DMs accompanied by more numerous or more diverse non-DMs?
2. What are the typical combination strategies

for DM and non-DM signals?

3. Are strategies and distributions general, or are they genre-specific?

Because of their lower information content, we hypothesize that polysemous DMs will exhibit a stronger connection with non-DM prevalence. We also anticipate that different genres will exhibit distinct preferences for specific types of signals for polysemous DMs when resolving DM ambiguity, in part because they involve different prior likelihoods of certain relations. We therefore expect the relationship between DM polysemy and the number and diversity of co-occurring non-DMs to vary by genre.

2 Related Work

Previous studies have demonstrated that discourse relations are frequently signaled not just by DMs, with over 80% of signaled relations exhibiting some other textual cues, both with and without the

presence of accompanying DMs (Taboada and Das, 2013; Das and Taboada, 2018a,b). Moreover, it has been found in many cases that multiple signals indicate discourse relations simultaneously (Das and Taboada, 2018b; Webber et al., 2019). Among these, the combined use of DMs and non-DM signals is particularly common and serves to signal a wide variety of relations (Das and Taboada, 2019). For instance, in the following example from the GUM corpus (Zeldes, 2017), ‘while’ functions as a typical DM for the CONCESSION relation, which is further reinforced by a lexical chain connecting existing ‘studies of the psychology of art’ with ‘no work’, creating a contrast between previous work that exists and a gap in academic literature:

- (1) [While studies of the psychology of art have focused on ... no work has been ...] [Relation: ADVERSATIVE-CONCESSION; DM: ‘While’; Signal: semantic (lexical chain)] (File: *GUM_academic_art*)

Although this pattern is very common in academic writing, little attention has been paid to the ways in which ambiguous DMs such as ‘while’ (which can also mean *during a time that...*) resolve to a unique interpretation thanks to co-occurring signals in this manner, and the joint use of DMs and signals remains a complex question.

Non-DM signals can 1) overlap with DMs in meaning, potentially leading to redundancy, 2) co-occur with DMs but function independently (potentially signaling multiple distinct relations), and 3) may complement DMs in specific types of relations and environments (Hoek et al., 2018). Recent studies have begun to explore the underlying triggers of the *DM + other signals* phenomenon. Das and Taboada (2019) suggested that such combinations may arise from the inherent ambiguity of certain DMs which can signal various relations. For example, the DM *and* can mark additive LIST and temporal SEQUENCE relations, among other options, as illustrated in the following examples from GUM:

- (2) [I came home last night **and** told you.] [Relation: JOINT-SEQUENCE] (File: *GUM_conversation_grounded*)
- (3) [... borders of our moral **and** ethical understanding.] [Relation: JOINT-LIST] (File: *GUM_essay_ghost*)

Building on this, researchers have introduced the

concept of *marking strength* or *signaling strength* of DMs, which can be assessed by the number and frequency of discourse relations they can signal (Asr and Demberg, 2012). Zeldes and Liu (2020) proposed the *delta-softmax* metric, which quantifies prediction accuracy degradation for a trained neural model when a word is removed to estimate its signaling strength for a relation, providing empirical validation of an intuitive graded *signalyness* phenomenon. For instance, ‘but’ could be significantly less ambiguous than ‘and’ as a DM, in that removing ‘but’ would make the relation much harder to predict than removing ‘and’.

This strength directly influences how DMs interact with non-DM signals: it has been suggested that DMs tend to co-occur more frequently with other signals when indicating a wide range of discourse relations (Das and Taboada, 2019). In such cases, non-DM signals can play a disambiguation role, helping to clarify the intended relation (Cribble and Demberg, 2020). However, although patterns might be typical of specific genres, for example if formal texts prefer stronger and less ambiguous DMs, the association between DMs and other signals has not been found to vary significantly across genres in previous work (Cribble, 2020).

In addition, combinations of DMs and non-DM signals vary across relation types, but they are not necessarily driven by inherent semantics (Das and Taboada, 2019). That is to say, certain relations tend to prefer either DM-only or DM-plus-signal combinations. This is partly influenced by the inherent semantics of the discourse relations themselves (e.g., weakly connected sentences), but also appears to reflect an independent pragmatic strategy for ensuring clarity of the writer’s intention.

While prior research has qualitatively identified some factors influencing the co-occurrences of DM and non-DM signals, a systematic analysis of how specific non-DM signals interact with ambiguous DMs across relation types and genres remains underexplored. In particular, the co-occurrence patterns between ambiguous DMs and accompanying signals have not been quantitatively mapped. Using the largest sample of annotated discourse relation signals to date, this study addresses this gap by investigating 1) the types and frequencies of non-DM signals that co-occur with ambiguous DMs, 2) how these combinations vary across genres, and 3) whether certain signal combinations contribute to disambiguating the intended discourse relation.

3 Data

This study uses the Georgetown University Multilayer (GUM) Corpus which consists of 16 spoken and written, informal and formal style English text types (Zeldes, 2017) (see corpus details in Appendix A). The corpus originally contained Rhetorical Structure Theory (RST) annotations, which were recently extended based on Enhanced Rhetorical Structure Theory (eRST, Zeldes et al. 2025), adding annotated DMs and seven types of non-DM signals based on the taxonomy proposed by Das and Taboada (2018a), as well as adding multiple concurrent and tree breaking relations edges to the initial RST trees. With over 250,000 tokens, this is currently the largest dataset annotated for DMs and non-DM signals of discourse relations.

Since this study relies on accurate discourse annotations, we also report how quality was assured. Inter-annotator agreement studies on GUM showed F1 scores of 92.3 for DM identification and 90 for relation association (36 docs, 32K tokens). For non-DM signals, many types were automatically derived from gold syntax and coreference annotations, with others manually corrected or added. On a subset of documents, human-human agreement yielded an F1 of about 0.80.

4 Polysemy of DMs

The ambiguous nature of DMs arises from their one-to-many relationship with discourse relations. DMs that can signal multiple relations, such as ‘and’, are often described as *weak signals* (Asr and Demberg, 2012; Das and Taboada, 2019; Crible, 2020), as they do not map consistently to a single meaning, in contrast to unambiguous DMs such as ‘despite’, which always marks a CONCESSION.

Going beyond previous categorical approaches to such polysemy, we adopt a graded, quantifiable definition of DM polysemy by calculating the Shannon Entropy (Shannon, 1951) of DM meanings, which measures how evenly a DM is distributed across multiple discourse relations. A high entropy score indicates that a DM appears equally in multiple relations, while a lower score means that a DM is used in only one or very few types of discourse relations, or with a strong predominant sense. We expect a high entropy score here for the most polysemous DMs, for example, a high value for DMs like ‘and’ or even ‘but’, and the lowest value for DMs like ‘despite’.

Shannon Entropy is calculated by measuring the

probability of the DM appearing in each discourse relation. The polysemy score is computed as follows:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad (1)$$

x_i is the possible discourse relation signaled by a DM, n is the number of distinct relations signaled by the DM, and $P(x_i)$ is the probability of the DM signaling the relation x_i .

5 DM-Signal Co-occurrences

5.1 General Distribution

Across 16 genres, 21,435 discourse relations are annotated in our data, of which 1,372 (6.4%) are indicated by both DMs and non-DM signals. This result aligns fairly closely with Das and Taboada (2018a)’s finding for Wall Street Journal news (7.55%, see also Liu and Zeldes 2019). However as suspected, we observe substantial variation across genres (see Figure 1): *essay* (8.7%), *bio* (8.6%), and *how* (how-to guides from Wikihow, 8.5%) show a higher proportion of DM-signal co-occurrence, whereas *conversation* has the lowest proportion (3.9%).

Among the 1,372 instances of DM-signal co-occurrence, 96% are marked by DM + 1 signal or DM + 2 signals, while just 3% are marked by three to four signals. Only a handful of cases include more than five signals (see Table 2).

The most commonly used DM in co-occurrence with other signals across genres is the connective ‘and’ (36.6%), which is generally the most frequently used DM as well. Almost all genres in our corpus employ ‘and’ in DM-signal co-occurrences, except for *academic* where the conjunction ‘by’ is the most common DM favoring non-DM signal accompaniment, as in example (4), where the DM signaling the MEANS relation is accompanied by the lexical signal ‘using’:

- (4) **by using** a second order Rao and Scott (1981) ... correction

The top three most frequently used signal types in co-occurrences are *semantic*, *syntactic*, and *lexical* across genres, though different genres favor different types of signals, in part due to the format of texts. In spoken genres such as *vlog*, *conversation*, and *court*, there is a large amount of *reference* signals used along with DMs to indicate relations

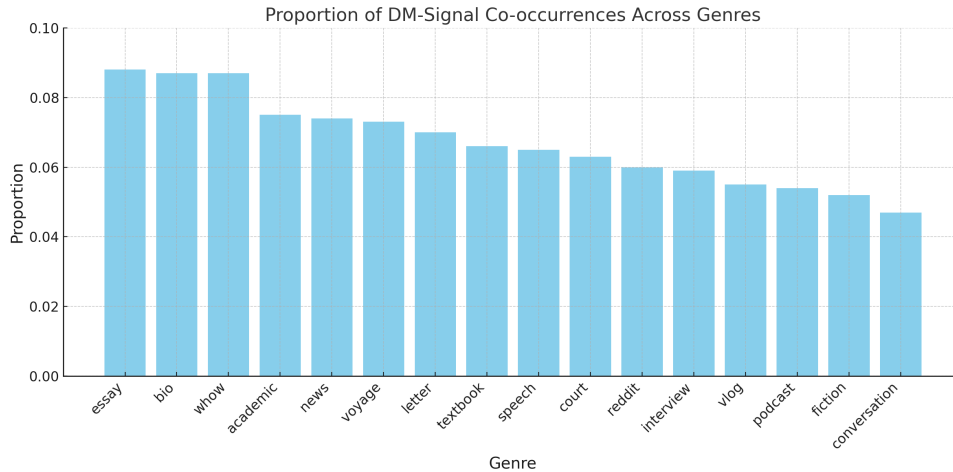


Figure 1: Proportion of DM-Signal co-occurrence across genres

	DM + 1 Signal	DM + 2 Signals	DM + 3 Signals	DM + 4 Signals	DM + 5 Signals	DM + 6 Signals	DM + 8 Signals
Total counts	1092	229	42	6	1	1	1
Proportion	79.55%	16.7%	3.1%	0.44%	0.07%	0.07%	0.07%

Table 2: Pattern of DM + signal combinations in co-occurrences

such as ELABORATION (see Figure 2). Trivially, *graphical* signals such as quotation marks to signal ATTRIBUTION cannot occur in spoken language and are restricted to written data.

5.2 Polysemous DMs and Signal Patterns

The DM ‘so’ has the highest polysemy score across all genres in our dataset, while the DM ‘for’ exhibits the most diverse range of accompanying signals (see Table 3). Here, *diversity*² refers to the number of distinct non-DM signal types that co-occur with a given DM, including individual signal types (e.g. *semantic*) and combinations of multiple types (e.g. *semantic + lexical*).

DM	non-DM signal diversity
<i>for</i>	29.50
<i>and</i>	26.64
<i>if</i>	25.00
<i>by</i>	20.80
<i>when</i>	19.00

Table 3: Top 5 DMs with the highest signal diversity

²Since DM frequency varies across genres, we normalized diversity by dividing the number of unique co-occurring signal types by the square root of total DM occurrences. This accounts for diminishing returns and prevents frequent DMs from being unfairly penalized.

This raises the question of whether more polysemous DMs tend to co-occur with a greater number of non-DM signals and exhibit more diverse signal patterns, on account of the less consistent mapping of their form to a specific meaning. To answer these questions, we employed fitted regression models to examine the relationship between DM polysemy (independent variable) and two dependent variables: (1) the total number of co-occurring non-DM signals and (2) the diversity of signal types associated with each DM.

Our results, based on both Pearson correlation and regression analyses (see details in Appendix C), suggest that polysemous DMs are more strongly associated with the diversity, rather than the quantity, of accompanying non-DM signals. While we observe a weak but statistically significant correlation between entropy and the total number of co-occurring signals ($r = 0.248, p < 0.05$), this association does not hold in a multiple regression model where both entropy score and total co-occurring signals are included as predictors of normalized signal diversity. In contrast, entropy remains a significant predictor of normalized diversity, even after controlling for signal quantity ($p < 0.001$). This supports the view that more polysemous DMs require more diverse signal patterns rather than just more signals to clarify their discourse functions.

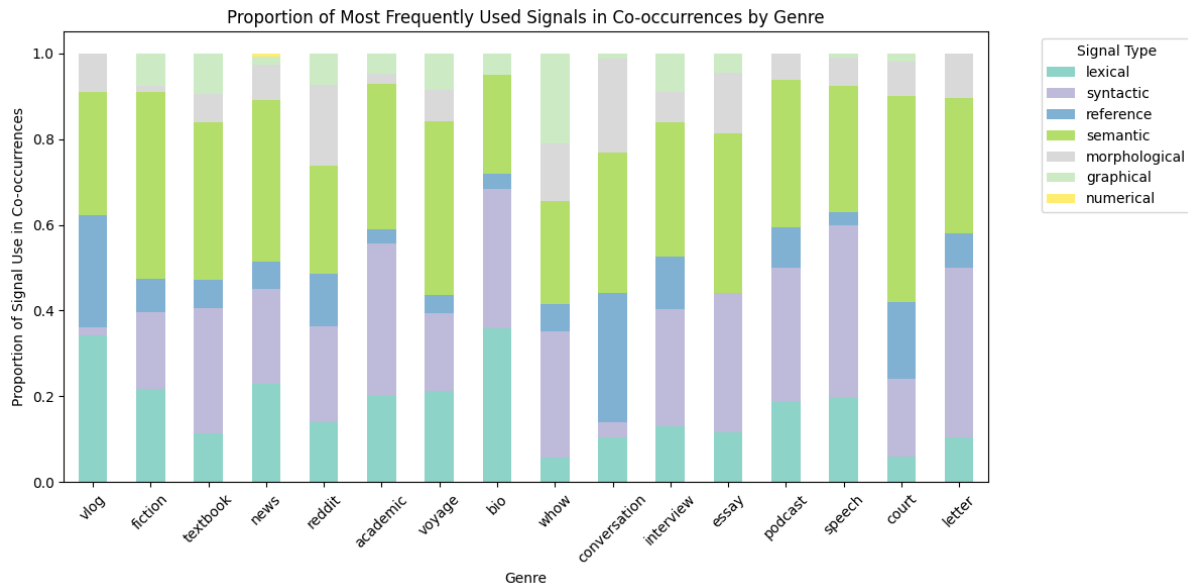


Figure 2: Proportion of most frequently co-occurring signals by genre

However, the overall explanatory power of entropy score alone is modest (adjusted $R^2 = 0.071$), suggesting that other factors may influence the relationship between DM polysemy and signal diversity. To further explore this, we considered genre as a variable. The regression model (see details in Appendix C) that includes genre and its interaction with entropy score significantly improved model fit ($p < 0.000001$, adjusted $R^2 = 0.090$), suggesting that the effect of DM polysemy on co-occurring signal patterns varies across genres. Notably, genres such as *vlogs* exhibited a significantly stronger positive relationship between DM polysemy and signal diversity, while others like *letter* showed a weaker or even negative trend. This variation highlights that the need for signal diversity in disambiguating polysemous DMs is not uniform, but shaped by genre-specific discourse norms. These genre-specific effects raise the question of what kinds of non-DM signal patterns are employed in each genre, which we address in the next section.

Looking at patterns rather than counts of signals in more detail, certain signal types consistently co-occur with highly polysemous DMs, suggesting that these signals play a crucial role in disambiguating them. For example, *lexical* and *syntactic* signals frequently appear across multiple cases and are more likely to be combined with other signal types, reinforcing their role in guiding interpretation (see Table 4).

In summary, our hypothesis is partially supported: polysemous DMs are more likely to exhibit

diverse combinations of non-DM signal, possibly due to their less stable mapping of form to meaning, but they do not consistently co-occur with a greater number of signals. Given prior evidence that signal co-occurrences vary in quantity across genres (Figure 1), we now turn to investigate the impact of genre variation, and examine the hypotheses within individual genres in the following section.

5.3 Signal Combinations and Genre Effects

According to the entropy scores, the most polysemous DMs within each genre are presented in Table 5³. Notably, the most ambiguous DMs within each genre differs from those identified as globally most ambiguous. The DM ‘and’ is the most polysemous in six genres, and DM ‘so’ and ‘as’ are the second most ambiguous DMs in eight genres. By contrast, ‘also’ is the most polysemous DM in only one genre.

The non-DM signals that co-occur with polysemous DMs exhibit diverse combination patterns, which vary across genres. A single DM may be more likely to be paired with entirely different signals depending on the genre. For example, the DM ‘and’ is most frequently used with *lexical_chain* signals (a subtype of *semantic*) signals, see example (5) in nearly all genres, except for *vlog*, *bio*, *whow*, *conversation*, and *podcast* (see Appendix Table 7). Here the lexical relation between the related items

³When comparing the polysemy across genres, we normalized the entropy score by dividing the raw entropy score by the maximum possible entropy for each DM in each genre.

DM	Top 3 co-occurring Types	Top 3 most frequent combinations
<i>so</i>	morphological, lexical, syntactic	(lexical + reference), (syntactic + reference + graphical)
<i>in</i>	syntactic, lexical	(syntactic + syntactic), (lexical + syntactic + syntactic)
<i>with</i>	semantic, graphical, syntactic	(reference + semantic), (syntactic + syntactic), (numerical + semantic + semantic)
<i>as</i>	syntactic, lexical, morphological	(lexical + semantic)
<i>and</i>	reference, lexical, semantic	(reference + graphical), (semantic + semantic), (lexical + syntactic)

Table 4: Top 5 Polysemous Discourse Markers and Co-occurring Signal Patterns

Genre	DM	Raw entropy	Normalized entropy
Court	and	2.85	0.61
Reddit	so	2.59	0.60
Conversation	and	2.63	0.58
News	as	2.55	0.57
Fiction	so	2.35	0.53
Voyage	as	2.33	0.53
Interview	and	2.34	0.52
Vlog	and	2.27	0.52
Speech	so	2.20	0.51
Wikipedia	so	2.25	0.50
Textbook	so	2.16	0.48
Podcast	and	2.15	0.48
Biography	also	1.90	0.44
Academic	as	1.92	0.44
Letter	as	1.84	0.42
Essay	and	1.64	0.40

Table 5: Entropy score of DMs per genre

‘information’ and ‘content’ forms a semantic signal next to ‘and’ to indicate that the two clauses are part of a list.

- (5) [The Penn State wiki was never proposed as a source of official information, **and** the university already hosts non-official content ...] [Relation: JOINT-LIST; DM: ‘and’; signal: semantic (lexical chain)](File: *GUM_letter_wiki*)

To further assess whether genre systematically affects the distribution of non-DM signals for polysemous DMs, we conducted Chi-Squared Goodness of Fit. For each genre, we compared the signal-type distribution to the global (genre-agnostic) distribution for the same set of DMs. After applying False Discovery Rate (FDR) correction, we found that all 16 genres show statistically significant deviations ($p_{\text{corrected}} < 0.05$), confirming that genre has a strong effect on the signaling strategies used to support polysemous discourse markers. Genres such as *vlog* and *conversation* exhibited the largest deviations, suggesting that signal use in these gen-

res is especially distinct from the overall norm.

This variation can be attributed to the nature of spoken genres such as *vlogs* and *conversations*, which emphasize audience interaction and shared common ground. In these contexts, indicative words and personal references are more commonly used to enhance engagement and coherence. Similarly, other spoken genres tend to favor *reference* signals, particularly *personal references*, using chains of pronouns to help the audience recall previously mentioned content. Semantic signals in the genre *podcast* show a particularly strong use of *meronymy*, using words in a part-whole relationship alongside the polysemous ‘and’ to indicate elaborations on complex information.

In addition, ‘and’ tends to use combined signals more frequently than other polysemous DMs, dovetailing with our initial hypothesis about non-DMs compensating for ambiguous DMs. Notably, in almost all genres where ‘and’ is the most polysemous DM, it co-occurs with multiple signals, except for the genre *essay*, where it primarily appears by itself or with a single signal type. Among all signal

combinations, the most frequent combined signal set for ‘and’ is *reference + semantic*, i.e. anaphora and lexical relations between words in the units joined by ‘and’. Interestingly, *letter* is the only genre where the most polysemous DM is ‘as’, yet it does not co-occur with any additional non-DM signals. Looking at its instances, nearly 65% are used to indicate MODE relations (manner/means), as opposed to only 32.2% in the rest of the corpus, suggesting that this usage may simply be more predictable as a default in *letters* – the most common sense in the remaining genres is indicating a temporal CIRCUMSTANCE, similarly to ‘when’.

Many DMs exhibit reduced polysemy within individual genres compared to their global scores, suggesting that their meaning is more specialized and thus less ambiguous in certain contexts. However, some DMs show substantial variation across genres, potentially requiring a greater variety or higher number of non-DM signals to aid interpretation in specific genres (see Figure 3).

To identify DMs whose polysemy varies the most across genres, we compared their normalized within-genre polysemy scores with their global polysemy scores. The top five discourse markers with the largest shifts are ‘so’, ‘in’, ‘with’, ‘given’, and ‘indeed’, which align with the overall polysemy ranking observed earlier. Highly polysemous DMs exhibit greater variance across genres, likely because their multiple meanings make them more adaptable to different discourse needs, which can be disambiguated either by non-DM signals, or simply by their use in a genre with strong priors on expected senses. In contrast to DMs with lower polysemy, which may serve more stable functions, highly polysemous DMs can shift more dramatically depending on genre-specific discourse structures, discourse relation compatibility, communicative conventions, and signaling strategies.

Among the genres, *academic*, *reddit*, and *court* seem to have larger variance, indicating that DMs used in these genres experience the most sizable shifts in polysemy compared to their global usage. These genres may have DMs that behave very differently in terms of polysemy compared to their global usage. In contrast, DMs in *fiction*, *podcast*, and *letter* appear to behave similarly locally and globally.

In addition, we examined the relationship between the number of co-occurring non-DM signals, the diversity of those signals, and the DM polysemy within each genre. Our global analysis

confirms that polysemous DMs tend to co-occur with more diverse signal patterns, however, since the frequency and variety of co-occurring signals differ across genres, we extended this investigation within individual genres to determine whether genre influences this phenomenon. The results indicate that spoken genres such as *court*, *podcast*, and *vlog*, DM polysemy strongly correlates with both the number and diversity of co-occurring non-DM signals, while other genres’ results almost align with our findings across genres, that the higher a DM’s polysemy score, the more diverse these signal combinations tend to be. This supports our hypothesis that specific genres, particularly spoken contexts and formal or unusual settings (e.g. courtroom transcripts or academic writing), adopt distinct non-DM signaling strategies which help in the disambiguation of polysemous DMs.

6 Conclusion

This study investigates the relationship between DM polysemy, the number and diversity of co-occurring non-DM signals, and the role of genre in these interactions. Our findings partially support the hypothesis that polysemous DMs exhibit more diverse non-DM signal patterns but do not necessarily co-occur with a greater number of non-DM signals. Moreover, genre greatly shapes DM polysemy, with significant variations in DM entropy and signal usage. Spoken genres (e.g. *court*, *podcast*, *vlog*) show a stronger dependence on non-DM signals to disambiguate polysemous DMs, while some written genres exhibit little or no correlation. This pattern likely reflects cognitive and interactional pressures in speech, where speakers must maintain fluency under real-time constraints (Clark, 2002) and often deploy additional cues to support coherence. Moreover, DMs in spoken discourse frequently serve interactive functions, such as managing turn-taking or structuring talk (Clark and Tree, 2002), which may further increase their co-occurrence with diverse signals.

These findings challenge theoretical views of DMs in frameworks that assume that DM marking means we do not need to consider other types of signals, such as in the Penn Discourse Treebank framework, where alternative lexicalizations marking a relation are generally only considered if a DM is absent (Prasad et al., 2018). They also suggest a consequence for treating DMs and other types of markers as categorically consistent across different

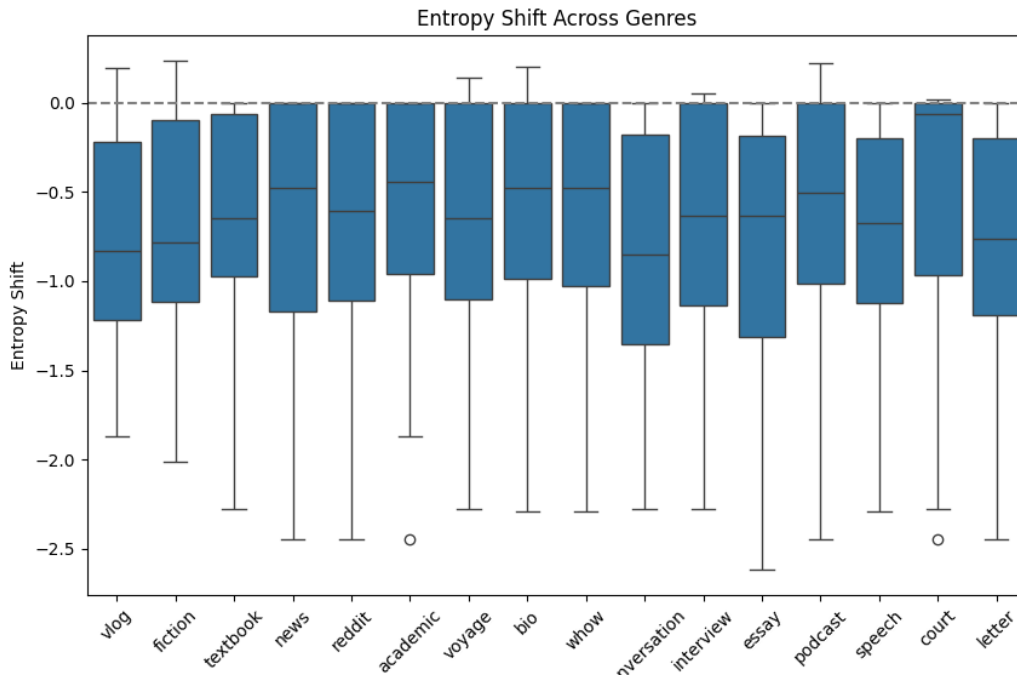


Figure 3: Entropy shift across genres

types of text: in practice, we find great variation in the extent and types of signaling present based on genre.

On the other hand, some genres exhibit little to no significant correlation among DM polysemy, the number and the diversity of non-DM signals, suggesting that different discourse contexts may impose different constraints on how DMs interact with non-DM signals. Additionally, we identified DMs whose polysemy scores are highly shifted across genres, such as, ‘so’, ‘in’, ‘with’, ‘given’, ‘indeed’, and ‘while’. This finding suggests that certain polysemous DMs are more sensitive to contextual variation, whereas others maintain stable meanings across different discourse settings. Further research is needed to understand the extent to which the picture of genre variation presented here is comprehensive, which could be carried out with new eRST data on unusual genres that has recently become available (for example in the GENTLE corpus, Aoyama et al. 2023, which includes annotations for poetry, legal writing, and more).

7 Discussion

This study does not fully account for the distribution of different discourse relations, which can further shape the observed patterns of polysemy and signal co-occurrence. Prior research has demon-

strated that certain non-DMs are more commonly used to disambiguate DMs in specific relations, such as *contrast* and *consequence* (Crible and Demberg, 2020), and different relations may vary in their sensitivity to signals, with some relations being more reliant on co-occurring non-DM cues for disambiguation. Moreover, the compatibility between DMs and specific signals may play a greater role in guiding interpretation than sheer signal frequency. Future work should therefore examine how relation type conditions the use of non-DM signals with polysemous DMs, and expand analysis to larger silver-standard multilayer corpora such as AMALGUM (A Machine Annotated Lookalike of GUM, Gessler et al. 2020), enriched with automatic annotation of discourse relations and signals, which would be less accurate, but mitigate the problem of data sparseness.

Beyond theoretical implications, these findings also have practical relevance for NLP. Current discourse parsers often treat explicit relations with DMs as straightforward, yet our results show that polysemous markers frequently rely on co-occurring signals. Incorporating such cues could improve discourse relation classification and domain adaptation, while also enhancing explainability in downstream tasks such as summarization or dialogue systems.

References

- Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. **GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation**. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.
- Fatemeh Torabi Asr and Vera Demberg. 2012. Measuring the strength of linguistic cues for discourse relations. *Proceedings of the Workshop on Advances in Discourse Analysis and Its Computational Aspects*, pages 33–42.
- Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. **DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.
- Herbert H Clark. 2002. Speaking in time. *Speech Communication*, 36(1-2):5–13.
- Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.
- Ludvine Crible. 2020. Weak and strong discourse markers in speech, chat, and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes*, 57(9):793–807.
- Ludvine Crible and Vera Demberg. 2020. The role of non-connective discourse cues and their interaction with connectives. *Pragmatics & Cognition*, 27(2):313–338.
- Debopam Das and Maite Taboada. 2018a. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149–184.
- Debopam Das and Maite Taboada. 2018b. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770.
- Debopam Das and Maite Taboada. 2019. Multiple signals of coherence relations. *Discours*, (24).
- Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23(1):55–106.
- Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. **AMALGUM – a free, balanced, multilayer English web corpus**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5267–5275, Marseille, France. European Language Resources Association.
- Cristina Grisot and Joanna Blochowiak. 2017. **Temporal connectives and verbal tenses as processing instructions: Evidence from French**. *Pragmatics & Cognition*, 24:404–440.
- Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2018. The linguistic marking of coherence relations: Interactions between connectives and segment-internal elements. *Pragmatics & Cognition*, 25(2):276–309.
- Wei Liu, Yi Fan, and Michael Strube. 2023. **HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification**. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.
- Yang Liu and Amir Zeldes. 2019. **Discourse relations and signaling information: Anchoring discourse signals in RST-DT**. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, volume 2, pages 314–317.
- William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. **Discourse annotation in the PDTB: The next generation**. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Alan Robaldo, Eleni Miltsakaki, Alan Lee, Rashmi Prasad, Nikhil Dinesh, Bonnie Webber, and Aravind Joshi. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. ELRA.
- Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.
- Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. *University of Pennsylvania*, 35:108.
- Bonnie Lynn Webber and Aravind K Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. *arXiv preprint cmp-lg/9806017*.
- Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. **GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection**. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*,

pages 133–143, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. [Can we fix the scope for coreference? problems and solutions for benchmarks beyond OntoNotes.](#) *Dialogue & Discourse*, 13(1):41–62.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. [eRST: A signaled graph theory of discourse relations and organization.](#) *Computational Linguistics*, 51(1):23–72.

Amir Zeldes and Yang Janet Liu. 2020. [A neural approach to discourse relation signal detection.](#) *Dialogue and Discourse*, 11(2):74–99.

Text type	Source	Docs	Tokens
Academic writing	Various	18	17,169
Biographies	Wikipedia	20	18,213
CC Vlogs	YouTube	15	16,864
Conversations	UCSB Corpus	15	17,932
Courtroom transcripts	Various	9	11,148
Essays	Various	9	10,842
Fiction	Various	19	17,511
Forum	reddit	18	16,364
How-to guides	wikiHow	19	17,081
Interviews	Wikinews	19	18,196
Letters	Various	12	9,989
News stories	Wikinews	24	17,186
Podcasts	Various	10	11,986
Political speeches	Various	15	16,720
Textbooks	OpenStax	15	16,693
Travel guides	Wikivoyage	18	16,515
Total GUM		255	250,409

Table 6: Overview of GUM corpus by text type.

A GUM Information

B Signal Patterns for "and" by Genre

C Regression Results

Table 7: Signal Patterns for "and" by Genre

Genre	Top 3 "and" + 1 signal		Top 3 "and" + multiple signals	
	Signal Type	Signal Subtype	Signal Type	Signal Subtype
vlog	lexical semantic reference	indicative_word lexical_chain personal_reference	reference + semantic reference + reference reference + semantic	personal_reference + lexical_chain oral_reference + propositional_reference personal_reference + synonymy
textbook	semantic semantic graphical semantic	lexical_chain meronymy semicolon lexical_chain	graphical + graphical graphical + reference semantic + semantic + semantic + semantic semantic + semantic	items_in_sequence + semicolon parentheses + personal_reference lexical_chain + lexical_chain + lexical_chain + lexical_chain lexical_chain + meronymy
reddit	lexical reference semantic	indicative_word personal_reference lexical_chain	reference + reference + semantic reference + semantic + semantic reference + semantic	personal_reference + propositional_reference + synonymy personal_reference + lexical_chain + repetition personal_reference + meronymy
academic	lexical semantic semantic	indicative_word meronymy lexical_chain	lexical + lexical graphical + graphical + semantic semantic + semantic	indicative_word + indicative_word items_in_sequence + semicolon + meronymy lexical_chain + lexical_chain
voyage	semantic lexical lexical	meronymy indicative_word indicative_word	lexical + lexical semantic + semantic lexical + lexical	indicative_phrase + indicative_word lexical_chain + meronymy indicative_word + indicative_word
bio	semantic lexical graphical	lexical_chain indicative_phrase items_in_sequence	lexical + lexical semantic + semantic semantic + semantic	indicative_phrase + indicative_word lexical_chain + meronymy lexical_chain + meronymy
whow	semantic reference reference	lexical_chain personal_reference personal_reference	graphical + semantic semantic + semantic reference + reference reference + semantic	items_in_sequence + lexical_chain lexical_chain + lexical_chain personal_reference + personal_reference personal_reference + synonymy
conversation	morphological semantic semantic	personal_reference tense lexical_chain lexical_chain	reference + semantic reference + semantic semantic + semantic graphical + lexical	personal_reference + personal_reference personal_reference + synonymy personal_reference + lexical_chain lexical_chain + meronymy
fiction	semantic lexical semantic	meronymy indicative_word lexical_chain	semantic + semantic semantic + semantic lexical + morphological	semicolon + indicative_word lexical_chain + lexical_chain lexical_chain + meronymy
news	semantic lexical semantic	meronymy indicative_phrase lexical_chain	semantic + semantic lexical + morphological semantic + semantic lexical + lexical + lexical	indicative_word + tense lexical_chain + lexical_chain indicative_word + indicative_word + indicative_word
interview	reference graphical	personal_reference semicolon	semantic + semantic + semantic semantic + semantic lexical + lexical + lexical	lexical_chain + lexical_chain + meronymy lexical_chain + synonymy indicative_word + indicative_word + indicative_word
essay	semantic semantic	lexical_chain meronymy	lexical + lexical + lexical	indicative_word + indicative_word + indicative_word
podcast	lexical reference lexical semantic	alternate_expression meronymy personal_reference indicative_word lexical_chain	reference + reference + semantic semantic + semantic lexical + lexical lexical + lexical	demonstrative_reference + personal_reference + meronymy lexical_chain + synonymy indicative_word + indicative_word indicative_word + indicative_word
speech	syntactic semantic semantic	parallel_synatactic_construction meronymy lexical_chain	reference + reference	personal_reference + personal_reference
court	reference semantic semantic	personal_reference negation lexical_chain	reference + semantic reference + reference + semantic reference + semantic reference + reference	demonstrative_reference + synonymy personal_reference + personal_reference + synonymy personal_reference + lexical_chain personal_reference + personal_reference
letter	reference semantic	personal_reference meronymy		

Table 8: Pearson Correlation: Entropy Score and Total Co-occurred Signals

Variable Pair	Correlation (r)	p-value
Entropy Score - Total Co-occurred Signals	0.248	0.0137

Table 9: Model 1: Regression of Entropy Score on Normalized Signal Diversity

	Coefficient	Std. Error	p-value
Intercept	0.850	0.040	<0.001
Entropy Score	0.112	0.039	0.005
R^2		0.081	
Adjusted R^2		0.071	
F-statistic	8.44	($p = 0.0046$)	

Table 10: Model 2: Regression of Entropy Score and Total Signals on Normalized Signal Diversity

	Coefficient	Std. Error	p-value
Intercept	0.851	0.040	<0.001
Entropy Score	0.123	0.040	0.003
Total Co-occurred Signals	-0.0005	0.0005	0.302
R^2		0.091	
Adjusted R^2		0.072	
F-statistic	4.76	($p = 0.0107$)	

Table 11: Model 3: Regression of Within-Genre Entropy and Genre Interaction on Normalized Signal Diversity

Coefficient	Coef.	Std. Error	p-value
Intercept	0.866	0.034	<0.001
Within-Genre Entropy	0.055	0.033	0.098
Entropy \times Genre[T.vlog]	0.132	0.050	0.009
Entropy \times Genre[T.letter]	-0.119	0.061	0.053
Entropy \times Genre[T.conversation]	0.075	0.047	0.116
Entropy \times Genre[T.reddit]	0.073	0.047	0.123
Entropy \times Genre[T.fiction]	0.077	0.055	0.162
Entropy \times Genre[T.speech]	-0.066	0.049	0.178
R^2		0.121	
Adjusted R^2		0.090	
F-statistic	3.97	($p < 0.000001$)	

Enhancing the Automatic Classification of Metadiscourse in Low-Proficiency Learners’ Spoken and Written English Texts Using XLNet

Wenwen Guan¹

Marijn Alta²

Jelke Bloem²

¹ Amsterdam Center for Language and Communication, University of Amsterdam

² Institute for Logic, Language and Computation, University of Amsterdam

w.guan@uva.nl, marijnalta@gmail.com, j.bloem@uva.nl

Abstract

This study aims to enhance the automatic identification and classification of metadiscourse markers in English texts, evaluating various large language models for the purpose. Metadiscourse is a commonly used rhetorical strategy in both written and spoken language to guide addressees through discourse. Due to its linguistic complexity and dependency on the context, automated metadiscourse classification is challenging. With a hypothesis that LLMs may handle complicated tasks more effectively than supervised machine learning approaches, we tune and evaluate seven encoder language models on the task using a dataset totalling 575,541 tokens and annotated with 24 labels. The results show a clear improvement over supervised machine learning approaches as well as an untuned Llama3.3-70B-Instruct baseline, with XLNet-large achieving an accuracy and F1-score of 0.91 and 0.93, respectively. However, four less frequent categories record F-scores below 0.5, highlighting the need for more balanced data representation.

1 Introduction

Metadiscourse (MD) is an essential rhetorical strategy in both speaking and writing that realizes two of the metafunctions of language proposed by Halliday (1994): the textual and interpersonal functions. MD that mainly has a textual function is used to form a cohesive and coherent text (Kopple, 1985). The textual dimension comprises transitions (e.g., *but, and, because*), frame markers (*firstly, in conclusion, the next point is ...*), code glosses (e.g., *in other words, namely, for example*), and so on. Textual MD markers often have fixed forms and consistent meanings, hence they pose relatively few challenges in automatic classification. Conversely, MD that is primarily interpersonal shows different features. Addressers use interpersonal MD to comment on the propositions and to involve the addressees in their discourse. Examples include

but are not limited to hedges (e.g., *may, probably, I’m not sure ...*), boosters (e.g., *certainly, must, I believe*), and addressing the addressees (e.g., *You may end up thinking that ..., You may ask..., Can you hear me?*). This dimension is linguistically more complex as it involves multiple syntactic classes and has fuzzy span boundaries. The complexity undoubtedly leads to difficulty in automatic classification. Previous research using supervised methods reveals the performance gap between the two broad dimensions (dos Santos Correia, 2018; Alharbi, 2016). Classification of textual MD has yielded satisfactory accuracy but classification of interpersonal MD is lacking.

Automatic MD classification has barely been studied. We are only aware of the two aforementioned SVM-based studies, where transformer-based Large Language Models (LLMs) have not been used. As LLMs encode a broader range of semantic, syntactic and contextual information due to the more complex architecture that is pretrained on various linguistic resources, this study aims to improve the state of the art in automatic MD classification with a transformer-based method.¹

The raw data used in this paper were sampled from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023). They effectively represent the diversity of natural language by containing both spoken and written English, as well as data from native speakers and learners with moderate proficiency in English. In linguistics, MD is often observed in non-native speakers’ language. It has been found that learners may use more MD markers compared to native speakers because they intentionally put more effort into linguistic and meta-linguistic matters (Ädel, 2006). However, the majority of existing work about MD has focused on proficient speakers. This

¹<https://github.com/W-Guan/Automatic-MD-annotation-with-XLNet>

study will pioneer research on MD used by learners with less advanced English proficiency.

Given that MD markers are captured in spans of varied lengths and are sensitive to context, we include a range of models in our study that are partly motivated by having a span-related training objective (SpanBERT: Joshi et al., 2020; ERNIE 2.0: Sun et al., 2020; XLNet: Yang et al., 2019), being state-of-the-art (ModernBERT: Warner et al., 2025), being a baseline for comparison (BERT: Devlin et al., 2019) and being common alternative choices (RoBERTa: Liu et al., 2019; ELECTRA: Clark et al., 2020). We also include Llama 3.3 70B-Instruct (Grattafiori et al., 2024) as an untuned decoder LLM baseline.

We find that the tuned encoder models show good performance for this task, with XLNet having the overall best accuracy and weighted F1-score, and ERNIE achieving a higher macro-F1 score, meaning better performance on low-frequency categories. Llama 3.3 70B-Instruct did not achieve adequate performance, even in a few-shot setting with per-category examples. Our models also outperform previous work, though this work used different MD categorization schemes and different kinds of text corpora.

2 Background

This section extends the concept of MD to its classification. Classification first refers to a theoretical taxonomy, which can be used when labelling the raw data. Two challenges have been identified in the theoretical classification and are anticipated to lead to difficulties in automatic classification.

2.1 Metadiscourse taxonomies

We introduced two main broad categories of MD. In linguistic practice, MD is classified into many categories but there is no uniform taxonomy. Along with the development of relevant research, scholars in the field came up with taxonomies from varied perspectives (Kopple, 1985; Crismore et al., 1993; Mauranen, 1993; Milne, 2003; Hyland, 2005; Ädel, 2006, 2010). Among them, Hyland’s (2005) taxonomy is the most commonly used one. It provides a list of discovered MD markers for English based on corpus data. However, we had concerns about adopting the taxonomy in this study. Above all, it might not be sufficient because it is built on the observations of formal written language, including textbooks, students’ writing, research articles, etc.

Thus, it may not include MD markers typical of spoken language, nor incorporate mistakes such as grammatical errors and misspellings often found in learners’ casual communication.

Ädel’s (2010) taxonomy is representative of MD in spoken language. More importantly, the identification of the categories in this taxonomy relies primarily on the functions of MD in the discourse. For example, how the topic is introduced, developed, and concluded. Nevertheless, this taxonomy also has a limitation that does not fully meet the research purposes of this study. Ädel’s (2010) taxonomy requires high explicitness of MD markers. If a text span does not contain deictic words that refer to the discourse or interlocutors, it will not be counted as MD in this taxonomy but may still be MD in other taxonomies. Therefore, we use Ädel’s (2010) taxonomy as the basis of our annotation scheme but some categories from Hyland’s (2005) taxonomy are added and adjusted. The final taxonomy we use is shown in Table 1. The dimensions of “Metalinguistic comments” and “Discourse organization” correspond to the textual MD, while the dimension “Writer-reader/speaker-listener interaction” aligns with the interpersonal MD.

The task of metadiscourse classification should not be confused with some related tasks that have been addressed in NLP. The task of dialogue act classification aims to label all of a dialogue according to their communicative function, such as ‘request’. This can also include metadiscourse acts, but standard dialogue act classification schemes have been criticized for their unsystematic annotation of metadiscourse acts (Verdonik, 2023). Dialogue act classification is a NLP task where tuned encoder models outperform autoregressive decoder LLMs, which sometimes fail to beat rule-based baselines (Qamar et al., 2025), though one study shows limited success using ChatGPT in multi-party boardgame dialogue with a four-class scheme (Martinenghi et al., 2024).

Our task is also related to epistemic stance detection, which identifies statements that mark the writer’s attitude towards the factuality of reliability of propositions (e.g. *I think that...*). Epistemic stance can be expressed through MD, but not all expressions of epistemic stance are MD markers. Epistemic stance detection also concerns propositions about others’ stances, which fall outside the scope of MD. Several categories from our taxonomy pertain to epistemic stances, specifically *Epistemic attitudes* (EPA), *Hedges* (HDG), *Boost-*

ers (BST) and *Speech act labels* (SAL). Eguchi and Kyle (2023) perform stance detection on an annotated corpus of student-written assignments as a span classification task, using the spaCy SpanCategorizer as a baseline and achieving best results with a RoBERTa-LSTM model, with results comparable to human inter-annotator agreement.

2.2 Challenges in MD classification

Two features of MD pose challenges in MD classification by human annotators and predictably also in automatic classification. Firstly, MD is highly context-sensitive. For instance, “so” functions as a MD marker when it indicates a causal relation between two clauses (Example 1.1). However, it is not a MD marker when it refers to a way something was described (Example 1.2). The ambiguity makes the identification of MD from propositional contents challenging.

Example 1.

- 1: *All people in the restaurant would be affected by smoking so it should be banned.*
- 2: *I don't think so.*

Secondly, a MD candidate may belong to more than one category. In Example 2, ‘I fully agree that’ is a MD marker to show the speaker’s attitude. Within it, the MD marker ‘fully’ is a booster.

Example 2.

I fully agree that smoking should be banned in restaurants.

To date, the annotation of MD still heavily relies on manual annotation. Research on automatic classification remains highly limited. Two relatively in-depth studies have been conducted, focusing on MD classification in academic lectures (Alharbi, 2016) and TED talk transcripts (dos Santos Correia, 2018), respectively. dos Santos Correia (2018) used Support Vector Machines (SVMs) and Conditional Random Fields (CRFs). His study classified 10 categories and F1-scores for nine categories are below 0.6. Alharbi (2016) also used SVMs for primary exploration and then improved MD classification using Continuous Bag-of-Words (CBOW) and Convolutional Neural Networks (CNNs). In his study, three out of 19 categories got F1-scores higher than 0.8. Nonetheless, the model’s predictions would not be reliable enough for further linguistic research. The biggest challenge of automatic classification lies in interpersonal MD because its

syntactic and semantic features are more inexplicit and flexible compared to textual MD. For example, the *Exemplifying* (EXP) category achieved an accuracy of over 0.8 in both studies, while *Anticipating the audience’s response* (AAR) got only 0.3 in Alharbi’s (2016) work and even lower in dos Santos Correia’s (2018) findings. We hypothesize that LLMs outperform these supervised methods.

Chan et al. (2024) also address metadiscourse using transformer models in the context of automated essay scoring. While they use a modified version of Hyland’s (2005) classification scheme for manual annotation, they only perform a token-level identification task with a binary classification scheme. They find little difference between the performance of BERT, DistilBERT and RoBERTa on the task, and focus on under/oversampling techniques and different classification algorithms such as multi-layered perceptrons and AdaBoost. While useful for automated essay scoring, this identification task has limited utility for linguists who wish to construct a metadiscourse-annotated corpus.

2.3 Selected LLMs

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019) introduced bidirectional encoding by masked language modeling (MLM) and next sentence prediction (NSP). BERT and its variants, including ModernBERT (Warner et al., 2025), SpanBERT (Joshi et al., 2020), and RoBERTa (Liu et al., 2019), consist of stacked encoders that are trained on unlabeled data to encode contextualized language representations. With a token classification head, they have been used for a wide range of token and span labeling tasks, such as named entity recognition and part-of-speech tagging. Metadiscourse classification is part of the same family of tasks.

SpanBERT is trained with a span-boundary training objective. This encourages the model to represent the relationships between tokens within a span. It predicts the entire span of tokens instead of individual tokens. This is useful for tasks that require representations of text chunks, such as our task of MD identification and classification. Although RoBERTa and ModernBERT are not pretrained with span-specific boundaries, the optimization of model architecture and training such as dynamic masking and larger datasets makes them outperform the traditional BERT model. It remains unknown if this general optimization would lead to a better performance than span-specific pretraining.

Dimension	Category	Label	Examples
Metalinguistic comments	Repairing	RPR	<i>I'm sorry...</i>
	Reformulating	RFL	<i>to put it differently...</i>
	Commenting	CMT	<i>... is a difficult question.</i>
	Clarifying	CLF	<i>I don't mean to say</i>
	Exemplifying	EXP	<i>for example</i>
Discourse organization	Managing topics	MNT	<i>I will focus on...</i>
	Organizing statements	ORS	<i>and; but; so</i>
	Providing evidentials	PED	<i>according to</i>
	Enumerating	ENM	<i>first; at last</i>
	Endophoric marking	EDP	<i>As we can see in Chapter III, ...</i>
	Previewing	PRV	<i>We will discuss...</i>
	Reviewing	RVW	<i>As I said last time, ...</i>
Writer-reader /Speaker-listener interaction	Epistemic attitudes	EPA	<i>I agree that...</i>
	Hedges	HDG	<i>perhaps; might</i>
	Boosters	BST	<i>definitely; should</i>
	Speech act labels	SAL	<i>I argue that...</i>
	Managing comprehension	MNC	<i>You know what I mean.</i>
	Managing channel/audience discipline	MCD	<i>Can you hear me?</i>
	Anticipating the audience's response	AAR	<i>You may ask...</i>
	Managing the message	MNM	<i>What I want to emphasize is...</i>
	Imagining scenarios	IMS	<i>Suppose you're giving a speech...</i>

Table 1: List of MD categories, labels, and examples

We selected two other models with pretraining objectives that have particular relevance to our task. ERNIE’s continual multi-task learning component includes two objectives relevant for our task. Firstly, there is a knowledge masking task, which requires the model to learn to predict masked spans and masked named entities rather than just tokens. Secondly, there is a discourse relation task, which relies on Sileo et al.’s (2019) discourse marker dataset to have the model predict the rhetorical relation between sentences during pretraining. This is related to metadiscourse, as discourse markers are mostly textual metadiscourse markers.

XLNet (Yang et al., 2019) is an autoregressive pretraining method of which the objectives include span-based prediction, where consecutive spans of up to five tokens are predicted rather than just single tokens. As textual MD markers are often spans, this may facilitate MD classification. However, MD spans go beyond the length of five tokens as well.

Lastly, ELECTRA (Clark et al., 2020) provides another alternative to masked language modelling pretraining by basing pretraining on a replaced token detection task. This approach is shown to outperform SpanBERT and perform similarly to XLNet on the somewhat related SQuAD benchmark

(Rajpurkar et al., 2016), where models select spans that answer questions. Therefore, we also expect good performance on our MD classification task.

In recent years, autoregressive decoder-only generative LLMs have shown impressive generalization performance on a range of NLP tasks with few-shot prompting, even to novel tasks and domains without fine-tuning. However, they show poor results in these settings on text classification tasks (Bucher and Martini, 2024), span labeling tasks such as named entity recognition (Keraghel et al., 2024) and other tasks related to ours such as dialogue act classification (Qamar et al., 2025) and implicit discourse relation annotation (Yung et al., 2024). Nevertheless, we include Llama-3.3-70B-Instruct (Grattafiori et al., 2024) as a decoder LLM baseline. We also include a SpaCy baseline.

3 Methods and data

Our data are English learners’ speaking and writing extracted from the ICNALE corpus. The selection of this corpus was initially motivated by an interest in its potential for subsequent qualitative analysis, specifically in examining learners’ use of MD. The corpus has four modules, namely spoken monologues (SM), spoken dialogues (SD), writing (WR),

Module	Texts	Tokens	Avg. Tokens
SM	999	128,989	129.12
SD	748	162,759	217.59
WR	1,100	254,064	230.97
EE	130	29,729	228.68

Table 2: The descriptive statistics of the dataset

and edited writing (EE). The extracted data concern the four modules and the groups whose first language is Chinese (Mandarin or Cantonese) and English, which includes CHN (Chinese mainland), HKG (Hong Kong), TWN (Taiwan), SIN (Singapore), and ENS (English native speakers). Random sampling was used to select half of the data for manual annotation. After sampling, the data comprises 2,977 texts totaling 575,541 tokens excluding punctuation. Table 2 shows detailed statistics.

3.1 Annotation

The annotation scheme consists of the 21 MD labels from the taxonomy in Table 1. Three additional labels pertaining to linguistic errors made by the writers and ambiguities, including Grammatical errors (ERR), Misuse (MIS), and Uncertainty (UCT), were also annotated but have been fully addressed in the gold standard corpus. Thus, they are excluded from the present experiments. Manual annotation was conducted using Prodigy (Honni-bal et al., 2024), an annotation tool for creating training and evaluation data for machine learning models. There are two annotators who have degrees in linguistics-related subjects. They were trained in the definition and classification of MD, difficult examples, and the use of Prodigy. Their annotation quality was evaluated by inter-annotator agreement (IAA) using the Cohen’s kappa coefficient (κ). This metric for pairwise agreement, which accounts for chance agreement, was computed per token rather than per span in order to allow for partial matching. Table 3 reports the IAA of the overall dataset, along with the label distribution which is visualized in Appendix 4. We observe strongly imbalanced class frequencies, which is a consequence of annotating natural language corpus data.

The macro average κ coefficient for all the MD categories is 0.79. This suggests that the majority of MD markers can be properly identified and classified. The disagreement is mainly attributed to the context-sensitive nature and fuzzy boundaries of MD. Taking ‘I think’ as an example, it is a

Label	κ	N-A1	N-A2
RPR	0.81	37	35
RFL	0.78	107	157
CMT	0.86	1,067	827
CLF	0.92	793	867
EXP	0.74	1,080	1,750
MNT	0.91	2,829	3,202
ORS	0.93	23,230	23,890
PED	0.88	2,077	2,488
ENM	0.95	4,887	5,076
EDP	0.79	152	173
PRV	0.94	788	801
RVW	0.89	783	866
EPA	0.88	15,039	14,791
HDG	0.85	9,330	9,547
BST	0.78	8,263	10,314
SAL	0.82	921	1,084
MNC	0.84	1,300	1,089
MCD	0.86	1,523	1,877
AAR	0.91	456	449
MNM	0.74	249	302
IMS	0.90	131	159
Macro Avg.	0.79		

Table 3: Pairwise Cohen’s κ coefficients of inter-annotator agreement (IAA). N refers to the number of annotated tokens in spans of the specific label, whereas A1 and A2 are annotators.

Hedge (HDG) marker when it appears at the end of a clause, but it is marked with *Epistemic attitudes* (EPA) when it starts a clause due to its neutral tone. Furthermore, fuzzy boundaries are found among EPA markers, such as ‘I agree (with)’ and ‘I agree (that)’. In this case, ‘with’ and ‘that’ should be included in the span. Label disagreements and inconsistent boundaries were resolved by discussion and the involvement of a third linguist.

We split the dataset into a 70/15/15 division for training, testing, and validation. We use splits with fixed random seeds for reproducibility and to ensure that every MD category is present in the validation and test set by re-splitting with a different fixed seed until this is the case.

3.2 Models

For the classification task, we use the base and large version of the aforementioned models: BERT, SpanBERT, ModernBERT, RoBERTa, ELECTRA, ERNIE and XLNet. We used cased models to facilitate the identification of sentence boundaries. We then tune these models on the task of MD classifi-

cation using the training portion of our corpus.

Specifically, we use these models to jointly perform the identification and classification tasks using token classifier heads tuned on the task. Predicting a span where no span was annotated is considered an incorrect prediction. In tuning, every token that is not covered by an annotated MD span in the training data, is considered an unlabeled (category ‘-’) span. For our evaluation metrics we consider both weighted F1 (also called micro F1) and macro F1, due to strong class imbalance. Macro F1 weighs all metadiscourse categories equally, including those with only a few instances. With our imbalanced dataset, this metric emphasizes performance on small categories. Weighted F1 is weighted by the frequency of the category, emphasizing performance on larger categories.

For hyperparameter tuning, we use Bayesian optimization with HyperOpt, and an Asynchronous Successive Halving scheduler to increase the efficiency of the process. We tune the learning rate $\sim \log\text{-uniform}(10^{-5}, 50^{-5})$, weight decay $\sim U(10^{-3}, 10^{-1})$, training batch size $\sim U\{4, 32\}$ and warmup steps $\sim U\{4, 32\}$. We perform the tuning with 60 sample trials, 20 initial points and using the weighted F1-score as a metric. We run the models for 25 epochs. For evaluation, we use a batch size of 16. Best obtained hyperparameter combinations can be found in Appendix G.

For the Llama-3.3-70B-Instruct baseline, we adapted the widely used GPT-NER (Wang et al., 2025) sequence generation approach for named entity recognition to the task of metadiscourse classification. Details on our few-shot prompting approach with class label explanations are described in Appendix B. We use the default temperature hyperparameter of 0.6.

For the SpaCy baseline, included to represent supervised classifiers as used in previous work, we experimented with the SpanCategorizer (spaCy, 2024) pipeline. Integrated in Prodigy, it is convenient for corpus linguists who are not proficient in programming to perform automatic annotation. The span categorizer uses Tok2Vec embeddings as features with a vocabulary of 5000 (1000 prefix, 2500 suffix), going into a Maxout Window Encoder. We used a hidden layer size of 128, four encoding layers and a max span size of 22. It was trained with a 70-30 split of data for 20 epochs with the Adam optimizer, a learning rate of 0.001 with 0.01 weight decay and 10% dropout.

4 Results

For the SpaCy baseline, we observed the effects of the class imbalance inherent in our task. Only the five most frequent labels (ORS, ENM, EPA, HDG, BST) showed non-zero performance, while other labels were not predicted. The baseline’s accuracy based solely on the non-zero values achieved 0.81 (computed per span), but drops to 0.19 when all categories are considered. This indicates that SpanCategorizer fails to generalize across all MD categories, and this result is inadequate for assisting linguists in semi-automatic MD classification. The complete results are presented in Appendix A.

The Llama-3.3-70B-Instruct baseline also performed poorly with an accuracy of 0.26, weighted F1-score of 0.41 and macro F1-score of 0.17 (Table 4). To confirm this result, we performed a follow-up experiment with the potentially more powerful GPT-4o model (Hurst et al., 2024) on 20 test documents and observed a weighted F1-score of 0.56 and a macro F1-score of 0.26, outperforming Llama-3.3-70B-Instruct but still trailing behind the fine-tuned encoder models. Further details on our decoder model baseline experiments can be found in Appendix B.

4.1 Model comparison

Table 4 shows the results of the base models on the metadiscourse classification task. In terms of accuracy and weighted F1-score, XLNet with its span-based prediction pretraining objective clearly outperforms the other models. However, its macro F1-score lags behind that of the other models, indicating that the model is relatively good at predicting common MD categories and relatively bad at predicting uncommon ones. In terms of macro F1, ERNIE-v2 shows the best performance.

In Table 5, results for the large versions of these models are shown. Results pattern similarly, with either no or very minor performance gains for most models. This indicates that the bottleneck for MD classification is in the classification head rather than in the base model, due to the relatively small amount of labeled data available.

Next, we examine the per-class performance for the best-performing XLNet-large model in Table 6. Unlike the spaCy and Llama-3.3 baselines, XLNet is able to classify all of the MD categories to some extent, even those with less than 50 labeled tokens in the test set. Nevertheless, we can observe performance issues due to class imbalance – some low

Model-base	Acc.	F1	MacroF1
BERT	0.871	0.901	0.751
SpanBERT	0.856	0.893	0.738
ModernBERT	0.858	0.897	0.752
RoBERTa	0.868	0.900	0.724
ELECTRA	0.863	0.898	0.739
ERNIE	0.870	0.904	0.766
XLNet	0.905	0.922	0.707
Llama-3.3-70B	0.257	0.409	0.168

Table 4: Evaluation results for different base models

Model-large	Acc.	F1	MacroF1
BERT	0.868	0.900	0.742
SpanBERT	0.857	0.894	0.743
ModernBERT	0.860	0.898	0.748
RoBERTa	0.869	0.901	0.741
Electra	0.846	0.890	0.754
ERNIE	0.866	0.900	0.769
XLNet	0.915	0.930	0.714

Table 5: Evaluation results for different large models

and mid-frequency categories exhibit F1-scores below 0.5. Figure 2 plots the relationship between the amount of instances in a category (frequency) versus the F1-score to visualize this pattern. We can observe a clear correlation between the two variables. All low-performing categories ($F1 < 0.88$) have 1000 tokens or less of support in the test set, reflecting the distribution in the training set.

This raises the question as to what causes these differences in performance in the lower frequency bracket. One potential explanation is lexical variability – categories that can be expressed by a larger range of words should be more difficult to classify.

4.2 Unique spans per category

An interesting property of MD categories is that some categories have far less variation than others. Textual MD markers are often grammaticalized and fixed in form, while interpersonal MD can be expressed in many ways, as discussed in the introduction. Therefore, we also examine the effect of MD variation on performance per category. Figure 3 plots each category in terms of their ratio of unique spans (variation), controlled for frequency, against the F1-score. Frequency is controlled by dividing the number of unique spans by the total amount of spans of that category, similar to how type/token ratio is computed. So, for example, the *Anticipating the audience’s response* (AAR) cate-

Label	P	R	F1	N
RPR	0.70	0.03	0.06	213
RFL	0.97	0.83	0.89	35
CMT	0.99	0.98	0.98	486
CLF	0.87	0.29	0.43	266
EXP	0.99	0.96	0.98	358
MNT	0.77	0.51	0.61	1,265
ORS	1.00	0.92	0.96	5,613
PED	0.98	0.86	0.92	623
ENM	0.92	0.92	0.92	696
EDP	1.00	0.67	0.80	15
PRV	0.68	0.33	0.44	234
RVW	0.98	0.97	0.97	289
EPA	0.99	0.99	0.99	17,130
HDG	0.99	0.93	0.96	4,358
BST	0.92	0.85	0.88	3,760
SAL	0.90	0.87	0.89	306
MNC	0.30	0.18	0.22	376
MCD	0.96	0.93	0.94	366
AAR	0.99	0.33	0.50	224
MNM	0.84	0.65	0.73	55
IMS	0.60	0.60	0.60	10
Acc.			0.91	36,678
Macro	0.83	0.67	0.71	36,678
Weighted	0.96	0.91	0.93	36,678

Table 6: Results per class for XLNet-large

gory has a ratio of 1, meaning that every instance of it uses a unique sequence of tokens.

We expect that the more varied categories are more difficult to classify, but in the top left corner we see that some of the most diverse categories also have high F1-scores, even when controlled for frequency. The *Reviewing* (RVW) and *Commenting* (CMT) categories show that even a highly variable categories can still get a high F1-score, while a category with somewhat lower variation, *Managing comprehension* (MNC) gets a lower F1-score while being similar in frequency to the aforementioned two categories. This indicates that the model has good generalization capabilities and does more than just remembering some common words for each category. We also note a pattern of more frequent categories, e.g., *Epistemic attitudes* (EPA) and *Boosters* (BST), having a lower variation ratio – this is tied to the frequency factor, as among more spans there are more likely to be repeated ones.

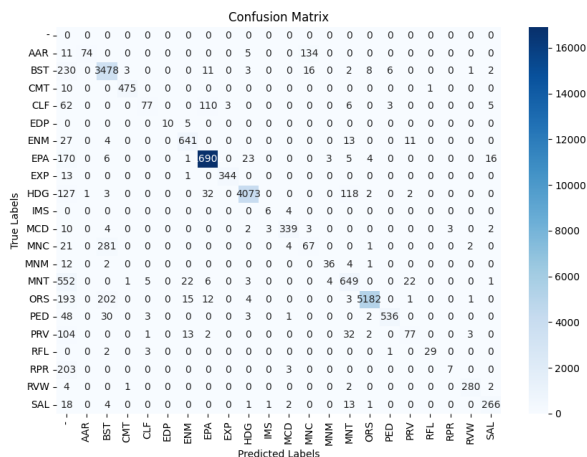


Figure 1: Confusion matrix for XLNet-large

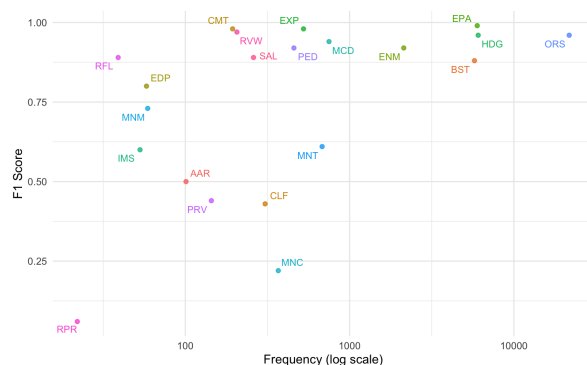


Figure 2: Relationship between label frequency and F1-score

4.3 Class ambiguity

Table 7 contains an overview of the most common misclassifications, highlighting elements from Figure 1. Such classification mixups indicate ambiguity between the categories, possibly because of similar MD spans. The overview includes misidentifications, where an annotated span is identified as not being MD — this is indicated with ‘-’. For each class, we show its frequency (N), the most common misclassification (Err1), what percentage of tokens of this class was misclassified in this way, and the second most common misclassification (Err2).

The first six categories are the relatively most commonly misclassified categories, and the correlation with category frequency is visible. For the least frequent class, *Repairing* (RPR), we see that almost all instances are misidentified. The bottom three categories are the three biggest categories. We can see that the most misclassified categories are misclassified as other uncommon categories, except *Clarifying* (CLF) which gets confused for *Epistemic attitudes* (EPA), the most common cat-



Figure 3: Relationship between the number of unique spans per category divided by frequency, and F1-score

egory. So, these confusions are more than a frequency effect.

Observed ambiguity instead appears to be caused by frequent overlapping of the two labels. They may co-occur in the same span a lot, as in our Example 2, where a *Boosters* (BST) occurs inside an *Epistemic attitudes* (EPA) indication. XLNet-large misclassifies a BST as EPA 11 times while classifying EPA as BST 6 times. Another situation caused by overlapped labels is the confusion between *Managing comprehension* (MNC) and BST. The misclassification particularly happens to the phrase ‘As we all know’, which is marked as both MNC and BST by human annotators.

Another potential cause of confusion is semantic ambiguity and similarity. For the ORS-BST confusion (*Organizing Statements-Boosters*), we observe that ORS spans that get misclassified as BST often contain words such as ‘especially’ and ‘even’, which are also used as boosters. To cite examples from the training data, the true label for ‘even still’ is BST, whereas that for ‘even if’ or ‘even though’ is ORS. The misclassification of *Anticipating the audience’s response* (AAR) as MNC is due to the same reason. Looking at some of the AAR markers (e.g., *you may say, if you ask me, you might say that*) and the MNC markers (e.g., *as you know, as you can see, you mean*), we could find all the occurrences of the personal pronoun ‘you’ refer to the readers/listeners and functions as the subject in the sentence. This semantic similarity and use of ‘you’ by both categories leads to confusion. Similarly, the confusion between *Clarifying* (CLF) and *Epistemic attitudes* is caused by the common use of ‘I’. These confusions are not observed in human annotators’ performance.

Cat	N	Err	%	Err2
RPR	213	-	95.31%	MCD
MNC	376	BST	74.73%	-
AAR	224	MNC	59.82%	-
PRV	234	-	44.44%	MNT
MNT	1,265	-	43.64%	ENM
CLF	266	EPA	41.35%	-
HDG	4,358	-	2.91%	MNT
ORS	5,613	BST	3.60%	-
EPA	17,130	-	0.99%	HDG

Table 7: Two most common misclassifications and misidentifications of categories for the 6 most misclassified categories and the 3 largest ones, by XLNet-large.

5 Discussion

Overall, we observed that tuned encoder LLM-based approaches are able to perform metadiscourse identification and classification well, clearly outperforming the spaCy baseline (weighted F1-score of 0.93 vs 0.81 only for the most frequent categories) and the Llama-3.3-70B-Instruct decoder LLM baseline (weighted F1-score of 0.93 vs 0.41). Only the encoder models are able to classify all MD categories, including the more challenging interpersonal ones. The best performance was reached with XLNet-large, but performance gains from using large versions of models were minimal.

Direct comparison to previous SVM-based work (Alharbi, 2016; dos Santos Correia, 2018) is difficult as different MD categorization schemes were used, and F1-scores were only reported per category. Broadly, dos Santos Correia (2018) reports F1-scores below 0.6 in 9 of 10 categories, while XLNet only goes below 0.6 in 5 of 21 categories. Alharbi (2016) reports 3 of 19 categories with F1-scores over 0.8, while XLNet-Large achieves this for 13 of our 21 categories. As a challenging example, the *Anticipating the audience’s response* (AAR) category got a F1-score of 0.3 in Alharbi’s (2016) work, while XLNet gets 0.5.

Improved LLM-based metadiscourse classification has promising applications. First, we can use this method to expand our annotated corpus, which is still under development, to overcome the research limitations imposed by the small amount of available data. Moreover, automatic MD classification is beneficial to language acquisition research, particularly for second language acquisition. It can help with tasks related to (automated) language assessment and language teaching. It also has po-

tential to be used in text analysis tools to provide language learners with concrete feedback on language coherence and interactionality. MD classification also provides an indication of the viability of the annotation of similar pragmatic and discursive properties of texts, in other words, linguistic items that are highly context-dependent and potentially challenging even to state-of-the-art NLP methods. Explicit MD representation may have other downstream potential for natural language processing tasks, such as in dialogue systems. It may also provide informative input features for related tasks such as stance detection, where stance can be expressed through metadiscourse, or dialogue act segmentation and classification, where some dialogue acts may be metadiscourse acts.

The practical applicability of our results is limited by poor performance on certain low-frequency classes. In future work, targeted supplemental annotation for the more challenging MD categories is the most promising next step. Another future research direction is improving generative decoder LLM performance on this task, such as through instruction tuning, prompt engineering or an agentic approach with domain-expert agents.

6 Conclusions

The model discussed in this paper achieved an accuracy and F1-score of 0.91 and 0.93 respectively on the task of metadiscourse identification and annotation, representing an improvement over related work, the SpaCy baseline and generative decoder LLM performance. Obtaining this performance with a fairly small annotated dataset shows that LLMs have potential to help speed up the annotation process even for fairly uncommon NLP tasks such as ours. The main open issue with this task is class imbalance. Annotating more conversations where these categories are present is the most promising approach to improve this.

7 Limitations

The main limitation of this project is the insufficient number of instances for the majority of categories. Annotating more conversations to make sure all categories have a significant number of instances is the most promising next step, as mentioned earlier. However, low frequency is inherent to some MD categories in natural language and small datasets with imbalanced category distributions are common in linguistic research. Alternative methods

should also be considered, for example, targeted annotating of only low-categories which would be more efficient.

A further limitation is that the study is based on a single dataset. Therefore, we cannot make claims about generalization to other learner populations or other discourse types. No other annotated MD datasets exist and it is costly to annotate them, though our results suggests that semi-automatic annotation can be performed after tuning on a relatively small dataset for future annotation of other corpora for MD.

The applicability of the method is limited by the fact that it requires tuning on MD annotation, a type of annotation that is not commonly available, especially in under-resourced languages. This limits the extent to which our method can be applied in diverse linguistic contexts and domain contexts. We were only able to demonstrate it for English. With our findings, semi-automatic annotation of more data can be explored, but only for languages where usable MD annotation already exists. Few-shot performance through in-context learning by a large generative LLM was insufficient to provide a viable alternative.

While MD spans can be embedded inside each other, the token classification approach adopted in this study can only assign one label per token. These cases are therefore not fully handled. The classifier can put a smaller span of category B inside a larger span of category A, but without the ability to assign multiple labels to one token, it is unspecified whether the category A span encompasses the category B span, or is actually two different category A spans, one before the category B span and one after.

We investigated the effect of span variation on classification performance, but our approach has some limitations. For example, if two instances of spans use the same words but in different order, or the same sequence of words but with one word omitted, they would count as different unique spans. Some sort of overlap or semantic similarity-based metric might give a better view of uniqueness of spans.

We discuss the application of our method to (semi)-automatic metadiscourse annotation. However, if it were to be used for this purpose, the annotations would be biased based on the mistakes that the classifier makes - categories that the classifier struggles with more, would be more poorly annotated. Annotators should pay particular attention to

these underperforming categories.

References

- Annelie Ädel. 2006. *Metadiscourse in L1 and L2 English*. John Benjamins.
- Annelie Ädel. 2010. Just to give you kind of a map of where we are going: A taxonomy of metadiscourse in spoken and written academic English. *Nordic Journal of English Studies*, 9(2):69–97.
- Ghada Alharbi. 2016. *Metadiscourse tagging in academic lectures*. Ph.D. thesis, University of Sheffield.
- Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.
- Sathena Chan, Manoranjan Sathiyamurthy, Chihiro Inoue, Michael Bax, Johnathan Jones, and John Oyekan. 2024. Integrating metadiscourse analysis with transformer-based models for enhancing construct representation and discourse competence assessment in L2 writing: A systemic multidisciplinary approach. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue):318–347.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. *Electra: Pre-training text encoders as discriminators rather than generators*. *Preprint*, arXiv:2003.10555.
- Avon Crismore, Raija Markkanen, and Margrat S. Steffensen. 1993. *Metadiscourse in persuasive writing: A study of texts written by american and finnish university students*. *Written Communication*, 10(1):39–71.
- 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.
- Rui Pedro dos Santos Correia. 2018. *Automatic Classification of Metadiscourse*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.
- Masaki Eguchi and Kristopher Kyle. 2023. Span identification of epistemic stance-taking in academic written English. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- M.A.K. Halliday. 1994. *An Introduction to Functional Grammar*. Hodder Arnold.
- Matthew Honnibal, Ines Montani, et al. 2024. Prodigy: An annotation tool for AI, Machine Learning & NLP. <https://prodi.gy>.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Ken Hyland. 2005. Metadiscourse: Exploring interaction in writing. *Continuum*.
- Shin'ichiro Ishikawa. 2023. *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.
- Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*.
- William Vande Kopple. 1985. Some exploratory discourse on metadiscourse. *College Composition & Communication*, 36(1):82–93.
- 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. *CoRR*, abs/1907.11692.
- Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, Dimitri Ognibene, et al. 2024. LLMs of Catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multi-party dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing@ LREC-COLING 2024*, pages 107–118. ELRA and ICCL.
- Anna Mauranen. 1993. Cultural differences in academic discourse—problems of a linguistic and cultural minority. *AFinLan vuosikirja*, pages 157–174.
- ED Milne. 2003. Metadiscourse revisited: a contrastive study of persuasive writing in professional discourse. regreso al metadiscursio: estudio contrastivo de la persuasión en el discurso profesional. *Estudios ingleses de la Universidad Complutense*, 11:29–52.
- Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation 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 3477–3486.
- spaCy. 2024. Span categorizer: Pipeline component. <https://spacy.io/api/spancategorizer>. Accessed: 2024-10-14.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.
- Darinka Verdonik. 2023. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2):144–171.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165.

A SpaCy baseline

Label	Precision	Recall	F1-score
RPR	0.00	0.00	0.00
RFL	0.00	0.00	0.00
CMT	0.00	0.00	0.00
CLF	0.00	0.00	0.00
EXP	0.00	0.00	0.00
MNT	0.00	0.00	0.00
ORS	0.89	0.88	0.88
PED	0.00	0.00	0.00
ENM	0.96	0.44	0.60
EDP	0.00	0.00	0.00
PRV	0.00	0.00	0.00
RVW	0.00	0.00	0.00
EPA	0.91	0.90	0.90
HDG	0.87	0.73	0.79
BST	0.88	0.70	0.78
SAL	0.00	0.00	0.00
MNC	0.00	0.00	0.00
MCD	0.00	0.00	0.00
AAR	0.00	0.00	0.00
MNM	0.00	0.00	0.00
IMS	0.00	0.00	0.00
Acc.			0.81
Macro (Non-zero)			0.79

Table 8: Classification accuracy of the SpaCy baseline

B Decoder LLM baseline

For the Llama-3.3-70B-Instruct baseline, we adapted the popular GPT-NER (Wang et al., 2025) prompting approach from the Named Entity Recognition (NER) task to our task. The model is prompted to apply labels in this format: *No, @@I don't think#EPA## so*. In this example, EPA is the label, @@ the span start token and ## the span end token. We prompt for the label to be generated at the end of the span due to the unidirectional nature of generative decoder LLMs. In our prompt, we added a definition of metadiscourse, the persona phrase “You are an excellent linguist.” from Wang et al. (2025), and short descriptions of all the categories with one example per category, which are the same as in Table 1. We used a three-shot setting, adding three randomly chosen full annotated documents from our training set (which was otherwise not used in the decoder LLM experiments). We also tried a zero-shot setting without the full annotated documents, but performance was very limited in these trials. The full three-shot prompt

template can be seen in Appendix F.

In Table 9, we show results per class for Llama-3.3-70B-Instruct. The overall metrics are notably lower than those for the tuned encoder models. Four categories are never labeled correctly. As might be expected in a setting without fine-tuning, per class performance does not correlate as clearly with the class’s frequency in our corpus — the rather frequent class of *Managing topics* (MNT) is never predicted by the model. The best classification performance is shown for classes that often consist of one or two words, such as *Enumerating* (ENM - *first, at last*), *Organizing statements* (ORS - *and, but*) and *Boosters* (BST - *definitely, should*). This is despite the fact that evaluation metrics are computed per token, and it suggesting difficulties in accurately marking boundaries of longer spans or in structure prediction more broadly.

As this result may be surprising to some readers, we performed an additional small-scale experiment with GPT-4o for verification. With a test set of 20 random documents (different from the example shots) and in the three-shot setting, GPT-4o achieved a weighted F1 score of 0.554 and a macro F1 score of 0.259, outperforming Llama-3.3-70B-Instruct but still underperforming compared to the tuned encoder models.

There are a few possible reasons for this poor performance. Firstly, our task is far less common than the NER task, so it is less likely to occur in training or instruction tuning data for these models. Secondly, our task has more categories than the NER task and the categories are domain-specific. They require specialized knowledge to interpret and are not frequently discussed outside of specialized literature. Even with the description and examples, the model does not seem to have an accurate representation of the categories.

Specifically, the model seems to experience interference from the more common tasks of discourse act labeling and discourse/conversation analysis. Even with the prompt stating that it’s a metadiscourse task, the model often tries to perform a discourse analysis task and interprets our categories as if they are part of such a task. For example, it marks disfluencies as RPR (*Repairing*), while only discourse about disfluencies (metadiscourse) should be marked as such. An example of this from our GPT-4o evaluation is: “*In university, most – most @@students#RPR## first goal is to study*“. This labeling is likely triggered by the repetition of *most*, but it is a repetition, not a repair (if we are do-

Label	P	R	F1	N
RPR	0.07	0.50	0.13	213
RFL	0.16	0.44	0.24	35
CMT	0.01	0.04	0.02	486
CLF	0.15	0.06	0.08	266
EXP	0.34	0.55	0.42	358
MNT	0	0	0	1,265
ORS	0.58	0.45	0.51	5,613
PED	0	0.02	0.01	623
ENM	0.70	0.58	0.63	696
EDP	0	0	0	15
PRV	0.04	0.05	0.04	234
RVW	0.19	0.30	0.23	289
EPA	0.30	0.40	0.34	17,130
HDG	0.46	0.20	0.28	4,358
BST	0.74	0.27	0.40	3,760
SAL	0.06	0.13	0.08	306
MNC	0.04	0.04	0.04	376
MCD	0.40	0.02	0.03	366
AAR	0.04	0.06	0.05	224
MNM	0	0	0	55
IMS	0	0	0	10
Acc.			0.26	36,678
Macro	0.20	0.20	0.17	36,678
Weighted	0.46	0.37	0.41	36,678

Table 9: Results per class for Llama-3.3-70B-Instruct

ing discourse analysis), and it is not metadiscourse about a repair. We also observe another issue in this example, which is that the label is applied to the word after the actual repair (most). This is likely due to the unidirectional nature of the model.

C Label frequency distribution

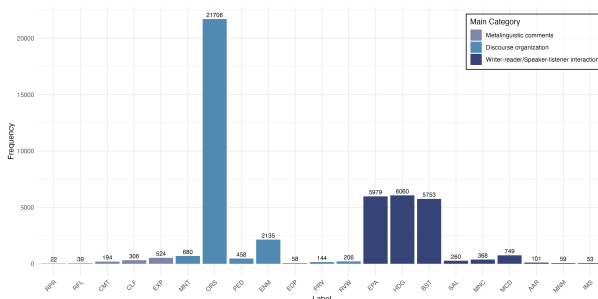


Figure 4: The distribution of MD labels in the gold-standard dataset

D Software specifications

Python: 3.11.4
numpy: 2.2.2

torch: 2.6.0+cu124

transformers: 4.48.2

All models apart from GPT-4o are available on HuggingFace:

google-bert/bert-base-cased

google-bert/bert-large-cased

SpanBERT/spanbert-base-cased

SpanBERT/spanbert-large-cased

answerdotai/ModernBERT-base

answerdotai/ModernBERT-large

FacebookAI/roberta-base

FacebookAI/roberta-large

google/electra-small-discriminator

google/electra-large-discriminator

nghuyong/ernie-2.0-base-en

nghuyong/ernie-2.0-large-en

xlnet/xlnet-base-cased

xlnet/xlnet-large-cased

meta-llama/Llama-3.3-70B-Instruct

The code can be found on this paper’s associated GitHub page: <https://github.com/W-Guan/Automatic-MD-annotation-with-XLNet>. The dataset is available upon request, as it was collected for another study that has not been published yet. Interested parties may contact the author directly to obtain access.

E Hardware specifications

GPU: NVidia L4

GPU Memory: 24GB

CPU: AMD 9445P

Total Number of Cores: 64

Memory: 384 GB

One model tuning run of 25 epochs takes about 30 minutes (ModernBERT-base) to 1 hour (XLnet-base) on this hardware.

F Prompt template

You are an excellent linguist. The task is to label metadiscourse spans - spans of words that guide the addressee through the discourse. Here are the possible categories that you can label spans as:

RPR: Repairing - Example: "I'm sorry..."
RFL: Reformulating - Example: "to put it differently..."
CMT: Commenting - Example: "... is a difficult question"
CLF: Clarifying - Example: "I don't mean to say"
EXP: Exemplifying - Example: "for example"
ORS: Organizing statements - Example: "and", "but", "so"
PED: Providing evidentials - Example: "according to"
ENM: Enumerating - Example: "first", "at last"
EDP: Endophoric marking - Example: "As we can see in Chapter III, ..."
PRV: Previewing - Example: "We will discuss..."
RVW: Reviewing - Example: "As I said last time, ..."
EPA: Epistemic attitudes - Example: "I agree that..."
HDG: Hedges - Example: "perhaps", "might"
BST: Boosters - Example: "definitely", "should"
SAL: Speech act labels - Example: "I argue that..."
MNC: Managing comprehension - Example: "You know what I mean."
MCD: Managing channel/audience discipline - Example: "Can you hear me?"
AAR: Anticipating the audience's response - Example: "You may ask..."
MNM: Managing the message - Example: "What I want to emphasize is..."
IMS: Imagining scenarios - Example: "Suppose you are giving a speech..."

If a span should be labeled, you will annotate in the format:
@@span#LABEL##

Do not output any other text apart from the annotated input text.

Document-level examples:

```
{example1}  
{example2}  
{example3}
```

Figure 5: Prompt template for decoder LLM baseline experiment

G Hyperparameters

Model	Epochs	Batch	Alpha	Decay	Warmup steps
BERT	25	6	3.46893934104582e-05	456	0.06331306513883898
SpanBERT	25	9	1.1235689536407466e-05	583	0.03328240000998865
ModernBERT	25	4	4.27605191279545e-05	352	0.07325555301079804
RoBERTa	25	17	4.1627214448844214e-05	419	0.03135747617843722
Electra	25	4	3.8988959827328105e-05	566	0.09975738929202359
ERNIE	25	14	2.7576890378412467e-05	204	0.016803254034765108
XLNet	25	4	3.784413058653172e-05	550	0.026749712474382164

Table 10: Best hyperparameters settings for base models.

Entity Tracking in Small Language Models: An Attention-Based Study of Parameter-Efficient Fine-Tuning

Sungho Jeon

Samsung Electronics
s4.jeon@samsung.com*

Michael Strube

Heidelberg Institute of Theoretical Studies
michael.strube@h-its.org

Abstract

The ability to track entities is fundamental for language understanding, yet the internal mechanisms governing this capability in Small Language Models (SLMs) are poorly understood. Previous studies often rely on indirect probing or complex interpretability methods, leaving a gap for lightweight diagnostics that connect model behavior to performance. To bridge this gap, we introduce a framework to analyze entity tracking by measuring the attention flow between entity and non-entity tokens within SLMs. We apply this to analyze models both before and after Parameter-Efficient Fine-Tuning (PEFT). Our analysis reveals two key findings. First, SLMs’ attentional strategies vary significantly with text type, but entities consistently receive a high degree of focus. Second, we show that PEFT – specifically QLoRA – dramatically improves classification performance on entity-centric tasks by increasing the model’s attentional focus on entity-related tokens. Our work provides direct evidence for how PEFT can refine a model’s internal mechanisms and establishes attention analysis as a valuable, lightweight diagnostic tool for interpreting and improving SLMs¹.

1 Introduction

A fundamental aspect of natural language understanding is the ability to track entities as a discourse unfolds (Grosz et al., 1995). This ability is a prerequisite for maintaining coherence, performing complex reasoning, and succeeding in a wide array of downstream natural language processing (NLP) tasks. For language models to generate coherent text or answer questions accurately, they must implicitly recognize entities and update their states based on new information (Grosz and Sidner, 1986).

*This work was conducted while Sungho Jeon was at Heidelberg Institute of Theoretical Studies.

¹Our code is available at https://github.com/sdeval4/codi25_entity_attn_tracking_slm

Despite the remarkable capabilities demonstrated by modern Large and Small Language Models (LLMs and SLMs) (Brown et al., 2020), the internal mechanisms by which these models manage and track entities remain largely unexplained (Li et al., 2024), especially in SLMs, which are often deployed for efficiency and on-device AI. These models are often treated as “black boxes”. While SLMs may replicate human-like output behavior, it is not clear whether they rely on linguistically grounded cues—such as noun phrases—or whether their performance stems from spurious correlations learned during pretraining.

Efforts to interpret model behavior typically fall into one of three categories: (i) evaluating input-output behaviors on benchmark tasks (Schuster and Linzen, 2022; Kim and Schuster, 2023), (ii) probing hidden state representations to see if they encode entity information (Loáiciga et al., 2022), or (iii) modifying architectures to better handle discourse entities (Fagnou et al., 2024). While these approaches provide valuable insights, they often leave a gap. They either do not directly inspect the internal mechanisms of standard architectures or they require complex, computationally intensive analysis. A direct, lightweight method for analyzing how the native attention mechanism facilitates entity tracking, particularly in the widely used Transformer architecture, is less explored.

This paper addresses this gap by proposing a novel framework to investigate entity tracking through the lens of attention scores. We treat attention as a direct, interpretable signal of the model’s focus during processing (Section 2.2). Our central hypothesis is that the allocation of attention to entity tokens is a direct correlate of a model’s entity tracking capability and that performance improvements from fine-tuning can be explained by specific, measurable shifts in these attention patterns. By systematically analyzing the attention scores between entity tokens and their surrounding

context, we aim to build a mechanistic bridge between an observable performance change and an internal model behavior.

This research makes the following contributions:

- A systematic analysis of entity-centric attention patterns in several modern SLMs, revealing how attentional strategies adapt to different text types and qualities.
- A key finding that Parameter-Efficient Fine-Tuning (PEFT) with QLoRA (Dettmers et al., 2023) substantially improves performance on an entity-centric classification task by mechanistically intensifying the model’s attention on entity tokens.
- A demonstration of attention analysis as a valuable and accessible diagnostic tool for understanding and explaining the effects of fine-tuning on a model’s internal mechanisms.

2 Related Work

Our work is situated at the intersection of three active research areas: entity tracking in language models, the use of attention for interpretability, and the mechanistic understanding of fine-tuning.

2.1 Probing Entity Representations in Language Models

The study of how language models manage entities has evolved from linguistic tests to sophisticated analyses of internal model states. Earlier work identified significant challenges, showing that even large models struggle with fundamental aspects of discourse, such as recognizing when a new entity is introduced (Schuster and Linzen, 2022). Subsequent research shifted from model outputs to internal representations, finding a disconnect between a model’s latent knowledge of entities and its ability to apply it effectively (Loáiciga et al., 2022). More recent work has created benchmarks to test dynamic entity tracking, discovering that this ability can be taught via fine-tuning (Kim and Schuster, 2023). Other studies propose architectural changes to better handle dynamic entity tracking (Fagnou et al., 2024). Unlike these prior approaches, our framework interprets entity tracking behavior through the model’s native attention weights, which directly reflect token-level interactions in Transformer models.

2.2 Attention as an Interpretability Tool

The attention mechanism, introduced as the core component of the Transformer architecture, was initially proposed as a window into the model’s reasoning process. Early work suggested that visualizing attention weights could serve as a proxy for interpreting model decisions. However, this view was contested by a line of research arguing that “attention is not explanation” (Jain and Wallace, 2019; Serrano and Smith, 2019). These studies demonstrated that attention weights could be manipulated without significantly affecting model output, suggesting they might be a symptom of the model’s reasoning rather than its cause.

Nevertheless, more recent work has revealed that specific attention heads often specialize in meaningful linguistic functions, including syntactic relations and coreference resolution (Clark et al., 2019). This suggests that attention, when interpreted systematically, offers insight into the model’s internal processing. Rather than treating attention as a complete explanation, we adopt a pragmatic perspective: we use it as a measurable correlate of focus, with a particular emphasis on how attention is distributed over discourse entities. In doing so, we aim to reconcile the interpretability of attention with its utility as a diagnostic signal.

2.3 Mechanistic Insights into Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2022) enable the adaptation of large pretrained models to specific tasks by modifying only a small subset of parameters. While PEFT methods are valued for their efficiency and scalability, their effect on the internal computations of language models has only recently begun to receive systematic attention.

Recent work attempts to reverse-engineer fine-tuned models using circuit analysis and other mechanistic tools (Wang et al., 2023; Prakash et al., 2024). These studies identify sub-network pathways responsible for specific behaviors, but their analyses are computationally intensive and often require considerable expertise. In contrast, our framework uses attention interactions to trace the effects of PEFT on entity focus directly. We show that LoRA fine-tuning leads to measurable shifts in attention toward entity tokens, which correspond to improved task performance. Our approach is both

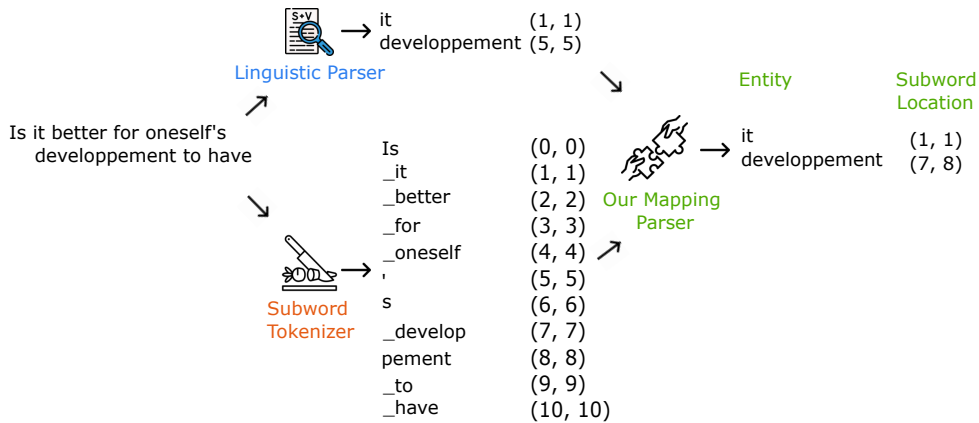


Figure 1: This example illustrates tokenization and mapping to noun phrases using the subword tokenizer of “google/gemma-2-2b-it”. The misspelling of “developpement” results in two subword tokens, “develop” and “pement”, a phenomenon commonly observed in real-world data.

computationally efficient and grounded in linguistic theory, making it suitable for broader adoption in small-model development and evaluation.

3 A Framework for Analyzing Entity-Centric Attention

In this section, we introduce a lightweight, linguistically motivated framework to analyze how Small Language Models (SLMs) allocate attention to entities during text processing. Our method is grounded in the assumption that coherent language understanding involves selectively focusing on salient discourse elements—primarily noun phrases—while integrating relevant context. We capture this behavior by systematically quantifying attention flows between entity and non-entity tokens.

3.1 Identifying and Mapping Entity Tokens

A primary challenge in analyzing the internal processing of linguistic phenomena is the discrepancy between human-readable words or phrases and the subword tokens that models actually operate on (Table 5). To bridge this gap, our framework employs a two-stage mapping process: first identifying linguistic units, then mapping them to model tokens.

3.1.1 Noun Phrase Extraction

For the purposes of this study, we define an “entity” as a noun phrase. This simplification provides a consistent and scalable method for identifying key subjects and objects across a large corpus. We use the Stanza constituency parser², which segments input texts into syntactic constituents and

²<https://stanfordnlp.github.io/stanza/>

extracts noun phrases based on their syntactic labels. We impose a constraint that limits the length of extracted noun phrases to a maximum of four words to reduce structural complexity and exclude deeply nested constructions. In cases of nested noun phrases, we retain only the outermost phrase to maintain consistent granularity across analyses.

3.1.2 Tokenization and Mapping

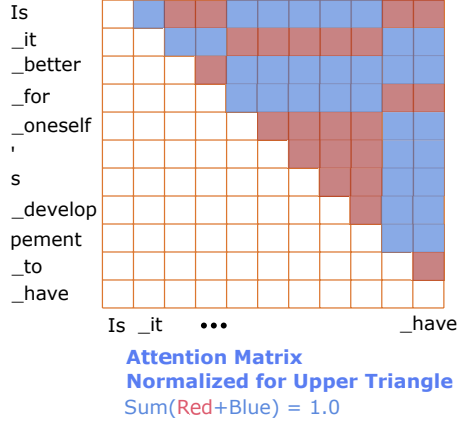
Once noun phrases are identified, we must align them with the subword tokens generated by the target model’s tokenizer. This alignment is a non-trivial task, as tokenization schemes like Byte-Pair Encoding (BPE) can fragment single words and handle whitespace in model-specific, often proprietary ways. To address this, we developed a mapping algorithm which uses the character-level spans of each noun phrase to identify all subword tokens within its boundaries. As illustrated in Figure 1, this process creates a definitive mapping from each linguistic entity to a set of token indices, a critical step that enables our subsequent analysis of attention flow.

3.2 Quantifying Attention Flow Across Linguistic Boundaries

With entities mapped to tokens, we can now quantify how the model allocates attention with respect to these linguistic categories (Figure 2).

3.2.1 Attention Score Extraction

We extract attention values from the final layer of the model’s Transformer architecture. This layer is chosen because it represents the culmination of the model’s processing, where representations are expected to be the most semantically rich and task-



$$Type1 = \text{Ratio}_{E-NE} = \sum_{t_a \in NP_{all}, t_b \in NonNP_{all}} (\bar{A}(t_a, t_b))$$

$$NP_{all} = \{ _it, _oneself, ', s, _develop, pement \}$$

$$NonNP_{all} = \{ Is, _better, _for, _to, _have \}$$

$$\begin{aligned}
 \text{Type1} &= \text{Sum}(\text{Blue}) \\
 &= A(_it, Is) + A(_it, better) + \dots \\
 &\quad + A(_oneself, Is) + \dots \\
 &\quad + A(', Is) + \dots \\
 &\quad + A(s, Is) + \dots \\
 &\quad + A(_develop, Is) + \dots \\
 &\quad + (pement, Is) + \dots
 \end{aligned}$$

Figure 2: Example of calculating Attention Type 1: between entities and non-entities. The word “developpement” is a typo found in a real TOEFL dataset, and it causes the subword tokenizer to split it into multiple subword tokens.

relevant. While individual attention heads may specialize in different functions (Clark et al., 2019), we average the attention scores across all heads in the final layer to obtain a holistic measure of the model’s aggregate focus. This provides a robust, high-level signal of information flow. The raw attention scores are normalized via softmax for each query token. The final averaged attention score between a query token t_a and a key token t_b is calculated as:

$$\bar{A}_{L_{last}}(t_a, t_b) = \frac{1}{|H|} \sum_{h \in H} A_{L_{last}, h}(t_a, t_b) \quad (1)$$

where L_{last} denotes the last layer, H is the set of all attention heads, and $A_{L_{last}, h}(t_a, t_b)$ is the attention score from token t_a to token t_b in the last layer and head h . Additionally, we investigate the different attention interaction patterns across various layers in our evaluation to provide a more comprehensive understanding.

3.3 Analysis of Attention Score Interactions

Using the extracted attention values and the LLM tokens that match noun phrases, we measure three distinct types of interactions. This helps us understand how the LLM processes context. For each interaction type, we define which tokens are involved and how we combine their attention scores.

In terms of formulation, for any input text, let N be the total number of LLM tokens: $T = \{t_1, t_2, \dots, t_N\}$. Let NP_k be the LLM tokens for the k -th noun phrase: $NP_k = \{t_{i_1}, t_{i_2}, \dots, t_{i_m}\}$. Let NP_{all} be the set of all tokens that are part of any noun phrase (entity): $NP_{all} = \bigcup_k NP_k$. Let $NonNP_{all}$ be the set of all tokens that are not part of any noun phrase (non-entities): $NonNP_{all} =$

$T \setminus NP_{all}$. The attention score from token t_a to token t_b is written as $\bar{A}_{L_{last}}(t_a, t_b)$. As detailed in Section 3.1, we focus on the last layer.

When investigating the interactions between different tokens, we focus on unique pairs of elements, effectively excluding self-attention (diagonal elements) and avoiding duplicate pairs (e.g., considering (t_a, t_b) and (t_b, t_a) as a single interaction). This is conceptually equivalent to considering only the upper triangle of the attention matrix and summing the attention for each unique pair. Each interaction type captures the ratio of attention taken by these specific pairs of tokens, normalized by the total amount of attention values between all distinct pairs of tokens in the sequence. Let $Attn_Total$ be the sum of all attention values between distinct token pairs in the sequence, considering both directions for each unique unordered pair:

$$\text{Attn_Total} = \sum_{t_x \in T} \sum_{t_y \in T, t_x < t_y} (\bar{A}(t_x, t_y)) \quad (2)$$

Our analysis focuses on three specific types of interactions: 1) between entities and non-entities, 2) between tokens of entities, and 3) between tokens of non-entities. This structured approach allows us to isolate and quantify specific linguistic phenomena, providing insights into how LLMs encode and leverage different types of relationships.

3.3.1 Type 1: Between entities and non-entities

This quantifies the attention flow between any subword token identified as part of an entity and any subword token identified as a non-entity. This captures how entities interact with their broader non-entity context within the sentence.

We calculate the average attention where tokens are from NP_{all} and tokens are from $NonNP_{all}$, then normalize by $TotalAttention$.

$$\text{Ratio}_{E-NE} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NP_{all} \\ t_b \in NonNP_{all}}} (\bar{A}(t_a, t_b)) \quad (3)$$

3.3.2 Type 2: Between tokens of entities

This measures the ratio of attention among subword tokens within the collective set of all entities, relative to the total attention in the sequence. It reflects the internal coherence and interconnectedness of all identified entities in the text.

We calculate the sum of attention between distinct tokens within NP_{all} , considering both directions for each unique unordered pair, then normalize by $TotalAttention$.

$$\text{Ratio}_{E-E} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NP_{all} \\ t_b \in NP_{all}, t_a < t_b}} (\bar{A}(t_a, t_b)) \quad (4)$$

3.3.3 Type 3: Between tokens of non-entities

This quantifies the ratio of attention among subword tokens within the collective set of all non-entities, relative to the total attention in the sequence. It reflects the internal coherence and interconnectedness of the non-entity context.

We calculate the sum of attention between distinct tokens within $NonNP_{all}$, considering both directions for each unique unordered pair, then normalize by $TotalAttention$.

$$\text{Ratio}_{NE-NE} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NonNP_{all} \\ t_b \in NonNP_{all}, t_a < t_b}} (\bar{A}(t_a, t_b)) \quad (5)$$

4 Experimental Setup

We evaluate our attention-based analysis framework in the context of two representative classification tasks using Small Language Models (SLMs). Our goal is to examine how SLMs allocate attention over entity and non-entity tokens across different discourse settings and how this distribution changes under Parameter-Efficient Fine-Tuning (PEFT). This section describes the datasets, models, and evaluation metrics used in our experiments.

4.1 Datasets

To ensure generalizability across different textual domains and discourse structures, we select two datasets that differ markedly in length, coherence structure, and task type.

- **SST-5 (Stanford Sentiment Treebank):** A benchmark for fine-grained sentiment analysis, consisting of 11,855 individual movie review sentences³ (Socher et al., 2013). The task involves assigning one of five sentiment labels: “very negative” to “very positive”. These short texts typically contain a small number of entities, often representing film titles or actors. Thus, SST-5 enables us to analyze attention patterns when entity information is concentrated in compact, sentiment-focused utterances.
- **TOEFL11:** A dataset for proficiency-level classification, composed of essays written by English language learners (Blanchard et al., 2013). Each essay is labeled with a language proficiency score (low, medium, or high). With an average length of over 400 words, the dataset provides a setting for analyzing long-form discourse. The essays include multiple entities and exhibit varied discourse organization, making it suitable for studying attention flow over extended contexts.

4.2 LLM Models for Evaluation

We perform our experiments on a representative set of modern, instruction-tuned SLMs available via the HuggingFace Hub. These models vary in parameter size, tokenizer behavior, and pretraining objectives, offering a diverse testbed for our attention analysis:

- google/gemma-2-2b-it
- meta-llama/Llama-3.2-1B-Instruct
- meta-llama/Llama-3.2-3B-Instruct
- microsoft/Phi-3.5-mini-instruct
- Qwen/Qwen2.5-1.5B-Instruct

To evaluate the effects of fine-tuning, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient tuning technique. LORA introduces trainable low-rank matrices into the model’s

³<https://huggingface.co/datasets/SetFit/sst5>

attention projections while keeping the original model weights frozen. We fine-tune each model on the SST-5 dataset using LoRA and analyze the resulting changes in both performance and attention allocation.

Hyperparameters used during fine-tuning (e.g., rank, learning rate, and epochs) are listed in Appendix A.2. All fine-tuning experiments are conducted using a consistent setup across models to ensure comparability.

4.3 Evaluation Metrics

We assess our framework using both interpretability metrics derived from attention interactions and standard performance metrics for classification.

- **Attention Analysis:** The core of our interpretability analysis relies on the three attention interaction ratios (Ratio_{E-NE} , Ratio_{E-E} , and Ratio_{NE-NE}) defined in Section 3.3. These values quantify the internal focus of the model and allow us to track systematic shifts in attention behavior across datasets and tuning conditions.
- **Classification Performance:** We measured model performance on the SST-5 test set using standard metrics for multi-class classification: Accuracy, Linear Weighted Kappa (κ_L), and Quadratic Weighted Kappa (κ_Q). Kappa scores are particularly important as they correct for agreement that could occur by chance and are sensitive to the ordinal nature of the sentiment labels (e.g., misclassifying “positive” as “very positive” is less of an error than misclassifying it as “negative”).

5 Evaluations

Our evaluation proceeds in three stages. First, we analyze baseline attention patterns in SLMs across different textual domains. Second, we examine how attention patterns vary with text granularity and writing quality. Finally, we investigate the impact of Parameter-Efficient Fine-Tuning (PEFT) using LoRA on both performance and attention allocation. Our findings demonstrate that entity-centric attention is a consistent and informative signal for tracking discourse focus and that LoRA fine-tuning meaningfully enhances this behavior.

For SST-5, we treat each review as a single unit, as reviews are typically single sentences. For TOEFL11, we analyze each sentence independently rather than encoding entire essays, allowing

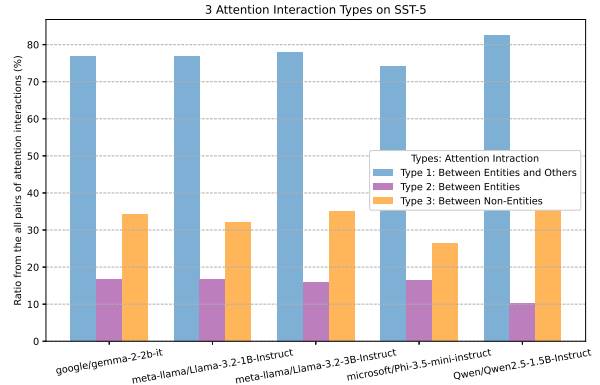


Figure 3: Attention allocation in pre-trained SLMs on the short-text SST-5 dataset. Entity-related interactions (Type 1) dominate.

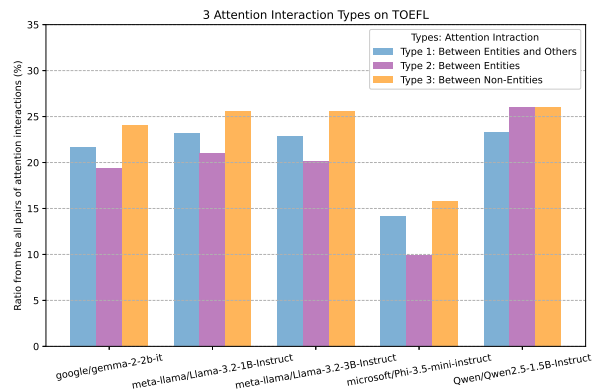


Figure 4: Attention allocation in pre-trained SLMs on the long-text TOEFL dataset. Attention is more distributed compared to SST-5.

us to capture fine-grained variations in local attention and entity focus across discourse units.

5.1 Entity Focus Depends on Text Length and Quality

We first examine how pretrained SLMs allocate attention scores across entity and non-entity tokens in two different textual settings: short-form reviews in SST-5 and long-form essays in TOEFL11. Our goal is to determine whether the model’s internal focus shifts based on text length and discourse complexity.

Dependence on Discourse Granularity: On the short, sentiment-focused sentences of the SST-5 dataset, all models dedicated the vast majority of their attention to interactions involving entities. As shown in Figure 3, the sum of Entity-NonEntity (Ratio_{E-NE}) and Entity-Entity (Ratio_{E-E}) interactions consistently accounts for over 70% of the total attention. This indicates that for concise, opinionated text, entities serve as the primary attentional

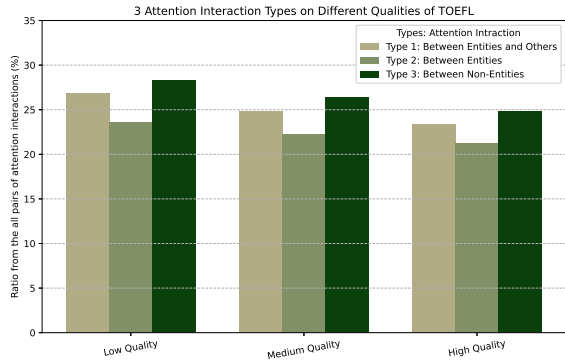


Figure 5: Attention patterns across different qualities of TOEFL essays. As text quality improves, the relative attention on Entity-NonEntity interactions (Type 1) slightly decreases.

anchors for the model. In contrast, on the long-form essays of the TOEFL dataset (Figure 4), attention is more distributed. Entity-related interactions still command a significant share but constitute a smaller portion of the total, ranging from 20% to 30%. This suggests that in complex, descriptive prose, models balance their focus between key entities and broader contextual and structural cues.

Effect of Writing Quality: To analyze whether attention patterns are sensitive to writing quality, we examine the TOEFL11 subset with labeled proficiency levels (“low”, “medium”, “high”). After controlling for essay length and sentence count, we observe a subtle inverse correlation: as writing quality improves, the proportion of Entity-NonEntity attention (Type 1) slightly decreases. Figure 5 illustrates this trend, with low-quality essays exhibiting approximately 26% Type 1 interaction, compared to 23% for high-quality essays.

This observation aligns with previous findings that well-written texts exhibit richer lexical diversity and syntactic variety (Louis and Nenkova, 2013), allowing models to rely on a broader set of discourse cues. Hence, entity tracking remains essential but is less dominant when more reliable and structured context is available.

5.2 Entities Receive Most Attention in Complex Texts

To better understand how SLMs process long-form texts, we conduct a fine-grained analysis of the TOEFL11 dataset by expanding our attention scope beyond noun phrases. Specifically, we compare attention interactions between entities and verb phrases (VPs), as well as between entities and other non-labeled tokens.

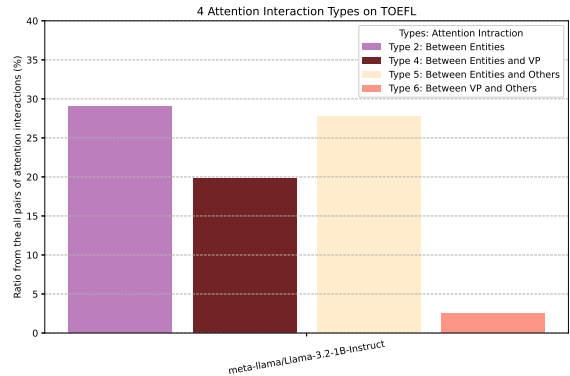


Figure 6: Attention patterns across different qualities of TOEFL essays. As text quality improves, the relative attention on Entity-NonEntity interactions (Type 1) slightly decreases.

Our results reveal a clear attentional hierarchy. As shown in Figure 6, interactions involving entity tokens (e.g., Entity-Entity, Entity-VP, Entity-Other) consistently account for more than 70% of total attention: the sum of Type 2, 4, and 5. By contrast, interactions between verb phrases and non-entity tokens are minimal (approximately 2.5%). This finding confirms that even in linguistically complex environments, SLMs focus on entities as central nodes in the discourse structure.

This behavior is in line with Centering Theory (Grosz et al., 1995), which posits that entities serve as coherence anchors during discourse progression. Our results suggest that pretrained SLMs implicitly adopt a similar processing strategy, prioritizing entities as focal elements in attention allocation.

5.3 PEFT Increases Entity Attention and Improves Accuracy

We next investigate whether QLoRA-based PEFT affects entity-focused attention behavior and model performance. The attention layers of all models are fine-tuned on the SST-5 training set (Appendix B). We then compare their attention distributions and classification accuracy on a balanced evaluation set, which was constructed by sampling 200 instances from each label of the test set to address class imbalance.

Performance Gains: Prior to fine-tuning, the models perform poorly on the 5-class sentiment classification task, with accuracy scores around 40%. After applying LoRA, we observe substantial improvements in classification accuracy and kappa scores (Table 1). In particular, PEFT let SLMs predict extreme emotions well, which was not possible

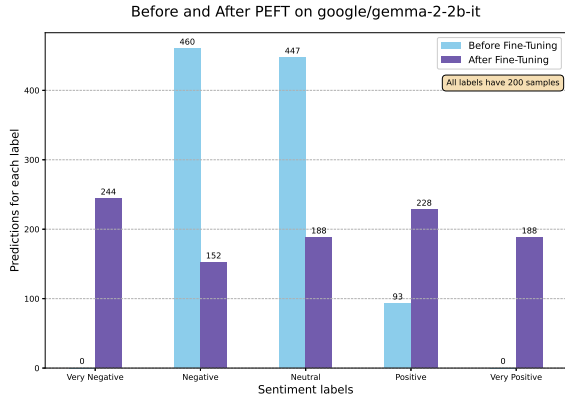


Figure 7: Error analysis: Predictions per label, before and after PEFT on SST5

before fine-tuning (Figure 7). The models become better at distinguishing closely related sentiment categories, such as “positive” vs. “very positive”, confirming that LoRA tuning effectively enhances task-specific capabilities.

Classification Performance before LORA			
	Acc	Kappa-L	Kappa-Q
gemma-2-2b-it	0.38	0.50	0.71
Llama-3.2-1B-it	0.26	0.16	0.27
Llama-3.2-3B-it	0.45	0.55	0.71
Qwen2.5-1.5B-it	0.41	0.54	0.71
Phi-3.5-mini-it	0.42	0.54	0.74

Classification Performance After LORA			
	Acc	Kappa-L	Kappa-Q
gemma-2-2b-it	0.60	0.73	0.88
Llama-3.2-1B-it	0.52	0.66	0.83
Llama-3.2-3B-it	0.52	0.66	0.83
Qwen2.5-1.5B-it	0.52	0.67	0.83
Phi-3.5-mini-it	0.61	0.74	0.88

Table 1: Classification performance on the SST-5 test set before and after LoRA fine-tuning. PEFT leads to substantial improvements in accuracy (Acc) and both Linear (κ_L) and Quadratic (κ_Q) Weighted Kappa scores.

Shifts in Attention Patterns: Crucially, these performance gains are accompanied by consistent and measurable shifts in attention allocation. Table 2 shows that after LoRA fine-tuning, the proportion of Type 2 interactions (Entity-Entity) increases across all models. This suggests that LoRA encourages the model to more explicitly model semantic relationships between entities. Simultaneously, the proportion of non-entity interactions (Type 3)

decreases, reflecting a redistribution of attention toward discourse-salient elements.

This result supports our central hypothesis: LoRA fine-tuning refines the model’s internal attention mechanisms by enhancing focus on linguistically meaningful units—specifically, entities. It also validates the use of our attention analysis framework as a lightweight, model-agnostic diagnostic tool for tracking internal behavioral changes induced by fine-tuning.

Model	Δ E-NE	Δ E-E	Δ NE-NE
gemma-2-2b-it	+0.95	-0.91	+0.63
Llama-3.2-1B-it	+0.52	-0.94	+3.22
Llama-3.2-3B-it	-0.27	-0.62	+0.26
Qwen2.5-1.5B-it	+0.40	-0.49	+0.32
Phi-3.5-mini-it	-0.02	+0.04	-0.16

Table 2: Change in attention allocation ratios (in percentage points, pp) on SST-5 after LoRA fine-tuning. The columns show the change in Entity-NonEntity (Δ E-NE), Entity-Entity (Δ E-E), and NonEntity-NonEntity (Δ NE-NE) attention.

6 Conclusion

Our findings suggest that attention weights – often dismissed as unreliable – can, when anchored in syntactic structure, serve not only as effective diagnostic tools but also as a valuable clue for model development. By tracing attention flow through entity representations, we provide an interpretable and lightweight method that not only probes the internal behavior of SLMs but also points toward directions for improving or tailoring such models to better capture entity-based coherence.

We emphasize that our findings should not be taken as a comprehensive explanation of how Small Language Models operate. The scope of our experiments is necessarily limited, and broader generalizations would require further study. Nonetheless, our work highlights an intriguing avenue: entity-focused attention analysis provides a promising perspective on model interpretability that may inspire future research. Extensions could include multi-sentence coherence modeling, cross-lingual entity behavior, or alignment of model outputs with formal discourse theories.

Limitations

This study, while providing clear findings, has several limitations that offer avenues for future research.

First, our definition of an “entity” as a noun phrase is a pragmatic simplification. This approach does not capture more abstract entities, such as events or concepts, and a more sophisticated entity identification method could yield further insights.

Second, our analysis treats attention as a diagnostic correlate, not a definitive causal mechanism. The final output of a Transformer layer is also influenced by the value vector transformations and the computations within the feed-forward networks. A complete mechanistic explanation would require analyzing the interplay between all these components, which was beyond the scope of this work.

Third, the scope of our study is confined to a specific set of SLMs and two classification tasks. While the consistency of our findings across multiple models is encouraging, they may not generalize to all model architectures (e.g., non-Transformers), significantly larger models (LLMs), or different task modalities, such as text generation.

Finally, our method of averaging attention scores across all heads in the final layer provides a high-level, aggregate view of the model’s focus. This approach necessarily obscures the diverse and specialized functions that individual attention heads are known to perform (Clark et al., 2019). A more granular, head-level analysis could reveal which specific heads are most affected by fine-tuning and what linguistic roles they play, representing a promising direction for future work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author had been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [BERT: A study of attention in BERT](#). 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 5717–5726, Hong Kong, China. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Erwan Fagnou, Paul Caillon, Blaise Delattre, and Alexandre Allauzen. 2024. [Chain and causal attention for efficient entity tracking](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13174–13188, Miami, Florida, USA. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. [Centering: A framework for modeling the local coherence of discourse](#). *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. [Attention, intentions, and the structure of discourse](#). *Computational Linguistics*, 12(3):175–204.

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*.

Sarthak Jain and Byron C. Wallace. 2019. [Attention is not explanation](#). 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Zichao Li, Yanshuai Cao, and Jackie Chi Kit Cheung. 2024. [Do LLMs build world representations? probing through the lens of state abstraction](#). In *The Twelfth International Conference on Learning Representations*.

Sharid Loáiciga, Anne Beyer, and David Schlangen. 2022. [New or old? exploring how pre-trained language models represent discourse entities](#). In

Proceedings of the 29th International Conference on Computational Linguistics, pages 875–886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. [What makes writing great? first experiments on article quality prediction in the science journalism domain](#). *Transactions of the Association for Computational Linguistics*, 1:341–352.

Bhavya Prakash, Tamar Rott Shaham, Tal Linzen, and Yonatan Belinkov. 2024. [Fine-tuning enhances existing mechanisms: A case study on entity tracking](#). In *The Twelfth International Conference on Learning Representations*.

Sebastian Schuster and Tal Linzen. 2022. [When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. [Is attention interpretable?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kevin Wang, Vikrant Varma, Neel Nanda, Jacob Steinhardt, and David McAllester. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations*.

A Appendix: Dataset Details

The TOEFL11 dataset consists of 12,378 essays written in response to eight distinct open-ended prompts, which are detailed in Table 4. On average, each essay is approximately 411 words long, with further statistics provided in Table 3. The dataset is labeled by proficiency, with a distribution of 1,308 low-quality, 6,568 medium-quality, and 4,502 high-quality essays.

For our sentence-level analysis, we began with a total of 128,549 sentences. We applied several filtering criteria to ensure data quality, excluding: 2,046 sentences that lacked any identifiable entities; 1,970 sentences that our entity-subword mapping parser could not process correctly; and 3,545

Dataset	#Texts	Avg len (Std)	Max len	Scores
T-P1	1,656	401 (97)	902	1-3
T-P2	1,562	423 (97)	902	1-3
T-P3	1,396	407 (102)	837	1-3
T-P4	1,509	405 (99)	852	1-3
T-P5	1,648	424 (101)	993	1-3
T-P6	960	425 (101)	925	1-3
T-P7	1,686	396 (87)	755	1-3
T-P8	1,683	407 (92)	795	1-3

Table 3: Dataset statistics on tokenization: each TOEFL prompt (T-P).

sentences shorter than five words. This filtering process resulted in a final set of 120,999 sentences used in our analysis.

The Stanford Sentiment Treebank (SST-5) dataset contains 5,992 movie reviews for 5-class sentiment classification (from “very negative” to “very positive”). The average sentence length is 23.44 subwords. From this initial set, we excluded 24 sentences shorter than five words and one sentence that failed parsing, resulting in a final analysis set of 5,967 sentences (99.6% of the original dataset).

B Appendix: LORA Hyperparameter Details

The LoRA fine-tuning was conducted using the HuggingFace PEFT library. We employed 4-bit quantization (QLoRA) with the nf4 data type and loaded the base models with fp16 precision. The target modules for LoRA were the attention layers of the SLMs: “q_proj”, “k_proj”, “v_proj”, “o_proj”. This results in 0.12% trainable parameters for google/gemma-2-2b-it, and 0.14% for meta-llama/llama-3.2-1B. The primary hyperparameters were set as follows: rank=16, alpha=32, lora_dropout=0.05, and a learning rate of $1e-4$ with AdamW optimizer. Models were trained for 2 epochs with a batch size of 4.

C Appendix: Subwords Tokenization as SLM

Our entity-subword mapping parser was designed to handle model-specific tokenization schemes. We observed that the SLMs in our study primarily use one of two conventions to mark word boundaries: a prefix _ (e.g., _word) or a special character “Ĉ” (e.g., “Ĉ”word). Our parser correctly interprets these conventions for each model to ensure accurate alignment between linguistic noun phrases and their corresponding subword tokens, as illustrated

T-Prompt 1	Agree or Disagree: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
T-Prompt 2	Agree or Disagree: Young people enjoy life more than older people do.
T-Prompt 3	Agree or Disagree: Young people nowadays do not give enough time to helping their communities.
T-Prompt 4	Agree or Disagree: Most advertisements make products seem much better than they really are.
T-Prompt 5	Agree or Disagree: In twenty years, there will be fewer cars in use than there are today.
T-Prompt 6	Agree or Disagree: The best way to travel is in a group led by a tour guide.
T-Prompt 7	Agree or Disagree: It is more important for students to understand ideas and concepts than it is for them to learn facts.
T-Prompt 8	Agree or Disagree: Successful people try new things and take risks rather than only doing what they already know how to do well.

Table 4: Topic description: TOEFL (T).

in Table 5.

Origin Text	Is it better for oneself’s developpement to have broad knowledge of many academic subjects than to specialize in one specific subject?
google/gemma-2-2b-it	
Tokenized-Ids	tensor([[2, 2437, 665, 2525, 604, 63320, 235303, 235256, 2115, 227070, 577, 791, 7209, 5567, 576, 1767, 15459, 12749, 1178, 577, 78292, 575, 974, 3724, 5091, 235336]])
Tokenized-Subwords	[‘<bos>’, ‘Is’, ‘_it’, ‘_better’, ‘_for’, ‘_oneself’, ‘’, ‘s’, ‘_develop’, ‘pement’, ‘_to’, ‘_have’, ‘_broad’, ‘_knowledge’, ‘_of’, ‘_many’, ‘_academic’, ‘_subjects’, ‘_than’, ‘_to’, ‘_specialize’, ‘_in’, ‘_one’, ‘_specific’, ‘_subject’, ‘?’]
meta-llama/Llama-3.2-1B	
Tokenized-Ids	tensor([[128000, 3957, 433, 2731, 369, 57669, 596, 2274, 79, 1133, 311, 617, 7353, 6677, 315, 1690, 14584, 15223, 1109, 311, 48444, 304, 832, 3230, 3917, 30]])
Tokenized-Subwords	[‘< begin_of_text >’, ‘Is’, ‘Ġit’, ‘Ġbetter’, ‘Ġfor’, ‘Ġoneself’, ‘’, ‘s’, ‘Ġdevelop’, ‘p’, ‘ement’, ‘Ġto’, ‘Ġhave’, ‘Ġbroad’, ‘Ġknowledge’, ‘Ġof’, ‘Ġmany’, ‘Ġacademic’, ‘Ġsubjects’, ‘Ġthan’, ‘Ġto’, ‘Ġspecialize’, ‘Ġin’, ‘Ġone’, ‘Ġspecific’, ‘Ġsubject’, ‘?’]

Table 5: Examples of different subword tokenization schemes deployed on SLMs.

Stance Detection on Nigerian 2023 Election Tweets Using BERT: A Low-Resource Transformer-Based Approach

Mahmoud Said Ahmad¹ and Habeebah A. Kakudi^{2*}

¹Federal University of Technology Babura, Jigawa, Nigeria

²Bayero University Kano, Kano, Nigeria

msahmad.cs@futb.edu.ng, hakakudi.cs@buk.edu.ng

*Corresponding author

Abstract

This study investigates stance detection on Nigerian 2023 election tweets by comparing transformer-based and classical machine learning models. A balanced dataset of 2,100 annotated tweets was constructed, and BERT-base-uncased was fine-tuned to perform stance classification into three categories: Favor, Neutral, and Against. The model achieved strong results, with 98.1% accuracy on a stratified 80/20 split and an F1-score of 96.9% under 5-fold cross-validation. To contextualize these outcomes, baseline models including Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machines (SVM) were also evaluated. While several baselines demonstrated competitive performance (with SVM reaching an F1-score of 97.6%), BERT proved more robust in handling noisy, sarcastic, and ambiguous text, making it better suited for real-world applications. The findings highlight both the competitiveness of classical methods on curated datasets and the scalability of transformer-based models in low-resource African NLP contexts.

1 Introduction

Democratic governance depends on citizen participation and empowerment. These elements play a vital role in addressing long-standing social, economic, and political imbalances (Bandyopadhyay and Green, 2012).

The rapid growth of social media has transformed how individuals express and disseminate political opinions. Platforms such as Twitter and Facebook provide quick and affordable means of gathering real-time perspectives from diverse groups (Ceron et al., 2014; Díaz et al., 2016). These platforms complement traditional data collection methods and are now widely applied in political prediction and analysis (Liu et al., 2021).

A central application of this trend is stance detection, which involves determining whether a user

supports, opposes, or remains neutral toward a specific topic (Küçük and Can, 2020). Unlike sentiment analysis, which measures emotional tone, stance detection explicitly links opinions to their targets. This distinction makes it especially valuable for monitoring misinformation, examining polarization, and analyzing the dynamics of political discourse (Hardalov et al., 2022; Zhao and Yang, 2020; Liu et al., 2024).

In highly polarized contexts such as Nigeria’s 2023 presidential election, stance detection offers critical insights into public opinion toward candidates and the broader nature of online debates.

Despite significant progress in natural language processing (NLP), African electoral contexts remain underrepresented in stance detection research. Previous studies highlight the need for localized datasets and tailored approaches to capture electoral behaviors in decentralized political systems (Khan et al., 2024).

However, notable challenges persist. The absence of annotated datasets, the prevalence of code-switching and informal discourse on social media, and limited computational resources restrict progress in this area. Moreover, large language models such as Mistral require substantial GPU infrastructure and are trained on English data that arguably under-represent African dialects.

While, at the moment, we can only speculate as to the reason, this paper provides evidence that at least one LLM performs very poorly in zero-shot as well as few-shot evaluations, making them unsuitable for low-resource environments. Nevertheless, LLMs do have a constructive role to play through supervised fine-tuning.

This study presents a CPU-efficient stance detection model for the 2023 Nigerian presidential election. A balanced dataset of 2,100 tweets was constructed, and the resource-efficient BERT-base-uncased model was fine-tuned to classify stances into *Favor*, *Neutral*, and *Against*. The

specific contributions of this work are as follows:

- Construction of a balanced dataset of 2,100 annotated tweets.
- Demonstration of effective stance detection using BERT on CPU-only hardware.
- Empirical evidence showing 98.1% accuracy with F1-scores above 0.98 across stance categories.

These contributions demonstrate that, with careful dataset curation and model selection, transformer-based models as simple as BERT can achieve high performance in resource-limited African NLP contexts. This research expands the reach of computational political analysis in under-represented regions. In the following section, related work on stance detection and transformer-based approaches is reviewed.

1.1 Problem Statement

Despite notable progress in natural language processing (NLP), stance detection remains an under-explored area in the context of African elections, particularly in Nigeria. The 2023 Nigerian presidential election generated extensive online discourse on platforms such as Twitter, often characterized by colloquial language, slang, and frequent code-switching. However, no locally annotated datasets or computationally optimized models currently exist to address this setting. Moreover, state-of-the-art large language models, such as Mistral 7B, require substantial GPU resources and, as we will show in Section 4.4, perform poorly in zero-shot and few-shot settings, rendering them impractical for low-resource environments.

This gap highlights the urgent need for an efficient and reliable stance detection system that can be trained using widely available CPU hardware while still achieving high accuracy in classifying political stances as *Favor*, *Neutral*, or *Against*.

2 Related Work

The stance detection task has gained growing interest in natural language processing (NLP), with the heightened role of social media in political discussion. Stance detection, unlike sentiment analysis that involves the assessment of emotional tone, involves determining if a speaker or author is supportive of, against, or neutral about a particular topic (Mohammad et al., 2016). This makes it highly

applicable to electoral research and political alignment studies (Al-Dayel and Magdy, 2021).

2.1 Traditional Methods

Early stance detection used classifiers like Support Vector Machines, logistic regression, and Naive Bayes (Mohammad et al., 2016). These relied on hand-crafted features such as n-grams and sentiment lexicons. While they worked well in some cases, they often struggled with sarcasm, slang, and the informal language commonly found on social media.

2.2 Transformer-based Architectures

The introduction of transformers, especially BERT (Devlin et al., 2018), improved stance detection by capturing the full context of sentences through self-attention. BERT has outperformed CNN, LSTM, and ensemble systems in benchmarks like SemEval-2016 and COVID-19 stance detection (Sirrianni and Zhang, 2021; Davydova and Dutta, 2024). It shows a strong ability to recognize implicit and subtle opinions.

2.3 New Large Language Models

Recent models like ChatGPT, LLaMA, and Mistral advance NLP, with frameworks such as COLA Lan et al. (2024) supporting multi-agent stance recognition. However, these systems need a lot of computational power, which limits their use in low-resource environments.

2.4 Zero-shot and Transfer Learning Approaches

Zero-shot and few-shot methods purport to reduce the need for large labeled datasets. Examples include Multi-Perspective Transferable Feature Fusion (Zhao et al., 2024, MTPFF) and Cross-Target with Text and Network embeddings (Khiani et al., 2024, CT-TN) which use both textual and network signals for stance detection across targets. While these methods are generally considered to be effective, they require complex prompts and careful tuning, making them more challenging to use in limited settings.

2.5 Model Selection: BERT

We chose BERT-base-uncased as our main model. We made this choice not because we believe it is the best overall option, but due to its practical benefits:

- It has shown strong previous results in stance detection studies (Sirrianni and Zhang, 2021);

- It works well in CPU-based environments.
- It performs reliably on medium-scale, balanced datasets.
- Hugging Face’s Trainer API provided a simple interface to batch train, validate, and log.

2.6 African NLP and Low-Resource Contexts

Beyond general stance detection, recent African NLP efforts such as Masakhane (Orife et al., 2020), MasakhaNER 2.0 (Adelani et al., 2022), and AfriSenti (Abdulmumin et al., 2023) have emphasized the importance of building datasets and benchmarks tailored to African languages. These initiatives highlight the challenges of low-resource settings, code-switching, and domain-specific biases, issues that are also evident in our Nigerian election dataset. Our work extends this line of research by focusing on stance detection in a politically charged African context.

This choice supports the need for resource-efficient NLP in African contexts. It shows how careful dataset preparation and thoughtful model selection can enable effective stance detection without the need for expensive infrastructure.

3 Dataset and Preprocessing

The study aims at stance analysis in Twitter posts about Nigeria’s 2023 presidential election, particularly tweets mentioning four principal candidates: Atiku Abubakar, Bola Ahmed Tinubu, Peter Obi, and Rabiu Kwankwaso. The methodological pipeline involved data collection, noise removal, dataset enlargement, model selection, and evaluation processes.

The resultant corpus contained 2,100 prepared tweets, balanced across three stance labels *favor*, *neutral*, and *against*. Tweets were scraped through focused hashtag searches and filtered using hand-engineered rules to remove off-topic or ambiguous posts.

3.1 Dataset Collection and Balancing Strategy

We collected tweets with candidate-specific hashtags such as #atiku4president, #tinubu2023, and #obidatti2023. The initial distribution revealed severe class imbalances, particularly in the under-representation of some stance categories for other candidates. Table 1 shows the skewed nature of the raw dataset.

Candidate	Total Tweets	Favor	Neutral	Against
Atiku	47,508	175	175	80
Tinubu	23,456	175	175	80
Peter Obi	59,212	199	—	—
Kwankwaso	8,702	171	—	—

Table 1: Initial distribution of scraped tweets showing class imbalance

Candidate	Favor	Neutral	Against
Atiku	175	175	175
Tinubu	175	175	175
Peter Obi	175	175	175
Kwankwaso	175	175	175

Table 2: Final balanced dataset following augmentation (Total: 2,100 tweets)

To address these imbalances and ensure that the dataset could be reliably used for training a supervised classifier, a set of balancing techniques was applied. These included heuristic labeling, rule-based annotation, and multiple data augmentation methods.

The final training dataset was uniformly structured, with each candidate having an equal number of tweets in each stance category, 175 per class. This resulted in a balanced dataset of 2,100 tweets in total. The complete breakdown is presented in Table 2.

For a clearer overview of this transformation, a pie chart (Figure 1) was included to illustrate the final stance distribution. Each class—Favor, Neutral, and Against—is represented equally, with 700 tweets each.

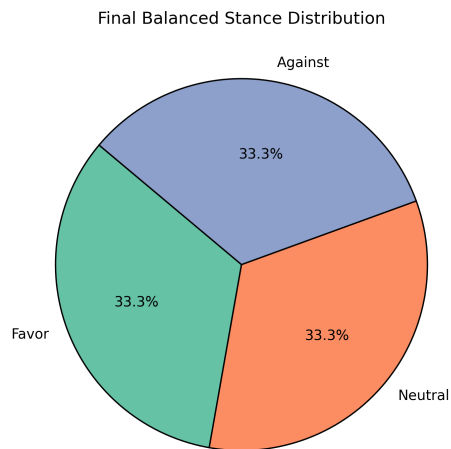


Figure 1: Final distribution of stance categories after dataset balancing

To achieve the target level of 700 tweets per candidate (175 per stance category), a sequence of enrichment and refinement processes was employed to expand the dataset:

- **Rule-Based Labeling:** Sentiment words, hashtags, and user mentions were used as heuristics to assign initial stance labels.
- **Keyword-Based Weak Supervision:** Tweets with overt expressions of support or disapproval were labeled "favor" or "against," while posts lacking explicit evaluative content were placed in the "neutral" category.
- **Data Augmentation:** A set of augmentation techniques was applied to increase the linguistic richness and balance of the dataset.

a. Expansion of the Dataset by Augmentation

To counteract low samples in some classes, most prominently "favor" and "against," the following augmentation methods were employed:

- **Synonym Substitution:** Synonyms were incorporated in tweets using WordNet and NLTK libraries to create natural variants.
- **Back-Translation:** Tweets were automatically translated into another language and back into English to generate paraphrased versions.
- **Template-Based Generation:** Stance-bias sentence templates were completed with candidate names and contextual phrases to increase diversity.

This multi-step approach ensured that the final dataset was not only balanced but also linguistically rich and representative of actual social media language.

3.2 Data Cleaning and Preprocessing

For the sake of data quality and interpretability of models, each tweet was preprocessed with uniform text preprocessing that consisted of:

- Normalization of all characters to lowercase
- Erasure of URLs, mentions, hashtags, punctuation, and redundant spaces
- Lexical analysis to identify richness and detect anomalies

It aided in holding input consistent and removing noise, which is particularly required in social media settings.

3.3 Tokenization and Data Formatting

Tweets were tokenized with the bert-base-uncased tokenizer, padding and truncating to a fixed maximum token length of 128. The stance labels were numerically encoded using LabelEncoder. The dataset was loaded into Hugging Face's Dataset format. A balanced split of the dataset into training and test sets in the ratio 80-20 was used to preserve even class distribution in both sets.

3.4 Model Configuration and Training

The stance classifier was built by fine-tuning the BERTForSequenceClassification model. Training was done using Hugging Face's Trainer class, with parameters to configuration tweaked for CPU-based systems:

```
TrainingArguments(  
    output_dir="./bert_stance_output",  
    num_train_epochs=2,  
    per_device_train_batch_size=16,  
    per_device_eval_batch_size=16,  
    logging_dir="./bert_logs",  
    logging_steps=10,  
    save_steps=100,  
    logging_strategy="steps",  
    load_best_model_at_end=False  
)
```

This setup allowed the model to be effectively trained without requiring access to GPUs.

3.5 Evaluation Framework

The model performance was compared to commonly used classification metrics:

- **Accuracy** – proportion of correct predictions
- **Precision** – precision among positive predictions
- **Recall** – proportion of actual positives correctly identified
- **F1-Score** – harmonic mean of precision and recall

Evaluation Metric	Result
Accuracy	98.10%
Precision	98.10%
Recall	98.10%
F1-Score	98.09%
Evaluation Loss	0.1433

3.6 Error Analysis of Predictions

The below detailed confusion matrix shows how accurately each stance class was predicted.

- **Against:** 139 correctly predicted, 1 mislabeled as Neutral.
- **Neutral:** 139 correct, 1 mislabeled as Favor.
- **Favor:** 134 correctly predicted; 4 were predicted as Against, 2 as Neutral.

While overall performance was good, the majority of misclassifications were between proximate categories (e.g., Favor and Neutral). That likely stems from the vagueness and informality of social media use. Nevertheless, the strength of the model in discriminating among fine-grained categories is very high.

Identified Challenge	Applied Resolution
Failure of Mistral 7B to make stance predictions	Replaced by BERT for local fine-tuning on labeled data
Imbalance in Favor and Against examples	Treated using multiple augmentation techniques (e.g., templates, synonyms)
GPU limitations on Google Colab	Fine-tuned on CPU with optimized parameters for learning in small batches
Noisy or inconsistent labeling in the primary dataset	Cleaned using rule-based heuristics and manual quality checks
Risk of overfitting due to reliance on a single split	Addressed by performing 5-fold cross-validation to confirm robustness

Table 3: Overview of experimental difficulties, corrective strategies, and validation measures

3.7 Model Overview

This study employed the bert-base-uncased model configuration within the Hugging Face Transformers library (Wolf et al., 2019). BERT’s

architecture includes 12 transformer encoder layers with multi-head self-attention to encode rich contextualized information from text input.

The modeling pipeline had the following necessary steps:

- **Tokenization:** Raw text of tweets was tokenized into subword units using a BERT-compatible tokenizer.
- **Embedding:** Tokens were converted into numerical vectors that represent lexical and positional context.
- **BERT Encoder:** A series of transformer layers was applied to the embeddings to learn contextualized relationships within each tweet.
- **Dropout:** A dropout layer with a rate of 0.1 was added to lower the danger of overfitting.
- **Classification Layer:** A Softmax over a linear output layer mapped BERT outputs to probabilities across the three classes.

Model training was performed using the Hugging Face-offered Trainer utility. The most significant training parameters were:

- Epochs: 2
- Batch size: 16
- Learning rate: 5e-5
- Optimizer: AdamW

Training was done using the cross-entropy loss function, which is widely used for multi-class classification problems. Despite utilizing only CPU resources to the fullest, high performance was achieved due to proper implementation, efficiency, and dataset readiness.

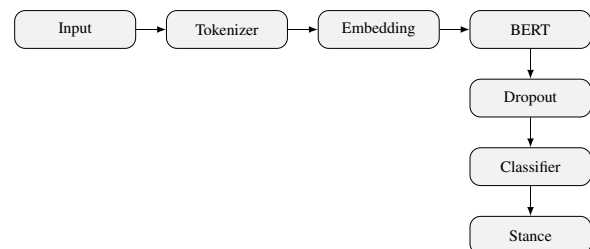


Figure 2: Compact Diagram of the Fine-Tuned BERT Pipeline

3.8 Training and Evaluation with Cross-Validation

For testing generalization, the data was split into training and test sets through a balanced 80/20 split. This meant that the proportionate distribution of the three stance labels—Favor, Neutral, and Against—was preserved in both partitions.

Training employed the Hugging Face Trainer framework, with the ability for model fine-tuning, evaluation, and logging. Training was executed only twice across two epochs, at a batch size of 16 with the AdamW optimizer at a learning rate of $5e-5$. The loss function employed was the categorical cross-entropy one, which suited addressing the three discrete stances.

Intermediate evaluation was carried out after each epoch. Logging and checkpointing routines were activated to help ensure training reproducibility and allow progress to be picked up in the event of need.

In addition to the 80/20 balanced split, we also used a 5-fold balanced cross-validation to further test the model’s strength. In this setup, we divided the dataset into five folds, each with equal class representation. Each fold acted as a test set once, while the other four were used for training. We averaged the model’s performance across the folds and reported the mean accuracy, precision, recall, and F1-scores along with standard deviations. This dual evaluation method helped us present both detailed single-split outcomes and broader cross-validation results.

3.11 Methodology Summary

For clarity, we summarize the methodological pipeline as follows:

Dataset

The dataset consisted of 2,100 tweets related to Nigeria’s 2023 presidential election. Tweets were heuristically labeled into three stance categories: Favor, Neutral, and Against. Data augmentation techniques such as synonym replacement, back-translation, and sentence templating were used to improve balance and diversity.

Dataset Split

We used two evaluation strategies:

1. A single 80/20 balanced split, chosen for reproducibility and comparability with prior studies.

2. Balanced 5-fold cross-validation, where the dataset was divided into five folds with equal class representation. Each fold was used once as the test set while the remaining four served as training data.

This dual setup allowed us to report both detailed single-split results and robust average performance across folds.

Model and Training Setup

We fine-tuned BERT-base-uncased using Hugging Face’s Trainer API. Training was run on CPU-only hardware to reflect resource-limited conditions. The key parameters were: learning rate 2×10^{-5} , batch size 16, and 2 epochs.

Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1 score (weighted across classes). Confusion matrices were generated for error analysis. For cross-validation, mean and standard deviation were reported across the five folds.

4 Results

4.1 Baseline Models

To provide context for BERT’s performance, we evaluated several classical machine learning baselines using TF-IDF features: Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). Table 4 summarizes their 5-fold cross-validation performance.

Model	Accuracy	F1-score
Naïve Bayes (5-fold CV)	94.7% (± 0.7)	0.947
Logistic Regression (5-fold CV)	96.6% (± 0.7)	0.966
Random Forest (5-fold CV)	97.0% (± 1.0)	0.970
SVM (5-fold CV)	97.6% (± 0.6)	0.976
BERT (5-fold CV)	96.9% (± 0.8)	0.969

Table 4: Comparison of classical ML baselines and BERT on stance detection using 5-fold cross-validation.

The classical baselines performed strongly, with Random Forest and SVM achieving F1-scores above 97%. BERT’s performance (96.9% F1) was comparable, but its main advantage lies in robustness to noise, sarcasm, and domain shift, making it more reliable for real-world deployment beyond the controlled dataset. These results highlight that while classical models remain competitive on balanced datasets, pretrained transformers provide scalability and adaptability.

4.2 Performance on Single Split

On the balanced 80/20 split, our BERT-base-uncased model achieved an accuracy of 98.1% with weighted F1-scores above 0.98 across all stance categories. The confusion matrix (Figure 3) showed that most misclassifications occurred in tweets with ambiguous or sarcastic language. Table 5 reports the detailed classification metrics.

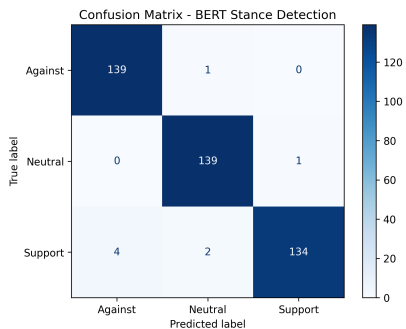


Figure 3: Confusion matrix on the 80/20 stratified split.

Class	Precision	Recall	F1-score
Against	0.97	0.99	0.98
Neutral	0.99	0.98	0.98
Support	0.99	0.97	0.98
Weighted Avg.	0.98	0.98	0.98

Table 5: Classification metrics on the 80/20 stratified split.

4.3 Cross-Validation Results

To further validate robustness, we performed 5-fold stratified cross-validation on the balanced dataset of 2,100 tweets. The model achieved a mean accuracy of 96.9% (± 0.8), precision of 96.9% (± 0.8), recall of 96.9% (± 0.8), and F1-score of 96.9% (± 0.8), as shown in Table 6.

Metric	Mean	Std Dev
Accuracy	96.9%	± 0.8
Precision	96.9%	± 0.8
Recall	96.9%	± 0.8
F1 Score	96.9%	± 0.8

Table 6: 5-fold cross-validation performance of BERT stance classifier.

The slight difference between the single-split result (98.1%) and the cross-validation mean (96.9%) highlights the sensitivity of performance to dataset partitioning. The single split demonstrates the model’s potential under a particular train–test scenario, while the cross-validation average provides a

more reliable estimate of real-world generalization across multiple data splits.

Taken together, the single-split and cross-validation experiments confirm that supervised fine-tuning of BERT provides consistent and robust performance across different partitions of the dataset. However, recent advances in large language models (LLMs) have made it possible to attempt stance detection without fine-tuning, through prompting alone. To investigate this alternative approach, we conducted few-shot prompting experiments, as described in the next subsection.

4.4 Few-Shot Prompting Experiments

To explore whether large language models can perform stance detection without supervised fine-tuning, we conducted few-shot prompting experiments using the Flan-T5-base model. The model was evaluated in 0-shot, 5-shot, 10-shot, 20-shot, and 60-shot settings. In each case, a small set of labeled examples was provided in the prompt as demonstrations before classifying unseen tweets. Table 7 summarizes the results.

Setup	Accuracy	Macro F1
0-shot	54%	0.42
5-shot	52%	0.41
10-shot	52%	0.41
20-shot	38%	0.18
60-shot	38%	0.18

Table 7: Few-shot prompting performance of Flan-T5 on stance detection.

The results indicate that few-shot prompting did not perform in the same league as supervised methods. The best performance was achieved in the 0-shot setting, with an accuracy of 54% and macro F1 of 0.42. Adding more examples (5-shot and 10-shot) yielded no improvement, while larger prompts (20-shot and 60-shot) performed significantly worse, likely due to input truncation from the model’s limited context window.

An initial effort at zero-shot stance classification using Mistral 7B Instruct, another large language model, was confronted with its own drawbacks:

- **Poor prediction scores:** All the evaluation metrics (precision, recall, and F1-score) had a value of zero for stance classes.
- **Total misclassification:** The model made no correct predictions on over 279,000 tweets.

- **Bias against "favor" class:** The model made no outputs tagged as "favor," likely due to biased prompt encoding or internal representation issues.
- **Overcomputing demands:** GPU memory limitations in freely available platforms like Google Colab rendered training impossible.
- **Stable operation:** Inference and loading cycles that were slow resulted in frequent failures and crashing of the sessions.

Furthermore, across all prompting conditions, the *Support* class was never predicted, highlighting class imbalance issues. These findings suggest that while instruction-tuned LLMs can perform stance detection without fine-tuning, their performance is inconsistent and substantially weaker than supervised approaches like BERT. This demonstrates the limitations of relying solely on prompting-based methods for nuanced political stance classification.

4.5 Error Analysis

Despite overall strong results, errors were observed in tweets that used indirect references, irony, or heavy code-switching between English and local languages. Such cases remain challenging for transformer models and indicate areas for future dataset expansion and multilingual model fine-tuning.

To illustrate these challenges more concretely, Table 8 presents several example tweets where the model made errors.

As shown, errors often arose from sarcasm, comparative statements, or mixed sentiments, which remain difficult even for transformer-based models. These examples highlight the importance of expanding datasets with more nuanced cases and considering multilingual or context-aware approaches in future work. We next interpret these results in detail in the following discussion section.

5 Discussion

The results show that fine-tuning a transformer model like BERT on a balanced and well-curated dataset can achieve strong classification performance in politically sensitive contexts. The model reached 98.1% accuracy on an 80/20 split and maintained stable results under 5-fold cross-validation (mean accuracy and F1-score of 96.9%). The small gap between the two estimates suggests consistent performance across dataset splits, with cross-

Tweet (anonymized)	True Label	Predicted	Comment
“So after all this, Obi still thinks he can win? Nigerians know better.”	Against	Neutral	Sarcasm confused the model.
“Tinubu has his flaws but at least he has experience.”	Favor	Neutral	Subtle support phrased cautiously.
“#Atiku2023 we deserve better leaders!”	Against	Favor	Hashtag misled model despite negative wording.
“Kwankwaso is not bad, but Obi remains my choice.”	Neutral	Favor	Mixed stance with comparative phrasing.
“I don’t care who wins, same story every time.”	Neutral	Against	Cynicism mistaken for opposition.

Table 8: Examples of challenging tweets where the model made errors. Tweets have been anonymized and paraphrased for clarity.

validation offering a more reliable measure of true generalization.

To contextualize these findings, we compared BERT with classical baselines. Interestingly, SVM performed competitively (F1 score of 97.6% under cross-validation), nearly matching BERT. This indicates that the dataset is relatively learnable for simpler models due to its balanced distribution and clear stance signals. However, BERT remains more scalable and robust, particularly in handling nuanced expressions, sarcasm, and noisy text common in political discourse.

Mistral 7B, while theoretically stronger, underperformed in practice. It struggled with zero-shot predictions and faced hardware limitations, including memory overflows in free-tier environments. By contrast, BERT-base-uncased proved efficient, resource-friendly, and easy to implement with widely available tools like Hugging Face’s Trainer.

Data preparation was a major challenge, as the original stance labels were skewed toward negative tweets, with fewer neutral or supportive examples. To mitigate this, we applied data augmentation techniques such as synonym substitution, template rewriting, and back-translation. These methods helped balance the dataset and improved

generalization.

Error analysis revealed that most misclassifications occurred between Neutral and Favor classes, reflecting the implicit nature of stance in political text. These errors were relatively minor and had little impact on overall accuracy. Preprocessing also played a key role: text normalization, removal of irrelevant tokens (e.g., links, mentions), and basic linguistic filtering improved input quality and ensured that the model learned from the most relevant features.

Despite strong results, limitations remain. The dataset includes only English tweets, while much Nigerian political discourse involves multiple languages and frequent code-switching. Moreover, real-world distributions are more skewed and unstable than the curated dataset used here, which may limit generalizability.

Overall, this study demonstrates that high-performance stance detection is achievable without large-scale hardware, provided the dataset is carefully prepared and models are fine-tuned. The comparison of classical baselines with transformer models highlights the complementary value of both approaches. Future work will extend this effort to code-switched and multilingual stance detection in Nigerian political discourse, building on African NLP initiatives such as Masakhane, MasakhaNER, and AfriSenti.

6 Conclusion

This study examined stance detection on Nigerian election tweets using BERT and classical machine learning baselines. The results show that fine-tuning BERT on a balanced and augmented dataset yields high accuracy, achieving 98.1% on a stratified 80/20 split and 96.9% F1 on 5-fold cross-validation. Classical baselines, including Logistic Regression, Random Forest, and SVM, also performed strongly, with SVM achieving 97.6% F1. These findings suggest that while the dataset is learnable with simpler models, transformers provide robustness to noisy and nuanced political language, offering better generalization potential.

Error analysis revealed that most misclassifications occurred between *Neutral* and *Support*, often due to sarcasm, subtlety, or code-switching. Although BERT proved efficient and effective, limitations remain: the dataset only covered English tweets, and political discourse in Nigeria frequently involves multiple languages and code-switching.

Future work will explore multilingual stance detection and context-aware transformers, building on recent African NLP initiatives such as Masakhane, MasakhaNER, and AfriSenti.

Overall, this research confirms that high-performance stance detection is possible without large-scale hardware, provided that data preparation is rigorous. Combining classical baselines with transformer models provides a comprehensive evaluation and demonstrates the potential of modern NLP approaches for political text classification in low-resource African settings.

6.1 Limitations and Future Work

Although this study demonstrates the feasibility of stance detection in a low-resource African electoral context, several limitations remain. First, the dataset consists of 2,100 tweets, which, while balanced, is relatively small. The reliance on heuristic labeling and data augmentation may also introduce noise, and further validation with human-annotated datasets would strengthen reliability.

Our experiments were restricted to English-language tweets and a CPU-only training setup. This excludes the widespread use of code-switching and indigenous languages in Nigerian political discourse, which may reduce real-world applicability.

While BERT-base-uncased performed consistently under cross-validation, the study did not compare fine-tuned large language models (LLMs) due to hardware constraints. Future research should explore multilingual transformer models, lightweight LLM adaptations (e.g., quantization, distillation), and larger annotated datasets to better capture the complexity of political conversations in Nigeria and other underrepresented regions.

Acknowledgments

The first author would like to express his deep gratitude to Prof. Gerald Penn for his invaluable feedback and for kindly offering mentorship that significantly improved the quality of this work.

References

1. Abdulmumin, D. I. Adelani, A. Awokoya, R. Gitau, , and 1 others. 2023. Afrisenti: A sentiment analysis benchmark for african languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

- D. I. Adelani, B. F. Dossou, J. Kreutzer, J. O. Alabi, S. H. Muhammad, and 1 others. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1462–1481.
- Abdulrahman Al-Dayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.
- Sanghamitra Bandyopadhyay and Elliott Green. 2012. The relevance of political decentralization in developing countries. *Development Policy Review*, 30(2):131–153.
- A. Ceron, L. Curini, and S. M. Iacus. 2014. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358.
- Kseniia Davydova and Pallavi Dutta. 2024. Bert-based stance detection on covid-19 twitter discussions. *Social Network Analysis and Mining*, 14(1):1–12.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, abs/1810.04805:1–15.
- Fernando Díaz, Michael Gamon, Jake M. Hofman, Emre Kıcıman, and David Rothschild. 2016. Online and social media data as an imperfect continuous panel survey. *PLOS ONE*, 11(1):e0145406.
- Momchil Hardalov, Preslav Nakov, and Ivan Koychev. 2022. Survey on stance detection. *ACM Computing Surveys*, 55(1):1–37.
- Nizamuddin Khan, Firoj Biswas, and Mostafijur Rahman. 2024. Dynamics of electoral behavior of panchayat election in nadia district, west bengal. *The Deccan geographer*, 61:266–282.
- Shayan Khiabani, Siyuan Chen, and William Yang Wang. 2024. Cross-target stance detection via text and network embeddings. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2597–2608.
- Dror Küçük and Funda Can. 2020. Stance detection: A survey. In *ACM Computing Surveys*, volume 53, page 1–37.
- Yuan Lan, Chao Huang, Wayne Xin Zhao, and Jun Li. 2024. Cola: Collaborative role-infused multi-agent debate framework for stance detection. *arXiv preprint arXiv:2403.01234*.
- Guan-Tong Liu, Yi-Jia Zhang, Chun-Ling Wang, Ming-Yu Lu, and Huan-Ling Tang. 2024. Comparative learning based stance agreement detection framework for multi-target stance detection. *Engineering Applications of Artificial Intelligence*, 133:108515.
- Qian Liu and 1 others. 2021. Automated pipeline for sentiment analysis of political tweets. In *Sentire@IJCNLP*, page –. Accuracy: 73.7% on 2020 US election tweets.
- Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.
- I. Orife, J. Kreutzer, D. I. Adelani, J. O. Alabi, S. H. Muhammad, A. Tapo, and 1 others. 2020. Masakhane: A grassroots nlp community for africa. *arXiv preprint arXiv:2003.11529*.
- Luca Sirrianni and Yu Zhang. 2021. Transformer-based models for stance detection: A comparative study. *Journal of Computational Social Science*, 4(2):225–238.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Ying Zhao and Qian Yang. 2020. Rumor detection and stance classification in social media. *Journal of Information Science*.
- Yuting Zhao, Lei Lin, and Xiaoming Liu. 2024. Mttf: A multi-perspective transferable feature fusion model for few-shot stance detection. *Information Processing & Management*, 61(2):103477.

Code-switching in Context: Investigating the Role of Discourse Topic in Bilingual Speech Production

Debasmita Bhattacharya and Anxin Yi and Siying Ding and Julia Hirschberg

Department of Computer Science

Columbia University

New York, NY, USA

debasmita.b@cs.columbia.edu, ay2616@columbia.edu, sd3609@barnard.edu, julia@cs.columbia.edu

Abstract

Code-switching (CSW) in speech is motivated by conversational factors across levels of linguistic analysis. While we know much about why speakers code-switch, there remains great scope for exploring how CSW occurs in speech, particularly within the discourse-level linguistic context. We build on prior work by asking: how are patterns of CSW influenced by different conversational contexts spanning *Academic*, *Cultural*, *Personal*, and *Professional* discourse topics? To answer this, we annotate a Mandarin-English spontaneous speech corpus, and analyze its discourse topics alongside various aspects of CSW production. We show that discourse topics interact significantly with utterance-level CSW, resulting in distinctive patterns of CSW presence, richness, language direction, and syntax that are uniquely associated with different contexts. Our work is the first to take such a context-sensitive approach to studying CSW, contributing to a broader understanding of the discourse topics that motivate speakers to code-switch in diverse ways.

1 Introduction

Code-switching (CSW) occurs when a multilingual speaker alternates between languages in speech or writing (Poplack, 1980). Speakers can code-switch between or within utterances across a variety of language pairs, producing a) syntactically simple *insertional* code-switches of single words or short phrases, or b) more syntactically complex *alternational* code-switches at grammatical clause boundaries (Muysken, 2000),¹ e.g.:

- (1) a. "让我拿出我的 calculator."
["Let me get out my calculator."]
- b. "我不懂 but the result isn't out yet."
["I don't understand but the result isn't out yet."]

¹Insertional and alternational code-switches are known as different *strategies* of code-switching.

Prior work has examined why speakers code-switch, showing the influence of various conversational factors: speaker competency, linguistic context, affective state of the speaker, the type of information the speaker wants to convey, and listener identity, among many others (Dornic, 1978; Bell, 1984; Gardner-Chloros, 2009; Broersma, 2009; Ferreira, 2017; Bhattacharya et al., 2024b). While much work has focused on the psycho-, socio-, and paralinguistic motivations for CSW, some studies have proposed alternative explanations of CSW based on discourse-level analysis. Early discourse-functional work on code-switched speech, e.g. Blom and Gumperz (1972); Auer (1998), suggested that CSW indicates a shift in topic during spontaneous conversations. This claim has held true in more recent studies across speech settings and language pairs (see Section 2). However, little is known about the *types* of topics that tend to elicit CSW, or how different genres of topic motivate distinctive patterns of downstream code-switched language production, particularly from a quantitative perspective across large-scale datasets of conversational speech. So, while we know much about *why* speakers code-switch in speech, there remains great scope for exploring *how* CSW occurs, especially within the discourse-level linguistic context.

We begin this research by studying the extent to which the topic of bilingual Mandarin-English conversations interacts with the presence, quantity, frequency, language direction, and syntactic complexity of CSW in spontaneous speech. We do so by examining an augmented version of the SEAME corpus of code-switched speech (Lyu et al., 2010) using statistical and unsupervised learning approaches, finding not only that differences in discourse topics interact significantly with CSW, but also that these interactions result in distinctive patterns of CSW features that can be used to distinguish between conversational contexts.

Our contributions include 1) producing a man-

ually annotated version of the SEAME corpus for new aspects of CSW, which we share at <https://tinyurl.com/3ac6jv2b>; 2) building a topic classifier for automatic annotation that is robust to both monolingual and multilingual Mandarin and English speech; and 3) identifying novel and nuanced quantitative and qualitative insight into the influence of discourse on multiple aspects of CSW. Overall, we contribute to a broader understanding of the conversational contexts that motivate speakers to code-switch in diverse ways. We hope that this work will inform innovation in robust spoken language technology that is capable of both understanding and producing code-switched speech grounded in naturalistic aspects of multilingual discourse function.

2 Prior Work

The earliest work on discourse aspects of CSW focused on defining taxonomies of when and why speakers code-switch. Notably, [Blom and Gumperz \(1972\)](#) proposed a dichotomy between *situational* CSW that indicates the topic of conversation, and *metaphorical* CSW for signaling emphasis; the combination of these allowed for the prediction of language choice among bilingual speakers in Norway. Though the precise boundary between situational and metaphorical CSW has been the subject of debate among authors such as [Auer](#) and [Wei](#), subsequent work has supported the claim that CSW serves as a discourse context cue that signals a semantic shift in the topic of Italian-German and Cantonese-English dialogue ([Auer, 1998](#); [Wei, 1998](#); [Auer, 2003](#)). Ethnographic studies of Spanish-English, e.g. [Lowi \(2005\)](#), have similarly shown that both intra- and inter-sentential CSW is used as a discourse feature to indicate change of topic among adult bilinguals of varying linguistic ability. These results generalize to the speech of children, in which topic and situation shift signaling is found to be a primary function of Spanish-English CSW ([Reyes, 2004](#)). Such a relationship between discourse framing and CSW has been observed in other language pairs, including Malaysian-English ([Ariffin and Rafik-Galea, 2009](#)), Bangla-English ([Das, 2012](#)), and Hindi-English ([Dey and Fung, 2014](#); [Begum et al., 2016](#)), and across spoken and written modalities.

By primarily conducting qualitative examinations of small-scale and hand-curated speech corpora, and analyzing coarse-grained CSW charac-

teristics (e.g. the number of code-switches in a given dialogue), the existing work described above has consistently established a link between discourse framing and CSW. However, few studies have extended this to consider finer-grained aspects of code-switched speech in relation to the nature of specific discourse topics. This is especially striking for the Mandarin-English language pair, given the vast number of global Mandarin native speakers,² many of whom are bilingual in English and code-switch regularly. Prior work on the high-level processing of topics in code-switched social media text and speech ([Peng et al., 2014](#); [Asnani and Pawar, 2016](#); [Rabinovich et al., 2019](#)) has targeted making topic modeling techniques robust to multilingual inputs, rather than identifying a deeper understanding of the types of topics that are associated with specific CSW behaviors. To address this gap, we ask **RQ**: How are specific discourse topics associated with patterns in the presence, quantity, frequency, direction, and syntactic complexity of CSW in a conversational domain?

3 Corpus

We examine the Mandarin-English Code-switching in South-East Asia (SEAME) corpus of spontaneous speech ([Lyu et al., 2010](#)). This corpus is made available by the [LDC User Agreement for Non-Members](#). SEAME consists of 192 hours of speech and 1,074,032 transcribed words across 256 dialogues. 156 unique speakers from Singapore and Malaysia are represented in the corpus. All dialogues are in an informal register, whether they were recorded in open-domain *conversation* settings or slightly more structured *interview* settings. Recordings comprise a mix of monolingual and code-switched utterances, the latter of which are Mandarin-dominant with inter-sentential code-switches to English. The corpus-level token ratio of Mandarin to English is 1.54:1.

4 Method

Data annotation and pre-processing: aspects of CSW. We annotate SEAME for the different aspects of CSW performed by speakers. First, we inspect the 110K utterance-level transcripts and automatically label each one for whether it is code-switched or monolingual, based on the Simplified Chinese and English orthographies used in the corpus. For the code-switched utterances, we also

²Almost 1 billion, per [Ethnologue](#) as of early 2025.

calculate utterance-level CSW quantity using the CSW ratio and M-index metrics (Soto et al., 2018; Barnett et al., 2000) and CSW frequency using the I-index metric (Guzman et al., 2016). We provide complete definitions of these metrics in Appendix A. We then perform additional manual annotations on the code-switched utterances to distinguish between insertional (I), alternational (A), and “other” (O) forms of CSW. We define “other” CSW similarly to tag-switching (Poplack, 1980), as the strategy used in any utterance where the code-switch is a filler word at the outset or end of a sentence. All annotation is performed by the second and third authors, who are native speakers of Mandarin with first-language proficiency in English. When the annotators disagree on a label, which occurs in less than 1% of utterances, they discuss their reasoning with each other until the disagreement is resolved.

We find that 48% of utterances in the corpus are monolingual. Among the 52% of code-switched corpus utterances, 89% use insertional CSW, 12% use alternational CSW, and 7% use “other”³ CSW.

Classifier construction: discourse topic labeling. Given the dataset’s size, instead of performing a second round of manual annotation of discourse topics over the entire corpus, we use a multi-class classifier to approximate utterance-level ground truth labels of discourse topics. Rather than unsupervised topic modeling, we use classification to approximate ground truth labels since we have some prior knowledge of the discourse topics present in the corpus. To do so, we train⁴ and evaluate four classifier models on a 10% sample of the corpus (11K utterances; the **ground truth** set), which we manually annotate for topic using the same protocol for resolving disagreement as above; label disagreements occur in less than 5% of utterances. Full task instructions are in Appendix A. We then apply the best-performing classifier for inference on the remainder of the corpus.

We begin with a rule-based approach and define a set of seed words in English and Mandarin as the lexicon associated with each of the following broad topic areas: *Academic*, *Cultural*, *Personal*, *Professional*.⁵ We choose this particular set of

topics based on those that were used to elicit speech during the collection of the corpus by Lyu et al., and our own observations of the data during our initial annotation pass, as these are most likely to reliably reflect the topics actually present in the data, and to avoid unnecessary complexity for an already time-intensive manual annotation task. We further justify this choice of topics experimentally in Appendix A. We define an additional *Other* topic to account for utterances that do not fall into any of the above topic areas. At inference time, we assign topic labels by identifying the number of exact matches with each topic’s lexicon and breaking ties at random. Any utterance with zero matches is labeled as *Other*. We assess the performance of this initial classifier on both the entire ground truth set and a test-time subset of it, for consistency with subsequent models.

We then refine our rule-based approach by expanding our handcrafted lexica with lexical and conceptual synonyms, applying Havaldar et al. (2024)’s method. This involves choosing the ten most similar neighbors per seed word using a cosine similarity threshold greater than or equal to 0.9 on pre-trained GloVe embeddings (Pennington et al., 2014).⁶ We incorporate these synonyms into our existing lexica, then perform inference with the expanded classifier and evaluate performance on the ground truth set and its test-time subset.

Next, we use `scikit-learn 1.6.1` to train a stacking ensemble (291K params.) combining three individually calibrated base learners: logistic regression, random forest, and gradient boosting. The ensemble takes as input utterance-level speaker gender, dialogue type, token length, presence of filler words (see list in Appendix A), presence of corpus-level frequent words, unigram tf-idf statistics, speaking rate, duration, and pause rate, in addition to weighted counts of seed words from the discourse topic lexica. We train this classifier using a logistic regression meta-learner on 80% (8.8K utterances) of the ground truth set, blending predictions with 3-fold cross-validation, and evaluate its performance on the remaining 20% of the ground truth set (2.2K utterances; the **held-out test** set).

Finally, we use a self-supervised learning approach, with a class-weighted logistic regression base model (17K params.). The input features to this model are the same as those used by the ensemble classifier, with additional bigram tf-idf statistics

³We retain all annotations of this CSW strategy in our augmented version of the SEAME corpus, but largely exclude this strategy from our subsequent statistical analyses for simplicity.

⁴Models are trained in about an hour on a Mac M1 chip.

⁵For detailed definitions and examples of each, as well as a complete list of seed words in each lexicon, see Appendix A.

⁶`glove-wiki-gigaword-50` accessed via Gensim.

and pre-trained sentence embeddings sourced from HuggingFace⁷ (33M frozen params.). This model iteratively generates pseudo-labels for unlabeled utterances in each of up to five rounds. For each discourse topic, we select up to a fixed quota of the highest-confidence predictions, using specific thresholds tuned per class that particularly include a dynamic adjustment for the relatively sparsely represented *Cultural* class. We treat 70% (7.7K utterances) of the ground truth set as the pseudo-train set and append newly pseudo-labeled examples to it before retraining and recalibrating the model. We monitor performance on a separate 10% subset of the ground truth set (1.1K utterances), using macro F1 to determine early stopping, and evaluate final model performance on the same held-out test set as above.⁸

Statistical analysis. Once the corpus is labeled for all CSW features of interest and discourse topics, we examine the relationship between these by using chi-squared and one-way ANOVA tests.

Clustering analysis. We build on significant statistical results by using `scikit-learn 1.6.1` to perform k -means clustering (10 params.) on vectors representing utterance-level binary CSW presence, strategy, and language direction, and CSW quantity and frequency, standardized to zero mean and unit variance. We then examine the resulting clusters and compare their composition over discourse topics and each CSW feature of interest.

5 Results

5.1 ML models outperform rule-based classifiers on discourse topic labeling.

We first calculate the expected blind guessing, i.e. random, baseline accuracy on our data, given the distribution of discourse topic labels in the held-out test set: 0.33. We subsequently use this value to contextualize the performance of our models.

We find that all four of our models significantly outperform the calculated baseline over the held-out test set (Table 1).⁹ As expected, the performance of our rule-based classifiers is generally inferior to that of the machine learning models, given that certain characteristics of the discourse topic cannot be captured by the raw content of an utter-

ance alone. Somewhat unexpectedly, the expanded lexicon-based classifier performs worse than the initial rule-based one, indicating that certain synonyms effectively *dilute* associations with specific discourse topics. On the other hand, the relative performance of the two machine learning models aligns with our expectations, as the self-supervised classifier demonstrates its unique ability to leverage large quantities of unlabeled data during learning. However, despite both machine learning models’ relative superior performance, we note the overall difficulty of utterance-level discourse topic labeling, reflected by the modest absolute value of all four classifiers’ task accuracy.¹⁰

Classifier	Accuracy	F1 Score
Initial lexicon-based	0.62	0.60
Expanded lexicon-based	0.60	0.59
Ensemble: LR, RF, GB	0.68	0.62
Self-supervised LR	0.72	0.71

Table 1: Classifiers’ accuracy and macro F1 score on discourse topic labeling over the held-out test set. We also report per-class performance metrics for the best-performing model in Table 14 in Appendix A.11.

Following training and evaluating the four classifiers on our sample of ground truth data, we use the self-supervised model to infer discourse topics for the remaining 90% of the corpus (99K utterances) that consists of unlabeled utterances. This results in the utterance-level distribution of discourse topics shown in Table 2, which suggests that certain conversational contexts are more popular than others.

Discourse topic	% of corpus
Academic	7.8
Cultural	0.1
Personal	28.1
Professional	2.4
Other	61.5

Table 2: Discourse topic label distribution across classes in the entire SEAME corpus. Please see Table 15 in Appendix A.12 for distributions across the ground truth and automatically-annotated subsets of the corpus.

We note that utterances on *Other* topics dominate the corpus, aligning with expectations for open-domain dialogue, and supporting our definition of this category to account for most utterances. To verify the absence of hidden clusters of topics within the *Other* category, we use exploratory LDA

⁷[sentence-transformers/all-MiniLM-L6-v2](#)

⁸Hyperparameter values for both the ensemble and self-supervised models are in Appendix A.

⁹For the rule-based classifiers’ performance over the entire ground truth set, see Appendix A.

¹⁰See ablation studies in Appendix A for relative contributions of different features to topic classification performance.

and BERTopic models (Blei et al., 2003; Grootendorst, 2022).¹¹ Both show that *Other* utterances are typically too short¹² to clearly denote any topic, and primarily consist of function words like wh-question words and deictic pronouns, fillers (e.g. “um”, “loh”), and temporal or connective markers (e.g. “then”, “如果”, “after”). While some functional groupings are present, we conclude that further subdividing the *Other* topic is not justified, as no additional coherent *semantic topics* emerge from this analysis.

The *Personal* topic is the next-most highly represented at the corpus-level, accounting for close to a third of utterances. This reflects how everyday conversations often revolve around personal anecdotes, thoughts, opinions, and feelings, validating our choice to examine it as a core discourse topic. The *Academic* and *Professional* topics are also present in the corpus, though their representation is relatively modest. And, while the *Cultural* topic is rare within the SEAME corpus, note that it still accounts for hundreds of unique utterances.

5.2 Discourse topics interact significantly with CSW presence, strategy, direction, quantity, and frequency.

Having a fully labeled corpus, we begin our statistical study of how discourse topics interact with aspects of CSW production in SEAME by considering differences in topic between monolingual and code-switched utterances. We control for utterances that are greater than 6 tokens in length across topics, since utterances in the *Other* topic are notably shorter than those in the other four topics.

Chi-squared tests comparing monolingual and code-switched utterances by topic all yield significant results, with clear patterns in associated odds ratios (Table 3). The odds that an utterance is about *Academic*, *Cultural*, *Personal*, or *Professional* topics, given it is code-switched, are at least two times those for a monolingual utterance. It seems that certain discourse contexts significantly lend themselves to multilingual, rather than monolingual, production, which is particularly noteworthy given their collective minority representation in the corpus overall. In contrast, the odds that an utterance is about any *Other* topic, given it is code-switched, are only about two-thirds of those for a monolin-

gual utterance, suggesting that *Other* topics are much better expressed in a monolingual fashion.

Topic	χ^2	<i>p</i> -val.	OR	95% CI
Academic	443.3	**	1.92	[1.81, 2.05]
Cultural	13.7	**	2.48	[1.52, 4.27]
Personal	430.2	**	2.49	[2.42, 2.56]
Professional	357.4	**	3.16	[2.78, 3.59]
Other	448.1	**	0.68	[0.66, 0.70]

Table 3: Chi-squared tests and odds ratios comparing topics in monolingual and code-switched utterances. Odds ratios >1 favor CSW. Odds ratios <1 favor monolinguality. *p*-values less than 0.01 are denoted by **.

Honing in specifically on CSW, utterances on *Academic*, *Professional*, or *Other* topics are more likely to be insertional code-switched than alternational code-switched (Table 4), reflecting how specific discourse contexts intersect with the strategy of CSW that is most represented in the corpus. Each of these topics also has greater representation of insertional CSW than the corpus overall (94% on average, compared to corpus-level 89%), reinforcing the influence of topic on CSW strategy. Interestingly, in *Cultural* and *Personal* topics, insertional CSW is equally and one-third as likely as alternational CSW, respectively, indicating that the relatively more complex CSW strategy is more suited to conversation topics that may require less subject knowledge to discuss, while the opposite was true for topics that may be more difficult to speak on. These patterns suggest that speakers might attempt to achieve a balance in complexity between the nature of the discourse topic under discussion and the CSW strategy used to express it when producing CSW. Given that both topics’ representation of alternational CSW is the same as the corpus overall (about 12% in all cases), these topics’ relative skew towards alternational CSW in our calculated odds ratios is even more striking.

Topic	χ^2	<i>p</i> -val.	OR	95% CI
Academic	28.5	**	1.25	[1.15, 1.36]
Cultural	435.0	*	1.03	[0.61, 1.86]
Personal	2029.2	**	0.31	[0.30, 0.33]
Professional	54.7	**	1.79	[1.53, 2.11]
Other	1430.0	**	2.98	[2.81, 3.16]

Table 4: Chi-squared tests and odds ratios comparing topics in insertional and alternational CSW. Odds ratios >1 favor insertional CSW. Odds ratios <1 favor alternational CSW. *p*-values less than 0.05 and 0.01 are denoted by * and **, respectively.

¹¹Hyperparameter details are in Appendix A.

¹²Mean and standard deviation token length for *Other* utterances are 6.5 and 6.2, respectively. Across all other utterances, these are 15.1 and 9.3, respectively.

With respect to language direction of CSW, *Academic* and *Professional* utterances are about two-thirds as likely to be code-switched from English to Mandarin, as from Mandarin to English, reflecting the frequency of insertion of English technical, jargon-like, and/or domain-specific terms into such multilingual utterances (Table 5). *Personal* and *Cultural* utterances have relatively higher odds of being code-switched from English to Mandarin, and are about equally likely to be code-switched from Mandarin to English. Only utterances on *Other* topics are significantly more likely to be code-switched from English to Mandarin at an odds ratio of 1.22, which is especially worth noting given the overall skew of the corpus towards Mandarin.

Topic	χ^2	<i>p</i> -val.	OR	95% CI
Academic	78.0	**	0.75	[0.70, 0.80]
Cultural	0.007	–	0.96	[0.61, 1.47]
Personal	0.5	–	0.99	[0.95, 1.03]
Professional	91.7	**	0.57	[0.51, 0.64]
Other	102.6	**	1.22	[1.17, 1.27]

Table 5: Chi-squared tests and odds ratios comparing topics in English-to-Mandarin (en \rightarrow zh) and Mandarin-to-English (zh \rightarrow en) CSW. Odds ratios >1 favor en \rightarrow zh. Odds ratios <1 favor zh \rightarrow en. *p*-values less than 0.01 are denoted by **. *p*-values more than 0.05 are denoted by –.

Next, we perform one-way ANOVA tests to compare CSW quantity and frequency metrics across the different discourse topics. For each of CSW ratio, M-index, and I-index, ANOVA tests show a strong and statistically significant ($p < 0.01$) association between each topic and the metric of CSW richness. This statistical significance holds even after applying Bonferroni correction to account for possible noise in discourse topic labels generated by our best-performing classifier from Section 5.1. These associations suggest that variations in discourse topic can distinguish CSW behavior in terms of both quantity and frequency of utterance-level CSW. That is, there are significant differences in CSW richness in utterances on different topics. More concretely, we find that the *Personal* and *Cultural* topics consistently rank the lowest in terms of mean CSW richness across metrics, while the *Professional* and *Academic* topics are the two most highly ranked across the board. The *Other* topic sits in the middle of the ranking in each case. These results provide further evidence of a relationship between discourse-level conversational context and various aspects of CSW behavior in SEAME.

Overall, the results of our statistical analysis reveal significant interactions between specific discourse topics and granular patterns of code-switched speech production in SEAME. Not only are utterances on *Academic*, *Cultural*, *Personal*, and *Professional* topics more likely to be expressed using CSW, but each of these topics also has a unique, typical CSW profile. Multilingual utterances on *Academic* and *Professional* topics are characterized by higher quantity and frequency of CSW, with the majority of such code-switches taking place from Mandarin to English in an insertional fashion. In contrast, *Personal* and *Cultural* utterances are characterized by fewer and less frequent alternational code-switches from Mandarin to English. Utterances on *Other* topics are overall less likely to be code-switched; when such utterances are expressed multilingually, these are less striking in their CSW quantity or frequency, but are more likely to involve insertional code-switches from English to Mandarin. These findings provide the motivation for the remainder of the work.

5.3 Unsupervised models learn many discourse-CSW relationships.

Having found multiple significant interactions between discourse topics and several fine-grained aspects of CSW behavior in SEAME, we further develop our investigation by assessing whether these relationships are salient enough to be learned by unsupervised models, and potentially in turn inform the downstream outputs of such models. Instead of methods like LDA that explicitly group datapoints by topic, we want to see if unsupervised models that do not have this specific topic-centric objective can still cluster utterances based on both topic and CSW information, as a stronger test for the validity of associated patterns. To do so, we implement *k*-means clustering, setting $k = 5$ to match the number of distinct discourse topic labels, with random starting points¹³ and principal component analysis. We then compare resulting cluster compositions across discourse topic and CSW presence, strategy, language direction, and richness, verifying the significance of these groupings using chi-squared tests. Throughout this section, we discuss only the comparisons yielding significant *p*-values.

We begin by comparing cluster compositions across topics and CSW presence, and find a clear separation between Cluster 2 and the remaining

¹³We motivate this design choice further in Appendix A.

Topic	C1	C2	C3	C4	C5
Acad.	6.3%	3.6%	13.2%	8.6%	10.0%
Cult.	0.2%	0.1%	0.2%	0.1%	0.9%
Pers.	29.8%	18.8%	39.0%	30.1%	62.2%
Prof.	1.2%	0.7%	4.8%	2.1%	2.7%
Other	62.5%	76.7%	42.7%	59.1%	24.3%

Table 6: Cluster composition by discourse topic.

CSW?	C1	C2	C3	C4	C5
No	0%	99.9%	0%	0%	0%
Yes	100%	0.1%	100%	100%	100%

Table 7: Cluster composition by CSW presence.

four clusters in terms of multilinguality of constituent utterances; this cluster is almost entirely dominated by non-code-switched utterances (Table 7), while also representing *Other* topics most highly (Table 6), in a clear reflection of the specific association between discourse and monolingual expression we have found in Section 5.2. Clusters 1, 3, 4, and 5 are all dominated by CSW, and each represents a mix of discourse topics. Cluster 5 is most representative of *Personal* utterances while Cluster 3 contains a combined majority mix of *Academic* and *Personal* topics. Although these patterns do not exactly align with our initial hopes of obtaining five distinct clusters, each of which is uniquely dominated by one of the discourse topics, these are still interesting as they mirror many of our earlier statistical findings. We hypothesize that the absence of a clear *Professional* or *Cultural* cluster may be due to the relatively lower representation of these discourse topics in the corpus overall (Table 2).

Strategy	C1	C2	C3	C4	C5
I	58.7%	0.1%	95.8%	82.0%	27.0%
A	12.5%	0.0%	1.6%	10.5%	3.6%
O	28.8%	0.0%	2.6%	7.4%	69.4%
None	0.0%	99.9%	0.0%	0.0%	0.0%

Table 8: Cluster composition by CSW strategy.

Considering cluster compositions in Table 8, we again find patterns of overlap between utterance-level CSW strategies and discourse topics that align with those found in Section 5.2. For instance, Cluster 3, which we have already noted for its representation of *Academic* utterances, while simultaneously representing the greatest proportion of *Professional* utterances relative to other clusters, also contains the greatest proportion of insertional CSW. This reinforces the strength of the interaction be-

tween discourse and CSW strategy for these topics. Similarly, Cluster 5, which is dominated by the *Personal* topic and contains the greatest proportion of the *Cultural* topic relative to other clusters, has the smallest gap in representation between insertional and alternational CSW. This aligns with our statistically significant observation that these topics are less associated with insertional than alternational CSW. However, we also note the overall lower proportion of alternational CSW in each cluster, and hypothesize that this may be due to the relative infrequency of this CSW strategy in the corpus compared to insertional CSW, as noted in Section 4.

Next, we compare cluster compositions across CSW language direction (Table 9) and metrics of CSW quantity and frequency (Table 10). In the case of the latter, we transform utterance-level CSW ratio, M-index, and I-index into binary variables by denoting values less than the median of each metric of CSW richness as “low” and values greater than or equal to the median as “high”.

CSW dir.	C1	C2	C3	C4	C5
en → zh	100%	0.1%	15.6%	0%	64.9%
zh → en	0%	0.0%	84.4%	100%	35.1%

Table 9: Cluster composition by CSW language direction: English-to-Mandarin or Mandarin-to-English.

With respect to CSW language direction, Cluster 1, which primarily consists of utterances on *Other* topics, is made up exclusively of code-switches from English to Mandarin. This is striking as we know the *Other* topic is the only one that is significantly more likely to be expressed in such a fashion. Cluster 3, whose combined majority topic representation is from the *Personal* and *Academic* topics, is dominated by Mandarin to English CSW, which also aligns with our previous finding, since the *Academic* topic in particular is more likely to be code-switched in this direction. Cluster 5’s English-to-Mandarin dominance is also interesting, and is likely due to the presence of the *Personal* and *Other* topics, the former of which is equally likely to be code-switched in either direction, and the latter of which is always more likely to be code-switched from English to Mandarin; their combination likely determines the overall cluster composition in terms of CSW language direction. Recall that Cluster 2 effectively contains no CSW (Table 7) and hence does not contain CSW in either direction.

Metric	C1	C2	C3	C4	C5
R:H	90.1%	0%	98.9%	90.0%	0%
R:L	9.9%	100%	1.1%	10.0%	100%
M:H	95.9%	0%	98.1%	87.5%	0%
M:L	4.1%	100%	1.9%	12.5%	100%
I:H	90.1%	0%	98.7%	90.0%	0%
I:L	9.9%	100%	1.2%	10.0%	100%

Table 10: Cluster composition by metrics of CSW richness: CSW ratio (R), M-index (M), and I-index (I), binned into high (H) and low (L) values.

Finally, we examine cluster composition across metrics of CSW quantity and frequency. We find that the distribution of high vs. low values of each metric in Cluster 5 supports our previous finding that *Personal* and *Cultural* topics always contain the lowest quantity and frequency of CSW. Similarly, Cluster 3 reinforces how the *Academic* topic always has the highest values across metrics. The composition of Cluster 4 also demonstrates how the *Other* and *Academic* topics, which we know are associated with mid to high levels of CSW richness, pull metric values up within the cluster.

Overall, our clustering model is able to group utterances according to both topic and CSW characteristics, which indicates that it can learn relationships between topics and CSW patterns in a reasonable way. These results demonstrate that many of the statistical relationships we have found between discourse topics and various fine-grained aspects of CSW behavior in SEAME are significant enough to be learned by unsupervised models, and may well inform their downstream outputs, though we leave a detailed investigation of the latter claim to future work. A random baseline analysis confirms this conclusion and validates that our current clusters particularly capture topic structure beyond chance.¹⁴ General agreement between our comparative clustering analyses and initial statistical findings lends validity to the latter, demonstrating their value in understanding *and* modeling CSW.

6 Discussion

We find that specific discourse topics have notable relationships with several fine-grained aspects of Mandarin-English CSW in SEAME. These utterance-level relationships are sufficiently strik-

¹⁴Cramer’s V measures show that the strength of association between topic labels and current clusters is several orders of magnitude greater than between topic labels and random, size-matched clusters (0.168 vs. 0.005; both p-values < 0.01).

ing as to produce distinct values of CSW features across dimensions of multilingual spoken behavior that effectively distinguish between the topics being discussed in those code-switched utterances.

Our exploration and subsequent findings on how discourse topics relate to the presence of CSW echo and validate prior work on code-switches functioning as signals of topic shift, e.g. Wei (1998); Auer (2003), while our study of CSW quantity, frequency, language direction, and strategy reveal novel associations with topic at a level of detail previously not attained in discourse-functional work on CSW. Specific CSW patterns that group the *Academic* and *Professional* topics, and the *Personal* and *Cultural* topics, are reminiscent of prior qualitative work on an emotional detachment effect in CSW scenarios (Ladegaard, 2018; Ferreira, 2017). We speculate that the affective properties of certain kinds of discourse topics may similarly help determine the CSW style used to express them. Using Mandarin and English emotion lexica (Mohammad and Turney, 2010, 2013), we preliminarily find that utterances on *Personal* and *Cultural* topics have significantly greater emotional intensity than those on *Academic* and *Professional* topics (details in Appendix A). This aligns with our results on the association between discourse topic complexity and CSW strategy, and suggests that affect may modulate this interaction, though further work is required to confirm this hypothesis. Separately, we show that more formal topics (i.e. *Academic*, *Professional*) can involve CSW in speech, unlike prior studies that have primarily noted CSW in informal contexts, e.g. Bhattacharya et al. (2023). Finally, we show that many of the relationships we find can even be learned and applied by a simple unsupervised clustering model, lending validity to our statistical findings in a clearly interpretable manner.

7 Conclusion

We extensively examine the relationship between discourse topic and patterns of spontaneous CSW in the SEAME corpus. We find that (1a) certain discourse topics are much more likely to be expressed in code-switched utterances than monolingual ones; (1b) those discourse topics have significant associations with multilingual language production across previously unexamined patterns of CSW strategy, language direction, quantity, and frequency; (2) these associations lend themselves towards the inference of unique CSW profiles linked to specific

(groups of) topics; (3) the statistical relationships found in (1) and (2) are salient enough to be learned and applied in part by unsupervised clustering models. We conclude that the nature of the discourse topic in conversation contributes meaningfully towards motivating diverse patterns of Mandarin-English CSW in speech. Our work’s novelty is based in its context-sensitive approach towards understanding a dataset that we augment with new annotations across features of discourse and CSW. We hope this work will serve as a first step towards building improved models of CSW comprehension and informing the generation of authentic and discourse-informed multilingual speech.

Limitations

Our work focuses on a single language pair in a single corpus of CSW, which is somewhat skewed towards Mandarin relative to English. Both languages are represented only in the forms in which they are typically spoken in Singapore and Malaysia, in contrast to the majority of Mandarin-English code-switched corpora that are sourced from Mainland Chinese speakers. We acknowledge the need to extend our methods to the same language pair within different cultural contexts, and to additional language pairs with varying levels of typological distance, to test the robustness of our findings. We plan to do so in future work. Due to lack of access to CSW datasets, particularly those containing highly time-intensive manual discourse-level annotations and/or less discourse topic sparsity than in SEAME, our work makes use of the best currently available resources and serves as a reasonable first step towards understanding the role of discourse topics on code-switched speech production. For the *Cultural* topic in particular, we acknowledge that the relative corpus-level representation of this discourse topic in SEAME makes the associated findings, though novel and insightful, difficult to generalize. We are very interested in ultimately replicating our analyses on other CSW datasets, but also note that direct comparisons may be difficult since the categories and distributions of topics may differ across datasets.

With respect to our discourse topic classifiers, we note the inherent limitation of a single utterance receiving only a single label in our multi-class setup. By definition, this model design choice ignores the possibility of certain utterances dealing with multiple topics at a time by collapsing predictions

into a single output label. Given the number of discourse topics we examine in this work, we believe this was nonetheless a reasonable design choice that prevented subsequent analyses from becoming overly complex.

Relatedly, it could have been helpful to incorporate additional features, such as Linguistic Inquiry and Word Count (LIWC) labels (Boyd et al., 2022), into our machine learning discourse topic classifiers. We speculate that such features covering psychological processes and personal concerns could have augmented the performance of our supervised models. However, it is difficult to reliably extract LIWC features from code-switched language, as this framework was originally developed for use in monolingual settings, and we leave this methodological extension to future work.

Finally, while our best-performing classifier achieves an accuracy of 72%, which is well above baseline performance, there remains 28% error in subsequently inferred discourse topic labels. This residual noise in the data could impact downstream statistical analyses. We handle this using error aware correction in our one-way ANOVA tests, and preliminarily find in Appendix A.12 that any remaining noise effectively has no impact on our current results. However, a fruitful direction for future work would be to replicate these downstream results by exploring alternative methods for deriving discourse topic labels, such as pre-trained multilingual transformer models and LLMs, e.g. mBERT or zero-shot GPT. We chose not to use these in the present work primarily in order to avoid issues arising from domain mismatch in pre-training data, which may not be sufficiently mitigated through fine-tuning due to a scarcity of appropriate code-switched data, as well as the relatively lower inherent transparency, interpretability, and modularity of these methods in comparison to each of our four classifiers. However, we acknowledge that in future work it may be worth trading off the drawbacks of these methods, as well as relevant cost and feasibility concerns, in favor of their potential to boost classification performance, which would increase the reliability of downstream analyses. Our deliberate design choice to avoid such models in the present work is particularly relevant since our main contribution is not to provide state-of-the-art model performance, but rather to leverage our current custom models to augment data and provide nuanced insights on that data.

Ethical considerations

This study was conducted exclusively on secondary data, and did not require human experiments. We did not access any information that could uniquely identify individual users within the corpus, as its original authors de-identified all speakers as outlined in the documentation of the dataset. Though we did not collect the data used in this work, we note that all participants in the original corpus had consented to sharing the data that we analyze in our study.

Acknowledgments

We thank Nicholas Deas, Lin Ai, and Chhavi Dixit for helpful discussions and feedback. This work was supported in part by the National Science Foundation under Grant IIS 2418307.

References

- Kamisah Ariffin and Shameem Rafik-Galea. 2009. Code-switching as a communication device in conversation. *Language & Society Newsletter*, 5(9):1–19.
- Kavita Asnani and Jyoti D Pawar. 2016. [Use of semantic knowledge base for enhancement of coherence of code-mixed topic-based aspect clusters](#). In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 259–266, Varanasi, India. NLP Association of India.
- JC Peter Auer. 2003. A conversation analytic approach to code-switching and transfer. In *The Bilingualism Reader*, pages 167–187. Routledge.
- Peter Auer. 1984. On the meaning of conversational code-switching.
- Peter Auer. 1998. *Bilingual Conversation Revisited*, pages 1–24. Routledge, London, UK.
- Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietske Wensing. 2000. [The LIDES Coding Manual: A document for preparing and analyzing language interaction data version 1.1—July, 1999](#). *International Journal of Bilingualism*, 4(2):131–270.
- Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. [Functions of code-switching in tweets: An annotation framework and some initial experiments](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).
- Allan Bell. 1984. [Language style as audience design](#). *Language in Society*, 13(2):145–204.
- Debasmita Bhattacharya, Jie Chi, Julia Hirschberg, and Peter Bell. 2023. [Capturing formality in speech across domains and languages](#). In *Interspeech 2023*, pages 1030–1034.
- Debasmita Bhattacharya, Siying Ding, Alayna Nguyen, and Julia Hirschberg. 2024a. [Measuring entrainment in spontaneous code-switched speech](#). 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 2865–2876, Mexico City, Mexico. Association for Computational Linguistics.
- Debasmita Bhattacharya, Eleanor Lin, Run Chen, and Julia Hirschberg. 2024b. [Switching tongues, sharing hearts: Identifying the relationship between empathy and code-switching in speech](#). In *Interspeech 2024*, pages 492–496.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

- Jan-Petter Blom and John J. Gumperz. 1972. *Social meaning in linguistic structure: code-switching in Norway*, pages 75–96. Routledge.
- Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. *The development and psychometric properties of LIWC-22*.
- Mirjam Broersma. 2009. *Triggered codeswitching between cognate languages*. *Bilingualism: Language and Cognition*, 12(4):447–462.
- Basudha Das. 2012. Code-switching as a communicative strategy in conversation. *Global Media Journal—Indian Edition*, 3(2):1–20.
- Anik Dey and Pascale Fung. 2014. *A Hindi-English code-switching corpus*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Stanislav Dornic. 1978. *The Bilingual’s Performance: Language Dominance, Stress, and Individual Differences*, pages 259–271. Springer US, Boston, MA.
- A. Virginia Acuña Ferreira. 2017. Code-switching and emotions display in Spanish/Galician bilingual conversation. *Text & Talk*, 37:47 – 69.
- Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. *Simple tools for exploring variation in code-switching for linguists*. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.
- Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. *Building knowledge-guided lexica to model cultural variation*. 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 211–226, Mexico City, Mexico. Association for Computational Linguistics.
- Hans J. Ladegaard. 2018. *Codeswitching and emotional alignment: Talking about abuse in domestic migrant-worker returnee narratives*. *Language in Society*, 47(5):693–714.
- Rosamina Lowi. 2005. Code switching: An examination of naturally occurring conversation. In *Proceedings of the 4th International Symposium on Bilingualism*, pages 1393–1406. Cascadilla Press Somerville, MA.
- Dau-Cheng Lyu, Tien-Ping Tan, Eng Chng, and Haizhou Li. 2010. *Mandarin–English code-switching speech corpus in South-East Asia: SEAME*. In *Language Resources and Evaluation*, volume 49, pages 1986–1989.
- Saif Mohammad and Peter Turney. 2010. *Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon*. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.
- Saif M. Mohammad and Peter D. Turney. 2013. *Crowd-sourcing a word-emotion association lexicon*. *Computational Intelligence*, 29(3):436–465.
- Pieter Muysken. 2000. *Bilingual speech: a typology of code-mixing*. Cambridge University Press.
- Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. *Learning polylingual topic models from code-switched social media documents*. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–679, Baltimore, Maryland. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. *GloVe: Global vectors for word representation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Shana Poplack. 1980. *Sometimes I’ll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching*. *Linguistics*, 18:581–618.
- Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. *CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums*. 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 4776–4786, Hong Kong, China. Association for Computational Linguistics.
- Iliana Reyes. 2004. *Functions of code switching in schoolchildren’s conversations*. *Bilingual Research Journal*, 28(1):77–98.
- Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. *The Role of Cognate words, POS Tags and Entrainment in Code-Switching*. In *Interspeech 2018*, pages 1938–1942.
- Li Wei. 1998. *The ‘Why’ and ‘How’ Questions in the Analysis of Conversational Code-Switching*, pages 156–176. Routledge, London, UK.

A Appendix

A.1 CSW quantity and frequency metric definitions.

CSW ratio measures the number of code-switches normalized by the token length of the utterance. This differs from M-index, which incorporates information about the utterance-level balance between language varieties present. Different from both metrics of CSW quantity, I-index reflects CSW frequency via the number of potential switch points in an utterance. All three metrics of CSW richness have a minimum value of 0, associated with monolingual utterances. The maximum value of CSW ratio approaches but does not equal to 1, while both M- and I-indices can achieve maximum values of 1, associated with a code-switched utterance evenly mixed between languages.

A.2 Discourse topic definitions and examples.

Academic: utterances discussing education (at primary, secondary, or university levels), research, studying, or coursework.

Example 1: 那种他 exam. ["The kind he takes on an exam."]

Example 2: 他是 full time tuition teacher. ["He is a full time tuition teacher."]

Example 3: 这样他还可以 graduate in four years 还蛮强的. ["This way he can graduate in four years, which is pretty impressive."]

Cultural: utterances discussing traditions, customs, festivals, cultural practices, holidays, or religious celebrations.

Example 1: 然后你会一直听到有人在唱 Christmas carols 然后很好玩. ["Then you'll keep hearing people singing Christmas carols and it's fun."]

Example 2: Oh Chinese New Year 每年都是一样的. ["Oh Chinese New Year is the same every year."]

Example 3: But 最少他还 preserve 他的那个 culture 他还是会 Hokkien. ["But at least he still preserves his culture and he still knows Hokkien."]

Personal: utterances discussing hobbies, day-to-day/habitual experiences, opinions, feelings, preferences, family, friends, or other relationships.

Example 1: 可是她的 boyfriend 是 like 做 ["But her boyfriend likes to do - "]

Example 2: 他就觉得 like 就跟那个 the feeling is not right anymore because 我明明知道你背叛过我这样. ["He feels like the feeling

is not right anymore because I clearly know you betrayed me.]

Example 3: I think when you break up 跟你们可以做继续做 friend 是因为 at that time 你们真的没有很在一起那种感觉. ["I think when you break up you can continue to be friends because at that time you didn't really feel like you were together.]

Professional: utterances discussing work life and technical aspects of a job, including the use of technology or programming.

Example 1: Actually what I did was to source for 他们客户的数据. ["Actually what I did was to source for their clients' data."]

Example 2: start 你的 business 你可以. ["You can start your business."]

Example 3: 就比如我 interested in supply chain 如果. ["For example, if I am interested in the supply chain."]

A.3 Instructions for discourse topic annotation task.

Please label the following utterances for topic of conversation discussed: Academic, Professional, Personal, Cultural, or Other. Below are guidelines to help you distinguish between topics.

- **Academic:** *Topics related to education, research, school, university, study, or coursework.*
 - *Examples: discussions about classes, research projects, GPA, or teachers.*
- **Professional:** *Topics related to work, technology, programming, or aspects of a job, especially technical ones.*
 - *Examples: discussions about work, company, office, salary, etc.*
- **Personal:** *Topics focused on personal life, hobbies, family, friends, feelings, or relationships.*
 - *Examples: conversations about family members, personal emotions, thoughts, preferences, or day-to-day/habitual experiences.*
- **Cultural:** *Topics related to traditions, festivals, cultural practices, holidays, or religious celebrations.*
 - *Examples: discussions about cultural holidays*/events, or traditional customs.*
**Not just mentions of vacations or trips.*

- Other: *If the utterance does not clearly fit into any of the above categories.*
 - Examples: *Short utterances will tend to fit into this topic.*

A.4 Validating the current set of core discourse topics.

We believe the current granularity of discourse topics studied represents a reasonable starting point for this work. To validate this, we perform additional exploratory topic modeling, with an automatic number of topics from a BERTopic model, on each predefined topic in the corpus to assess whether additional, finer-grained topics emerge from any (see Appendix A.13 for implementation details).

While each of the *Personal*, *Academic*, and *Professional* topics demonstrate related subdivisions, (discussing relationships with friends/family vs. thoughts/feelings/preferences; studying for exams vs. specific school subjects; technical jargon vs. professional roles and responsibilities), none of these is distinct enough from its parent topic to warrant defining a distinct new topic; see Appendix A.5 for detailed examples. Thus, we confirm the appropriate granularity of the set of topics we choose to study.

A.5 Examples of related but non-distinctive sub-topic clusters within *Academic*, *Personal*, and *Professional* topics.

A.5.1 *Academic*.

- Sub-topic 1 (discussing exam preparation): 学校, exam, study, 备考, studying.
- Sub-topic 2 (discussing specific school subjects): school, lecture, maths, science, german.

Other *Academic* sub-topics are defined mainly by functional words, indicating that more granular topics do not exist within the *Academic* topic, e.g. Sub-topic 3: 我们, then, 没有, 那些, 东西.

A.5.2 *Personal*.

- Sub-topic 3 (discussing thoughts and preferences): 觉得, 这样, 比较, like, think.
- Sub-topic 5 (discussing interpersonal relationships): 他们, my, friend, 我的, parents.

Other *Personal* subtopics are also defined mainly by functional words, indicating that more granular topics do not exist within the *Personal* topic, e.g. Sub-topic 1: that, is, it, and, then.

A.5.3 *Professional*.

- Sub-topic 3 (discussing professional roles and responsibilities): lead, business, project, manager, fulltime.
- Sub-topic 5 (discussing jargon-like aspects): processing, 一个, job, software, data.

Other *Professional* subtopics are defined with related words, but in a less clearly cohesive way, indicating that more granular topics do not exist within the *Professional* topic, e.g. Sub-topic 4: part, time, 我们, 现在, 做工.

A.6 Seed words used for initial rule-based classifier.

- *Academic*: course, class, unit, lesson, lecture, batch, final, review, conference, presentation, reference, archive, result, semester, sem, academic, student, admit, scholar, teacher, prof, report, learn, uni, school, book, chapter, syllabus, read, paper, essay, econ, math, physics, chem, bio, science, psychology, grade, point, score, credit, fail, committee, major, master, phd, thesis, module, subject, average, analyze, analyse, honours, junior, question, lab, diploma, percent, quiz, exam, tutor, tuition, enroll, prove, uniform, graduate, orientation, levels, recess, homework, primary, secondary, year, study, engineering, gpa, studies, college, research, education, edu, pre, 课程, 班级, 单元, 课程, 讲座, 批次, 期末, 复习, 会议, 演示, 参考, 存档, 结果, 学期, 学期, 学术, 学生, 录取, 学者, 老师, 教授, 报告, 学习, 大学, 学校, 书, 章节, 教学大纲, 阅读, 论文, 文章, 经济, 数学, 物理, 化学, 生物, 科学, 心理学, 成绩, 分数, 得分, 学分, 不及格, 委员会, 专业, 硕士, 博士, 论文, 模块, 科目, 平均, 分析, 分析, 荣誉, 初级, 问题, 实验室, 文凭, 百分比, 测验, 考试, 导师, 学费, 注册, 证明, 制服, 毕业, 迎新, 水平, 休息, 家庭作业, 小学, 中学, 年级, 学习, 班, 工程, 课, 章节, 章, 学院, 研究, 科研, 项目, 教育.
- *Cultural*: christmas, carol, halloween, new year, typical, traditional, pray, buddha, jesus, islam, church, god, goddess, red envelope, gathering, taoist, bible, scripture, christian, orthodox, holiday, holidays, religion, lantern, china, mid-autumn, 圣诞节, 圣诞, 万圣, 颂歌, 万圣节, 新年, 典型, 传统, 祈祷, 佛, 耶稣, 伊斯兰教, 教堂, 神, 女神, 红包, 聚会, 道教, 圣经, 经文, 基督教, 正统, 假日, 节

假日,元宵,元宵节,放假,宗教,红包,拜年,鞭炮,花灯,灯谜,中国,春节,过年.

- *Personal*: think, feel, feeling, understand, believe, trust, like, know, prefer, want, miss, worry, regret, stress, remember, happy, sad, afraid, miserable, excite, mad, anger, angry, friend, cousin, mother, father, mom, dad, parent, child, brother, sister, sibling, uncle, aunt, daughter, son, family, relative, husband, wife, boyfriend, girlfriend, fun, together, individual, personal, relationship, play, piano, rugby, tennis, badminton, soccer, football, basketball, hobby, usual, often, home, house, room, live, birthday, community, facebook, show, name, person, people, identity, favourite, favorite, swim, self, 思考, 感觉, 感受, 理解, 相信, 信任, 喜欢, 知道, 更喜欢, 想要, 想念, 担心, 后悔, 压力, 记住, 快乐, 悲伤, 害怕, 痛苦, 兴奋, 生气, 愤怒, 朋友, 表亲, 母亲, 父亲, 妈妈, 爸爸, 父母, 孩子, 兄弟, 姐妹, 兄弟姐妹, 叔叔, 阿姨, 女儿, 儿子, 家庭, 亲戚, 丈夫, 妻子, 男朋友, 女朋友, 乐趣, 一起, 个人, 私人, 关系, 玩, 钢琴, 橄榄球, 网球, 羽毛球, 足球, 篮球, 爱好, 平常, 经常, 家, 房子, 房间, 生活, 生日, 社区, 脸书, 表演, 姓名, 人物, 人物, 身份, 最喜欢, 游泳, 哥哥, 哥, 姐姐, 姐, 妹妹, 妹, 弟弟, 弟, 爸爸, 爸, 妈妈, 妈, 爷爷, 奶奶, 外公, 外婆, 祖父, 祖母, 自己, 目前, 开心, 伤心, 难过, 悲伤, 悲哀, 恐怖.
- *Professional*: freelance, position, job, part-time, occupation, apply, application, work, interview, dollar, cent, salary, technology, program, boss, colleague, staff, regression, model, correlation, correlate, download, sensor, email, database, internet, website, system, algorithm, bug, update, server, warranty, business, service, user, experience, audit, consult, career, manage, stock, portfolio, project, system, procedure, develop, design, quality, team, equipment, lead, produce, function, tool, skill, consumer, customer, employee, contract, information, solve, solution, profit, design, machine, paperwork, training, zoom, company, firm, hierarchy, maintain, chip, weld, manufacture, manufacturing, property, properties, special, identify, admin, bank, software, tech, troubleshoot, industry, data, recruit, hire, offer, trademark, market, competition, government, capital, promotion, reboot, protocol, profit, commission, downstream, commercial, indus-

trial, stage, tutorial, manage, manager, interface, 自由职业, 职位, 工作, 兼职, 职业, 申请, 申请, 工作, 面试, 美元, 分, 薪水, 技术, 程序, 老板, 同事, 员工, 回归, 模型, 相关性, 相关, 下载, 传感器, 电子邮件, 数据库, 互联网, 网站, 系统, 算法, 错误, 更新, 服务器, 保修, 业务, 服务, 用户, 经验, 审计, 咨询, 职业, 管理, 股票, 投资组合, 项目, 系统, 程序, 开发, 设计, 质量, 团队, 设备, 领导, 生产, 功能, 工具, 技能, 消费者, 客户, 员工, 合同, 信息, 解决, 解决方案, 利润, 设计, 机器, 文书工作, 培训, 缩放, 公司, 公司, 层次结构, 维护, 芯片, 焊接, 制造, 制造业, 财产, 属性, 特殊, 识别, 管理员, 银行, 软件, 技术, 故障排除, 行业, 数据, 招聘, 雇用, 提供, 商标, 市场, 竞争, 政府, 资本, 促销, 重新启动, 协议, 利润, 佣金, 下游, 商业, 工业, 阶段, 做工, 做, 工, 教程, 经理, 界面.

A.7 Performance of rule-based classifiers on full ground truth set.

Classifier	Accuracy	F1 Score
Initial lexicon-based	0.61	0.54
Expanded lexicon-based	0.59	0.53

Table 11: Rule-based classifiers’ accuracy and macro F1 score on discourse topic labeling over the entire ground truth set.

A.8 Filler words used in ensemble and self-supervised classifiers.

Um, ah, uh, eh, er, hmm, mm, mmm, umm, ar, hm, 啊, 呃, 呃, 啊, 呃, 嗯, 嗯, 哼.

A.9 Classifier models’ hyperparameter settings.

In the **ensemble classifier**, the `TfidfVectorizer` used has its maximum features set to 2000 and its analyzer set to ‘word’. All other parameters are left at default values. All numeric features used in this model are normalized to zero mean and unit variance using `StandardScaler`. The ensemble architecture consists of three base models and one meta model. The first base model is a logistic regression model with the following hyperparameters: $C = 0.9$, class weight = ‘balanced’, maximum iterations = 3000, and solver = ‘saga’. The second base model is a random forest classifier with the following hyperparameters: number of estimators = 300, maximum depth = 20, class weight = ‘balanced’, and random state = 57. The

third base model is a gradient boosting classifier with the following hyperparameters: number of estimators = 250, learning rate = 0.03, maximum depth = 6, subsample = 0.8, and random state = 57. All base models are calibrated using `CalibratedClassifierCV` with `cv = 3`. Calibration of output probabilities uses Platt scaling. The meta model is a logistic regression model with default hyperparameter settings. The train/test split for the ensemble is determined using random state = 57.

In the **self-supervised classifier**, the `TfidfVectorizer` used has its maximum features set to 3000, ngram range set to (1, 2), and minimum df set to 3. As with the ensemble model, the numeric features used in this classifier are normalized to zero mean and unit variance using `StandardScaler`. The sentence embeddings used have output dimension = 384. These are converted to sparse CSR and stacked horizontally. The base model for the self-supervised classifier is a logistic regression with solver = ‘liblinear’, class weight = ‘balanced’, and maximum iterations = 3000 (2000 in loop iterations). This base model is calibrated using `CalibratedClassifierCV` with `cv = 5` and method = ‘sigmoid’. Output probabilities are also calibrated using Platt scaling. As with the ensemble classifier, data splits for training and evaluation are determined with random state set to 57. In addition, the stratify parameter is set to y . The self-training loop has its number of iterations set to 5, with base per-class confidence thresholds as follows: *Academic* = 0.6, *Cultural* = 0.3, *Other* = 0.9, *Personal* = 0.8, *Professional* = 0.9. The maximum pseudo-labeled samples per class per iteration are as follows: *Academic* = 100, *Cultural* = 200, *Other* = 100, *Personal* = 100, *Professional* = 100. For the *Cultural* class in particular, we implement dynamic thresholding, which we adjust using the 90th percentile probabilities for the class. This can be lowered slightly if insufficient samples are added, using dynamic adjustment = 0.05. We also implement hard negative mining for the *Cultural* class. After self-training, misclassified *Cultural* samples in the validation set are oversampled threefold and added back into the training set. Re-training occurs with these hard examples.

For both of the above models, we manually tuned hyperparameter settings until we found a good set of values that produced reasonable per-topic performance, especially on minority classes in the data.

A.10 Ablation studies on ensemble and self-supervised classifiers.

For the ensemble classifier, sentence embeddings, tf-idf, and lexicon count features contribute slightly positively to model performance and to roughly equal degrees, as demonstrated by the small decreases in accuracy resulting from each of their exclusion from the model pipeline (Table 12). In contrast, the exclusion of acoustic-prosodic and other lexical features from the model improves model performance, suggesting that these features are detrimental to accurate classification decisions.

Excluded feature group	Accuracy	F1 score
–	0.68	0.62
tf-idf	0.67	0.67
Lexicon seed word counts	0.67	0.67
Lexical	0.69	0.69
Acoustic-prosodic	0.71	0.70
Sentence embeddings	0.67	0.67

Table 12: Comparing ensemble classifier accuracy and macro F1 score across subsets of the entire feature set. The first row, where no features are excluded, denotes the performance of the model on the entire feature set, as originally shown in Table 1.

Excluded feature group	Accuracy	F1 score
–	0.72	0.71
tf-idf	0.70	0.69
Lexicon seed word counts	0.70	0.69
Lexical	0.71	0.71
Acoustic-prosodic	0.72	0.71
Sentence embeddings	0.68	0.67

Table 13: Comparing self-supervised classifier accuracy and macro F1 score across subsets of the entire feature set. The first row, where no features are excluded, denotes the performance of the model on the entire feature set, as originally shown in Table 1.

For the self-supervised classifier, sentence embedding features are the single biggest positive contributor to model performance, as demonstrated by the drop in accuracy from its exclusion (Table 13). Tf-idf and lexicon count features also contribute positively and roughly equally. On the other hand, the exclusion of acoustic-prosodic and other lexical features from the model does not affect performance, indicating that these may act as a source of noise instead of a model signal. These patterns are generally consistent with those from ablations over the ensemble classifier’s features.

A.11 Per-class performance of our best-performing (self-supervised) topic classifier.

We report per-class performance metrics and corresponding confusion matrix statistics for our best-performing topic classifier to derive further insight into which classes drive the overall performance of this self-supervised model (Table 14 and Figure 1).

Class	Precision	Recall	F1	Support
Academic	0.78	0.67	0.72	227
Cultural	0.85	0.71	0.77	24
Personal	0.67	0.66	0.67	685
Professional	0.74	0.53	0.62	189
Other	0.73	0.81	0.77	914

Table 14: Self-supervised classifier performance over the held-out test set, stratified by class, i.e. discourse topic. Recall that our best-performing topic classifier achieves an overall accuracy of 0.72, corresponding to a macro F1 score of 0.71 (Table 1). Given the respective class-level support values, we calculate the contribution to model accuracy of each class, in order: 0.07 (*Academic*), 0.01 (*Cultural*), 0.22 *Personal*, 0.05 (*Professional*), and 0.36 (*Other*). Thus, it appears that the overall performance of the model is primarily driven by the *Other* class (corresponding to 36.3% of correct predictions) and the *Personal* class (corresponding to 22.2% of correct predictions), while the other three discourse topic classes contribute relatively little.

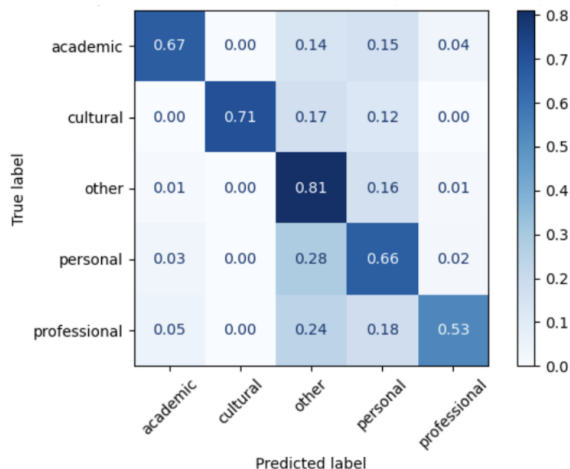


Figure 1: Row-normalized confusion matrix associated with our best-performing topic classifier. This provides additional support for the *Other* topic driving the major part of this self-supervised model’s performance.

A.12 Discourse topic label distribution in subsets of the corpus.

Discourse topic	% of corpus	% of GT	% of AA
Academic	7.8	11.0	7.5
Cultural	0.1	1.2	0.1
Personal	28.1	34.1	27.5
Professional	2.4	9.1	1.7
Other	61.5	44.6	63.2

Table 15: Discourse topic label distribution across classes in the entire SEAME corpus, the ground truth (GT) subset of the corpus, and the automatically-annotated (AA) subset of the corpus. This breakdown of results helps to diagnose a potential source of class imbalance in the fully-annotated corpus and may point to some distributional shift induced by our best-performing classifier. However, in spite of this, note that we replicate all the patterns described in Section 5.2 when we run the same analyses on 1) data from the ground truth set only and 2) data from high confidence subsets ($P \geq 0.7$, $P \geq 0.8$, $P \geq 0.9$) of the automatically-annotated portion of the corpus; the subsequent results remain statistically significant with the same general trends by discourse topic for CSW presence, strategy, direction, quantity, and frequency, though the exact values of odds ratios are slightly different. Similarly, we also replicate the majority of clustering patterns described in Section 5.3 using both 1) the ground truth subset and 2) the high-confidence subsets of the automatically-annotated portion of the corpus; cluster means in each case closely match those corresponding to clusters in Section 5.3. Combined, these replications reinforce the reliability of our current findings.

A.13 Exploratory LDA and BERTopic model hyperparameter settings.

For our LDA analysis on the *Other* topic reported in Section 5.1, we defined a custom list of English and Mandarin filler words (see Appendix A.8) based on prior work (Bhattacharya et al., 2024a). We treated these as stopwords. When building the model vocabulary, we ignored terms that had a document frequency higher than 0.9, and used a cut-off value of 50. We used 5 as the number of topics, which agreed with the automatic number of topics yielded by the BERTopic model with all-MiniLM-L6-v2 embedding for the same analysis. We used the default ‘batch’ learning method for LDA. For the exploratory sub-topic analyses on the *Academic*, *Personal*, and *Professional* topics (see Appendix A.4), we use the same model implementation details.

A.14 Unsupervised clustering initialization.

We select starting points at random to ensure that final k -means clustering results in Section 5.3 are not unduly influenced by initial points. We verify the validity of the conclusions that directly follow this design choice by re-running our current clustering implementation across 30 additional random seeds using different random starting points. We also try an initialization setting that uses one utterance from each topic as starting points for clustering. In each case, the resulting clusters retain the overall trends in cluster compositions that we report in Section 5.3, though the exact proportions in each cluster change slightly. This demonstrates that our qualitative conclusions are robust to initialization and validates our design choice.

A.15 Investigating the interplay of affect and discourse topic.

We conduct exploratory analysis to follow up on our hypothesis of a relationship between affect and discourse topic that goes on to shape the CSW patterns used to express those topics. To begin to verify the extent to which each discourse topic uses affective language, we combine English and Mandarin emotion lexica from [Mohammad and Turney \(2010, 2013\)](#) and calculate utterance-level normalized emotional intensity scores across eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) and two categories of sentiment (positive, negative). We bin utterances into high- or low-emotional intensity, based on the corpus-level median normalized emotional intensity score. We then perform chi-squared tests to compare emotional intensity across topics (Table 16). On average, the *Personal* and *Cultural* topics have greater odds of expressing high emotional intensity than low emotional intensity, relative to the *Academic* and *Professional* topics. Based on these results, we posit that greater affect is linked to discourse topics that require less up-front subject knowledge to discuss, i.e. less complex topics, which in turn allows for their expression using more structured and complex CSW strategies.

Topic	χ^2	<i>p</i> -val.	OR	95% CI
Academic	185.8	**	1.36	[1.30, 1.42]
Cultural	4.0	*	1.39	[1.01, 1.91]
Personal	3950.7	**	2.34	[2.28, 2.41]
Professional	231.8	**	1.80	[1.67, 1.95]
Other	4982.2	**	0.41	[0.40, 0.42]

Table 16: Chi-squared tests and odds ratios comparing topics in high and low emotional intensity utterances. p -values less than 0.05 are denoted by *. p -values less than 0.01 are denoted by **.

Otherwise in Context: Exploring Discourse Functions with Language Models

Guifu Liu¹ and Bonnie Webber¹ and Hannah Rohde²

¹School of Informatics, University of Edinburgh

²School of Philosophy, Psychology and Language Sciences, University of Edinburgh
Guifu.Liu@uni-saarland.de {Bonnie.Webber, Hannah.Rohde}@ed.ac.uk

Abstract

Discourse adverbials are key features of discourse coherence, but their function is often ambiguous. In this work, we investigate how the discourse function of *otherwise* varies in different contexts. We revise the function set in Rohde et al. (2018b) to account for a new meaning we have encountered. In turn, we create the *otherwise* corpus, a dataset of naturally occurring passages annotated for discourse functions, and identify lexical signals that make a function available with a corpus study. We define *continuation acceptability*, a metric based on surprisal to probe language models for what they take the function of *otherwise* to be in a given context. Our experiments show that one can improve function inference by focusing solely on tokens up to and including the head verb of the continuation (i.e., *otherwise* clause) that have the most varied surprisal across function-disambiguating discourse markers. Lastly, we observe that some of these tokens confirm lexical signals we found in our earlier corpus study, which provides some promising evidence to motivate future pragmatic studies in language models.¹

1 Introduction

Discourse coherence helps us understand what a speaker or writer is trying to say in placing one segment of text next to another (Kehler, 2006). In this paper, we focus on a key aspect of discourse coherence: the discourse adverbial *otherwise*, a word whose function in discourse depends on both its lexical semantics and a pragmatic understanding of the context. As seen in Figure 1, *otherwise* can convey 1) CONSEQUENCE: what would happen when a situation doesn't occur, 2) ENUMERATION: what is another option to achieve some goal, and 3) EXCEPTION: what is usually the case given that the clause left of *otherwise*, or left hand side [LHS] conveys an exception.

¹Code and data are available in <https://github.com/GuifuLiu/otherwise>

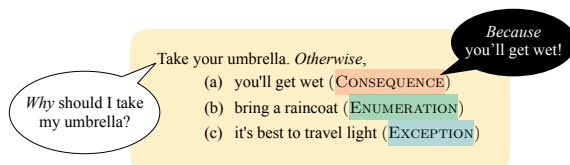


Figure 1: An example of *otherwise* functions.

Being able to distinguish these discourse functions is important for downstream applications in natural language understanding. For example, when asked “Why should I take an umbrella?”, a question-answering system should apply clause (a) and respond with “to avoid getting wet”, rather than clause (b) “to avoid bringing a raincoat”.

Although Rohde et al. (2018b) have shown that human participants can distinguish these discourse functions in a provided context, we still don't understand the signals that make such a function available, and the extent to which language models can infer these discourse functions. In addition, previous work has been limited to small-scale, researcher-constructed examples in the context of psycholinguistic studies (Rohde et al., 2016, 2018a).

To shine light on these questions, we introduce the *otherwise* corpus, a dataset of 294 naturally-occurring passages annotated for discourse functions, and a revised *otherwise* function set, to account for a new meaning that is not discussed in Rohde et al. (2018b). Through corpus study, we find that these respective functions are associated with the presence of distinctive lexical cues such as negation markers, modal triggers, and conjunctions.

To study how language models infer the function of *otherwise* in context, we define *continuation acceptability*: we replace *otherwise* with a set of candidate discourse markers that are distinctive of a function (e.g., *alternatively* for ENUMERATION). We expect that the model will accept the one that

best fits the context by assigning low *surprisal* (a word’s negative log probability in context) to the *continuation* (the text segment after a candidate marker, or right hand side [RHS]). We validate this metric by showing that it can infer the annotators’ assigned function better than a majority baseline, though its ability to do so varies across discourse functions of the passage and models.

We then explore alternative aggregation methods beyond the per-token average used in *continuation acceptability* to identify key tokens that help signal the function of *otherwise*. We find that solely focusing on tokens up to and including the head verb of the continuation that have the most varied surprisal across discourse markers shows convincing performance improvement, despite ignoring other tokens. In addition, some of these tokens confirm lexical signals identified in the corpus study, suggesting that when the model infers the *otherwise* function, these signals are indeed relevant.

Our contributions are (i) the *otherwise* corpus, a dataset of naturally-occurring *otherwise* passages annotated for discourse functions (§3.2), (ii) *continuation acceptability*, a new metric based on language models to probe for their most accepted discourse function (§3.4), (iii) insights into how lexical signals help make a discourse function available (§3.3), and (iv) results showing how language models are affected by certain aspects of the context (§4.2, §4.3).

2 Related Work

Theories of discourse coherence shape our research questions and inform our experimental design. In what follows, we begin by outlining prior work on interpreting *otherwise* in context (§2.1). We then discuss the application of language models in discourse research (§2.2).

2.1 *Otherwise* in Context

Knott (1996) studied the semantics of *otherwise* in relation to other discourse markers with a substitution test to discover when a writer is willing to substitute *otherwise* for another discourse marker. *Otherwise* was found to be synonymous with *if not*, a hyponym of *or* and *or else*, and contingently substitutable with *alternatively*. The finding suggests that *otherwise* exhibits granularity in its semantic meaning.

Webber et al. (1999) noted that *otherwise* is compatible with additional discourse relations, such as

an unmarked *because* in “If the light is red, stop. Otherwise, you might get run over”. Likewise, Rohde et al. (2016) have shown that, in the presence of *otherwise*, people infer additional discourse relations that hold jointly with those associated with the adverbial, by inserting connectives *because*, *or*, *but* before *otherwise*. For instance, in Figure 1, one may insert *because* to indicate inference of ARGUMENTATION for (a), *or* to indicate inference of ENUMERATION for (b), and *but* to indicate inference of EXCEPTION for (c).

Rohde et al. (2018b) subsequently provide empirical evidence for why conjunctions inserted before *otherwise* split among these three. They have found that variability in the choice of conjunctions arises from the lexical semantics of *otherwise*, combined with inferences of its discourse function (to be discussed in §3.1).

Our work builds on previous findings by scrutinizing the lexical signals that make a function available, using large-scale, naturally occurring examples that represent how a speaker or writer uses the discourse adverbial, and examining how language models infer the functions of *otherwise*.

2.2 Discourse and Language Models

The use of language models in discourse is an active research area. Recent work on discourse markers and language models has taken two main approaches: (i) using **cloze tasks** with masked language models to predict connectives (Kurfali and Östling, 2021; Pandia et al., 2021; Stodden et al., 2023; Dong et al., 2024), and (ii) using **prompting** to insert discourse connectives for implicit discourse relation annotation (Yung et al., 2024) and to uncover the function of discourse particle *actually* (Sadlier-Brown et al., 2024) and *just* (Sheffield et al.).

While standard masked language models may be limited in predicting multi-token discourse markers without additional training (Kalinsky et al., 2023), prompting also has several shortcomings. In particular, small variations in the prompt are shown to affect model outputs (Salinas and Morstatter, 2024; Mizrahi et al., 2024). To avoid the drawbacks of prompting, we use surprisal scores of language models to infer the discourse function of *otherwise*. There is also an increasing interest in the use of surprisal to account for a wide range of linguistic phenomena, such as sentence processing (Wilcox et al., 2018), utterance predictability (Giulianelli et al.,

2023), and discourse structure (Tsipidi et al., 2024). In our study, we apply surprisal to investigate the discourse function of an ambiguous adverbial.

Surprisal has also been used to test the effect of discourse connectives on discourse coherence. Zhou et al. (2010) constructed synthetic passages by inserting a candidate implicit connective between a pair of arguments. A language model is then used to calculate the perplexity of every token in the constructed passage. The connective from the passage with the lowest mean surprisal is chosen as the best implicit connective for the argument pair.

Cong et al. (2023) used controlled psycholinguistic stimuli and calculated the surprisal of a critical word to test the effect of discourse connectives *even so* and *however* on reversing the expectations about an event. Similarly, we measure how the expectation for the continuation is influenced by candidate discourse markers that disambiguate *otherwise* functions, which may be coherent or not depending on the context. The main differences are that the discourse functions, discourse markers, and the context we investigate are more diverse and complex than those used in psycholinguistic stimuli, which require the model to understand a wider context.

3 Methodology

3.1 Revised Function Set of *Otherwise*

Rohde et al. (2018b) define three functions of *otherwise* based on both the lexical semantics of *otherwise* and the relation that humans infer between two segments in the passage. They are shown in Figure 1. One function is ARGUMENTATION, where the clause to the right of *otherwise*, [RHS] shows what the result will be if certain advice in [LHS] is not followed, as in (a). Another function is ENUMERATION. When the speaker provides two equally viable options to fulfill a shared goal, [RHS] introduces an alternative option, as in (b). A third function is EXCEPTION, where [RHS] expresses what is usually the case, while [LHS] specifies an exception to it, as in (c).

However, we have encountered an additional meaning of *otherwise* that does not fit into this function set:

(d) I like you too. Otherwise, we wouldn't be friends.

(e) Of course I mean it. Otherwise, I wouldn't ask.

For these passages, [RHS] doesn't provide a reason for the claim in [LHS], an equally viable option, or a description of what generally holds. Instead, the *otherwise* clause describes a logical conclusion if the situation in the [LHS] **doesn't** arise. We name this new function CONSEQUENCE.

All ARGUMENTATION passages fulfill the definition of CONSEQUENCE, as their *otherwise* clauses describe an undesirable or negative outcome that can be avoided if the advice in the main clause is followed. However, the opposite is not necessarily true. Therefore, we define ARGUMENTATION as a subordinate function of CONSEQUENCE. However, when we mention CONSEQUENCE as a passage label in the following sections, we refer to passages that are CONSEQUENCE but not ARGUMENTATION.

We provide definitions and examples of the revised function set in Table 1.

3.2 The *otherwise* corpus

The *otherwise* corpus consists of 294 passages with sentence-initial *otherwise* that are annotated for our revised discourse functions in §3.1. These passages are randomly sampled from the *Corpus of Contemporary American English* or COCA (Davies, 2008), and the *British National Corpus* or BNC (BNC Consortium, 2007) and span a wide array of sentence constructions (e.g., declarative, imperative, question), genres and modalities (e.g., blogs, academic, fiction, TV, movies). All passages are contextually contained so that the context provided in the passage is sufficient to infer the discourse function.

To identify the discourse function that is operative in a passage, we use a paraphrase task: for each passage, every function of *otherwise* is assigned a paraphrase to convey the lexical semantics of that function (Paraphrase in Table 1), and participants infer a valid paraphrase.

The final dataset contains 294 human-annotated *otherwise* passages and their discourse functions (Table 2). Each passage was annotated by a researcher. In addition, one-fifth of the dataset was also labeled by four participants who are native or near-native adult English speakers. The average inter-annotator agreement between researcher and participant is $\kappa = 0.87$. Details on dataset construction and annotation are in Appendix A.

3.3 Function Signals

Our *otherwise* corpus contains naturally-occurring passages that are useful for corpus study. Particu-

Function	Definition	Paraphrase	Example
CONSEQUENCE	If the situation in [LHS] doesn't occur, the situation in [RHS] would arise.	[LHS] because if not [LHS], [RHS]	[I like you too.] Otherwise, [we wouldn't be friends.]
↳ ARGUMENTATION	[RHS] is <i>undesirable</i> and can be possibly avoided by following [LHS]	To avoid [RHS], [LHS].	[We have to operate immediately.] Otherwise, [she will die.]
ENUMERATION	It doesn't take the failure of [LHS] to consider [RHS] as another option.	There is more than one option for [goal]. They are 1) [LHS] and 2) [RHS].	[I like a nice curry.] Otherwise, [I'll nibble on fruit.]
EXCEPTION	[LHS] is an exception to [RHS]	Generally [RHS], an exception is that [LHS].	[Some people are riding horses.] Otherwise, [people are traveling on foot.]

Table 1: Revised *otherwise* function set, its description, the paraphrase used to identify a function, and examples from the *otherwise* corpus. [LHS] and [RHS] correspond to the clause that is left and right of *otherwise*.

CONSQ.	ARG.	ENUM.	EXCPT.
.19	.45	.13	.26

Table 2: Function Distribution of the *Otherwise* corpus.

larly, we are interested in finding the signals that make a function available. We calculate pointwise mutual information (Torabi Asr and Demberg, 2013) for each word token w and discourse function l ,

$$pmi(w, l) = \log \frac{p(w, l)}{p(w)p(l)}$$

A high PMI score indicates that word token w is highly associated with discourse function l , making the token a strong candidate for a lexical signal for that function. We only consider word tokens that occur in more than 15 passages to avoid overfitting the contents of the corpus (Zeldes and Liu, 2020).

Our results show that **modals** make up the largest group of signals. The functions they co-occur with depend on the modal type and its position: **Priority modals** (e.g., *need*, *must*, *should*) indicate how important and desirable an event is by the speaker (Pyatkin et al., 2021) and often occur in [LHS] to signal ARGUMENTATION:

- (1) Consumers should be told the whole truth. Otherwise, it amounts to fraud (CONSEQUENCE).

Plausibility modals (e.g., *could*, *can*, *may*), on the other hand, indicate how likely an event will happen given assumptions in the context (Pyatkin et al., 2021). Their appearance in [LHS] often indicates viability of an option and signals ENUMERATION, while *might*, *would*, *may* that appear in

[RHS] often indicate the likelihood of an outcome and signal CONSEQUENCE or EXCEPTION:

- (2) The public can visit an exhibition to share their feedback. Otherwise, the public can submit feedback forms on the website. (ENUMERATION)

- (3) It's a good thing to ride horses at home and not at the racecourse. Otherwise, you might have been much more badly hurt. (CONSEQUENCE)

Other function signals include **negation markers** and **downward-entailing predicates**² in either [LHS] or [RHS] that indicate CONSEQUENCE and ARGUMENTATION:

- (4) She was nervous. Otherwise, she wouldn't be rambling. (CONSEQUENCE)

- (5) You keep your mouth shut, you never contact Nasry again, you don't lawyer up, this affidavit stays in a vault, and the video disappears. Otherwise, the charge will be murder. (ARGUMENTATION)

- (6) Generally, eating problems can be avoided by being flexible with your puppy from the start, varying the eating location, alternating types of dog food, and changing feeding times. Otherwise, be prepared for your dog to become accidentally conditioned by circumstances that lend new significance to the sound of the dinner bell. (ARGUMENTATION)

Focus particle *only* in either [LHS] or [RHS] that indicates ARGUMENTATION and EXCEPTION:

- (7) He spent only two years at school. Otherwise, he was educated at home. (EXCEPTION)

²Downward entailing constructions support valid reasoning from a set to a member. For example, John doesn't own a dog to John doesn't own a beagle (Webber, 2013).

Connectives that appear at in [LHS] and not attached to *otherwise*³: *or* for ENUMERATION and *but* for CONSEQUENCE.

(8) Treatment may also be available in a Young Chronic Sick Unit or in a Geriatric Unit in a hospital. Otherwise, the patient might spend some time in a private nursing home. (ENUMERATION)

(9) Clearly, he resented Gavin but also had empathy towards him. Otherwise, he wouldn't have lent him money in the first place. (CONSEQUENCE)

We also found that several of these signals appear in function-bearing passages beyond those with sentence-initial *otherwise*. We gathered a substantially larger sample of passages ($n = 2656$) that contain *because otherwise*, *alternatively*, and phrases with *exception*⁴, each marking distinct discourse functions—namely, CONSEQUENCE, ENUMERATION, and EXCEPTION. Across these passages, modal triggers, negation markers, and the connective *or* **remain** function signals.

While our data-driven method extracts words that co-occur with some *otherwise* functions, the method falls short in identifying discourse signals in the context that surrounds them, and establishing whether a comprehender might actually use them when inferring a function. To address this, we analyze the linguistic characteristics of tokens that a language model identifies as distinctive of a function (§4.3). As we will show, the model is sensitive to the context of an *otherwise* passage. In inferring the appropriate function, the model confirms the utility of several lexical signals we have identified in this section.

3.4 Metric: Continuation Acceptability

To study the capability of language models to distinguish *otherwise* functions, we propose a variant of surprisal-based metric. *Continuation acceptability* selects a discourse marker that indicates a distinct function and makes a continuation, [RHS], more likely given prior context, [LHS].

Definition Let D be a set of candidate discourse markers that are distinctive of a discourse function, (a_1, a_2) , the [LHS] and [RHS] clause (or continuation) of a passage s in our *otherwise* corpus (with

³In our corpus, we only consider sentence-initial bare *otherwise* without additional connective attached to it (e.g. *or otherwise*).

⁴Phrases to mark EXCEPTION are: *with the exception that*, *except for the fact that*, *an exception is that*, *one exception is that*, *as an exception*.

sentence-initial *otherwise* removed). We construct $\{(a_1 + d + a_2) | d \in D\}$, the set of variations of s , where $(a_1 + d + a_2)$ denotes string concatenation. d is the most acceptable discourse marker if its variation $(a_1 + d + a_2)$ has the lowest surprisal of continuation in a language model θ :

$$\arg \min_{d \in D} I[(a_1 + d + a_2); \theta]$$

where I is the surprisal of continuation:

$$I(s; \theta) = -\frac{1}{|A|} \sum_{i \in A} \log p_{\theta}(t_i | t_{<i})$$

with A the set of indices of tokens in continuation, t_i the tokens in passage s .

For example, consider the passage s , *Take your umbrella. Otherwise, you'll get wet*. Suppose we have candidate discourse markers *because otherwise* for CONSEQUENCE and *alternatively* for ENUMERATION. Then the respective variations for s are:

- (1) Take your umbrella. *Because otherwise*, you'll get wet.
- (2) Take your umbrella. *Alternatively*, you'll get wet.

The continuation acceptability for the first variant is calculated as the surprisal of [RHS], *you'll get wet*, conditioned on [LHS], *Take your umbrella*, and the candidate marker, *Because otherwise*. We expect the model to assign lower surprisal scores for the continuation when it is conditioned on the candidate marker of the correct function. Therefore, the model should assign lower surprisal to variation (1) than (2).

Notice that in all variations constructed, both the prior context and the continuation are kept the same. The only change is the candidate marker, which allows us to test its facilitating effect on the expectation for the continuation. Per-token surprisal also allows us to examine how language models respond to specific aspects of the continuation (§4.2, §4.3), which are more difficult to capture through mask-filling or prompting.

We use the average of per-token surprisal as indicated by the formula, but we also consider other aggregate functions for per-token surprisal in §4.2.

4 Experiments

In Rohde et al. (2018b), human participants infer the discourse functions of an adverbial, which vary across passages. We raise the question of whether a

Function	Discourse Markers			
CONSEQUENCE	because if not	because [PRON] \neg [AUX]	because otherwise	
↳ ARGUMENTATION	unless this is done	(when/ by) failing to do so	for fear that	lest
ENUMERATION	alternatively	as an alternative	in addition	on the other hand
EXCEPTION	but mostly	but usually	but other than that	
CONTROL	otherwise			

Table 3: Candidate discourse markers and their corresponding function. [PRON] and \neg [AUX] correspond to the pronoun and negated auxiliary verb of [LHS].

language model can also discern varied interpretations of *otherwise* and accept annotators’ assigned function. The experiments below use *continuation acceptability* (§3.4) to evaluate the capability of language models to do so. First, we validate that *continuation acceptability* captures the models’ understanding of *otherwise* (§4.1). Then, exploiting per-token surprisal score, we explore what aspects of the continuation are important to identify the discourse function: we use alternative aggregate functions for per-token surprisal (§4.2) and linguistic annotations on tokens found to be distinctive of a function (§4.3).

4.1 Can continuation acceptability identify *otherwise* function?

Experimental setup. We used autoregressive language models of increasing size without further fine-tuning: GPT-2 Base (Radford et al., 2019), with 124 million parameters, and GPT-Neo (Black et al., 2021), with 1.3 billion parameters, and Mistral-7B-v0.1 (Jiang et al., 2023). We selected the GPT family because they are the standard models for testing psycholinguistic predictive power, allowing for comparability with prior work. We additionally included a newer open-weight model, Mistral-7B-v0.1, to access per-token surprisal.

We applied *continuation acceptability* (§3.4) to these models and our *otherwise* corpus.⁵ Specifically, for a passage in the corpus and all candidate discourse markers, we calculate the *continuation acceptability* score of words in the continuation.

As shown in Table 3, we choose candidate discourse markers that are both relatively frequent and generally successful at capturing a unique *otherwise* function in naturally-occurring examples we sampled from COCA and BNC, which are also used for constructing the *otherwise* corpus. We select three or more candidate markers for each *otherwise* function to reduce the bias of syntactic con-

⁵We use the discourse function assigned by the researcher as reference label.

	Researcher Label			
	Consq	Arg	Enum	Excpt
GPT2	.07	.11	.05	.04
GPT-Neo-1.3B	.16	.08	.05	.13
Mistral-7B-v0.1	.11	.11	.08	.21

Table 4: The proportion of passages predicting CONTROL (accept *otherwise* as Top 1 candidate discourse marker), corresponding to the researcher label in bold.

straints and specificity of one discourse marker (Pyatkin et al., 2023). We optionally allow *otherwise* to be a candidate and label its function CONTROL. When *otherwise* is chosen, the model doesn’t prefer any marker that explicitly realizes a single function that we defined, but prefers the adverbial as is, in its original form.

For each candidate discourse marker [DM], we allow both inter-sentential *Arg1*. [DM] *Arg2* and intra-sentential concatenation *Arg1*(,) [DM] *Arg2* of a marker and optionally includes comma after [DM] if suitable. We choose the concatenation that is most accepted by the model for that marker.

Results and Discussion. We consider the function to be correctly identified if a candidate marker of that function appears in the top $k = \{1, 3, 5\}$ accepted discourse markers. For example, a best-performing model will accept *but mostly*, *but usually*, and *but other than that* as top 3 markers for an EXCEPTION passage.

We first show that using surprisal to identify the *otherwise* function is not trivial. All models accept candidate discourse markers that explicitly realize a function more often than bare *otherwise* (i.e., without an additional conjunction before), which is the original discourse marker that appears in the passage (Table 4). The result suggests that these models do not simply memorize the sentence-initial *otherwise*, and that each model favors a distinct function to be CONTROL except for ENUMERATION. All models accept bare *otherwise* less fre-

$k =$		1	3	5
Majority		.45		
Mask Scoring	T5-Base	.45	.76	.93
Continuation Acceptability	GPT2	.51	.74	.85
	GPT-Neo-1.3B	.56	.78	.88
	Mistral-7B-v0.1	.59	.80	.89

Table 5: Overall passage accuracy using top $k = \{1, 3, 5\}$ predictions of discourse markers. **Majority** corresponds to assigning the majority function of the dataset to all passages.

$k =$	GPT2			GPT-Neo-1.3B			Mistral-7B-v0.1		
	1	3	5	1	3	5	1	3	5
Conseq	.36	.52	.61	.39	.64	.75	.41	.68	.84
Arg	.45	.75	.89	.44	.71	.94	.41	.73	.93
Enum	.41	.68	.81	.43	.78	.86	.59	.73	.84
Excpt	.40	.65	.77	.43	.64	.77	.48	.72	.79

Table 6: Per-function passage accuracy using top $k = \{1, 3, 5\}$ predictions of discourse markers.

quently for ENUMERATION than other functions. In other words, they accept candidate markers that explicitly indicate a function much more frequently (e.g., *alternatively* or *in addition*) than *otherwise*.

As there currently exists no system for identifying the discourse functions of *otherwise*, we use the majority function of the corpus as a baseline. In addition, we compare *continuation acceptability* against a mask-scoring baseline⁶, defined as the model probability of inserting the connective d at the mask token between the arguments (a_1, a_2) (Kurfali and Östling, 2021; Stodden et al., 2023):

$$P(d|a_1, a_2) \propto P_\theta([\text{MASK}] = d|a_1[\text{MASK}]a_2)$$

We report accuracy where the function of any of the top- k predicted discourse markers matches the gold function. Table 5 and Table 6 show overall and per-function accuracy. Models using *continuation acceptability* outperform the majority baseline, and also surpass mask-scoring at top-1 prediction. Upon inspection, mask-scoring under-predicts EXCEPTION, achieving only 18.6% accuracy on EXCEPTION passages with top-1 predictions.⁷ Moreover, in roughly 78% of cases, the model’s top-1 prediction is bare *otherwise*, suggesting that T5-Base may have encountered and memorized these

⁶We keep the same experimental setup but use T5-Base (Raffel et al., 2020), which is trained on a masked language modeling objective. We select this model because it supports multi-word predictions at the mask token, making it compatible with our candidate connectives. The top k candidate connectives with the highest probabilities are selected.

⁷When calculating per-function accuracy, we disregard predictions that are bare *otherwise*.

examples during training. However, mask-scoring achieves higher accuracy in identifying the gold function with top-5 predictions.

We also observe that a larger model doesn’t guarantee better accuracy in prediction. Despite Mistral-7B-v0.1 having 7 times more parameters than GPT-Neo, its overall improvement is marginal. Additionally, Mistral seems to be biased toward interpreting *otherwise* as EXCEPTION at $k = 1$: we observe that Mistral infers the most EXCEPTION passages than other models (Table 7). It also most frequently infers bare *otherwise* in EXCEPTION passages across functions (Table 4).

Across all models, the most prevalent function of the Top 1 scoring marker corresponds to the researcher label (Table 7). Nevertheless, CONSEQUENCE passages are the most challenging of all functions, as all models predict CONSEQUENCE correctly less frequently at $k = 1$ compared to other functions. It is often confused with ARGUMENTATION, which is expected as ARGUMENTATION is a subordinate function of CONSEQUENCE, and they are semantically similar.

One concern is that a model may inherently prefer specific candidate markers, regardless of the passage, which would complicate our analysis of model competence in inferring a discourse function. We demonstrate that this is generally not the case in Appendix B.

Our results have shown that *continuation acceptability* can be used to identify *otherwise* function, though the success varies across discourse functions of the passage and models.

4.2 Are all tokens in continuation equally important to identify *otherwise* function?

Fang et al. (2025) have shown that for long context understanding, not all tokens are equally important to identify the answer token. Similarly, our corpus study (§3.3) finds that many lexical signals that help make functions available, such as modals and negation markers, often appear before the main verb. We hypothesize that not all tokens in a continuation are equally important for identifying the *otherwise* function and that mean surprisal (i.e., token-level surprisal aggregate used in §3.4) may not be representative enough in predicting discourse function.

Experimental setup. In this experiment, we explore alternative aggregates for per-token surprisal and compare them with mean surprisal. We test

		Researcher Label															
		Consq				Arg				Enum				Excpt			
Top 1 Label		Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt	Consq	Arg	Enum	Excpt
GPT2		.36	.32	.14	.18	.21	.45	.22	.12	.08	.35	.41	.16	.20	.23	.17	.40
GPT-Neo-1.3B		.39	.21	.27	.12	.29	.44	.17	.10	.16	.24	.43	.16	.12	.19	.27	.43
Mistral-7B-v0.1		.41	.30	.20	.09	.30	.41	.23	.07	.03	.27	.59	.11	.04	.27	.21	.48

Table 7: The distribution of Top 1 predicted label. A green cell indicates that the predicted function is acceptable, while a red cell indicates that it is unacceptable. We allow CONSEQUENCE candidate markers for ARGUMENTATION passages.

both previously proposed aggregates in the literature (*superlinear*, *maximum*, *variance*, and *difference*; see Appendix C for full definitions) and new aggregates designed for our task. Specifically, we select key tokens using two criteria and average their per-token surprisal: 1) **Pre-root**: consider tokens up to the root (as defined in syntactic dependency) or head verb of the continuation and 2) **Most varied tokens (MVT)**: tokens with the largest variance in surprisal across variations of different discourse markers (that make a function available). We believe MVT allows us to pinpoint the exact location where the model diverges on its interpretation of *otherwise* function, given that the candidate marker is the only element that varies in our experiment. We consider the Top 3 most varied tokens. In addition to testing them separately, we also combine these two criteria.

We use GPT-Neo-1.3B for subsequent experiments, as it has comparable performance to Mistral-7B in our previous experiment and is often used in the psycholinguistic literature, which can shed some light on token-level understanding of the continuation.

Results and Discussion. Giulianelli et al. (2023) has shown that superlinear surprisal aggregate highly correlates with human acceptability judgments on an upcoming turn in dialogue from *Switchboard* (Godfrey et al., 1992) and *DailyDialog* (Li et al., 2017). Besides just dialogues, our results, which are tested on a wide range of genres (§3.2), confirm that superlinear is the best performing aggregate among what has been proposed in past literature, particularly in the context of continuation acceptability contingent on a function-indicating discourse marker.

Additionally, our proposed aggregates, **Pre-ROOT** or **Top 3 MVT**, obtain comparable or better performance when compared to mean surprisal, despite considering fewer tokens (around 3 tokens on

Suprisal Metric	Acc.			Avg # Token	
	k = 1	3	5		
Mean	.56	.78	.88	13.45	
Superlinear ($n = 0.5$)	.63	.83	.91		
Maximum	.51	.73	.85		
Variance	.52	.72	.85		
Difference	.54	.75	.87		
Ours					
Pre-ROOT	Top 3 MVT				
✓	✓	.61	.81	.91	3.54
	✓	.54	.75	.88	3
✓	✓	.64	.84	.93	2.73

Table 8: Passage Top k accuracy where $k = \{1, 3, 5\}$ predictions of discourse markers with GPT-Neo-1.3B, using various per-token surprisal aggregates

average as opposed to ≈ 13 tokens). Particularly, when combining **Pre-Root** and **Top 3 MVT** criteria, the performance exceeds that of superlinear (Table 8). We also see that **Top 3 MVT** criteria itself doesn’t filter tokens in a way to better identify *otherwise* function compared to mean surprisal. Upon examining the relationship between two criteria, we have found that on average 48% of the **Top 3 MVT** appear pre-root, and more so for CONSEQUENCE and ARGUMENTATION (51% and 49%): two functions that are associated with most types of lexical signals. Thus, we hypothesize that combining criteria **Pre-Root + MVT** provides linguistic cues to identify lexical signals that are predictive of a function. In what follows, we test this hypothesis by investigating linguistic information of tokens that fulfill these two criteria and compare their characteristics with those of function signals found in our earlier corpus study (§3.3).

4.3 What lexical signals are predictive of *otherwise* function?

We would like to assume that the key tokens selected by our criteria are in fact relevant to the model’s decision-making in predicting the *otherwise* function. In this experiment, we analyze the linguistic characteristics of tokens that the model identifies as distinctive of a function. We observe that some of these tokens confirm lexical signals we previously identified in the corpus study (§3.3), and they provide promising evidence on how model behavior, such as surprisal, can be useful for studying discourse signals.

Experimental setup. For each passage, we extract the following linguistic annotation for each token that is both **Pre-Root** and **Top MVT** (§4.2)⁸: 1) word type, 2) part of speech, and 3) dependency tag. For each type of linguistic annotation i , we calculate PMI score $pmi(i, l)$ as in §3.3, but extend i from word type to other linguistic information. For example, given the token *looking*, we calculate a score for its word type *look*, part of speech tag *gerund or present participle*, and dependency tag *root*.

A high PMI score indicates that the linguistic information i is highly associated with discourse function l as seen by the model.

Results and Discussion. We find that both word types of **modal tokens** and part-of-speech tag *modal* are high-scoring signals. PLAUSIBILITY modals (as defined in §3.3) *may*, *might*, *will* as a word type signal both CONSEQUENCE and ARGUMENTATION, while *can* and *could* signal ARGUMENTATION and ENUMERATION respectively. PRIORITY modals *need* and *must* signal ENUMERATION and EXCEPTION respectively. Interestingly, *modal* as a part of speech tag is only high-scoring for CONSEQUENCE and ARGUMENTATION.

We have found some other lexical signals that confirm those from the corpus study: **Negation** as a dependency tag signals consequence, while *no* and *nothing* as a word type signal EXCEPTION, and *not* signals CONSEQUENCE. **Focus particle** *only* as a word type signals EXCEPTION.

We also found lexical signals that were not previously discovered in the corpus study. For instance, the word type *become* is found to signal ARGUMENTATION, and there are eight of such instances

⁸with using spaCy en_core_web_sm pipeline, see <https://spacy.io/usage/processing-pipelines>

where *become* occurs in [RHS] to express a new state when the situation in [LHS] doesn’t arise:

(1) It was essential that people try to connect. Otherwise, we would become a society of strangers.
(ARGUMENTATION)

Because the language model we have chosen is auto-regressive (i.e., generates a continuation that is conditioned on previous context), we are unable to apply the same analysis on tokens in [LHS]. Nevertheless, it is reassuring to see that some key tokens extracted by the model confirm lexical signals we found in the corpus study, especially given the model likely has been exposed to far more *otherwise* passages than our corpus. More importantly, we show that the model has learned frequency-correlated cues during pre-training and assigns more varied surprisal on these tokens across candidate discourse markers that license a function. These findings provide some promising evidence that token-level surprisal may offer helpful information for future pragmatic studies. As a next step, stronger evidence for function signals could be obtained by directly manipulating them in the passage (e.g., ablation) while preserving the passage’s meaning.

5 Conclusion

In this paper, we study the discourse functions of *otherwise* through language models. To do so, we introduce a new dataset (the *otherwise* corpus) and metric (*continuation acceptability*). With these tools, we show that language models exhibit some capability of inferring *otherwise* function, though their extent to do so varies across functions of the passage and the model. Additionally, we identify the types of lexical signals that influence the availability of specific discourse functions and reveal that the model attends to some of these signals when inferring the discourse function. We hope our findings open new doors for study on adverbial and discourse coherence in both psycholinguistic and computational research, and inspire developing pragmatically competent models.

Limitations

We acknowledge that our study has some limitations. First, our dataset only considers sentence-initial *otherwise*. This helps us ensure the adverbial serves a discourse function and simplifies our data collection process. We recognize that this may not represent all use cases of the adverbial. It may also

affect syntactic patterns and lexical signals of passages we have analyzed. For future research, we plan to collect passages where *otherwise* within a sentence serves a discourse function.

Second, our analysis was based on the assumption that surprisal scores from language models reflect human behavioral patterns such as reading time. Recent work has shown that as the model size of language models increases, when using surprisal, their psycholinguistic predictive power decreases. This may be because these models are exposed to much more data than humans are. We have chosen models that are highly correlated with human reading time in past studies (Cong et al., 2023) or are of moderate size. Nevertheless, more direct evidence for discourse coherence and surprisal could be obtained by collecting reading time data (with an emphasis on function signals and preverbal tokens) or calibrating large-size models with temperature-scaling (Liu et al., 2024), so that they are more predictive of human behavioral patterns.

Lastly, although there is clearly value in studying discourse functions of adverbials in the interest of discourse parsing and other natural language understanding systems, we have not pursued other potential roles of discourse function inferences. An extended study may examine the influence of adverbials and their discourse functions on other semantic and pragmatic phenomena such as conditionals, anaphora resolutions, and presupposition, all of which we believe to be relevant to *otherwise*.

References

- Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow](#).
- BNC Consortium. 2007. [The British National Corpus, XML Edition](#). Oxford Text Archive.
- Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Philippe Blache. 2023. [Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives, Even So They Are Surprised](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 222–232, Singapore. Association for Computational Linguistics.
- Mark Davies. 2008. [The Corpus of Contemporary American English \(COCA\)](#).
- Yunfang Dong, Xixian Liao, and Bonnie Webber. 2024. [Syntactic Preposing and Discourse Relations](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2790–2802, St. Julian’s, Malta. Association for Computational Linguistics.
- Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2025. [What is Wrong with Perplexity for Long-context Language Modeling?](#) ArXiv:2410.23771 [cs].
- Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. [Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.
- J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. [Switchboard: telephone speech corpus for research and development](#). In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. [Simple and Effective Multi-Token Completion from Masked Language Models](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.
- Andrew Kehler. 2006. Discourse coherence. *The handbook of pragmatics*, pages 241–265.
- Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.
- Murathan Kurfalı and Robert Östling. 2021. [Let’s be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction](#). In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

- Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. [Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the “right reasons”?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9598–9619, Bangkok, Thailand. Association for Computational Linguistics.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of What Art? A Call for Multi-Prompt LLM Evaluation](#). *Transactions of the Association for Computational Linguistics*, 12:933–949. Place: Cambridge, MA Publisher: MIT Press.
- Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. [Pragmatic competence of pre-trained language models through the lens of discourse connectives](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.
- Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. [The Possible, the Plausible, and the Desirable: Event-Based Modality Detection for Language Processing](#). 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 953–965, Online. Association for Computational Linguistics.
- Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. [Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design](#). ArXiv:2304.00815 [cs].
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. [Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task](#). In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.
- Hannah Rohde, Anna Dickinson, Nathan Schneider, Annie Louis, and Bonnie Webber. 2018a. Exploring substitutability through discourse adverbials and multiple judgments. In *IWCS 2017-12th International Conference on Computational Semantics*. ACL Anthology.
- Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018b. [Discourse Coherence: Concurrent Explicit and Implicit Relations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.
- Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. [How Useful is Context, Actually? Comparing LLMs and Humans on Discourse Marker Prediction](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241, Bangkok, Thailand. Association for Computational Linguistics.
- Abel Salinas and Fred Morstatter. 2024. [The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651, Bangkok, Thailand. Association for Computational Linguistics.
- William Berkeley Sheffield, Kanishka Misra, Valentina Pyatkin, Ashwini Deo, Kyle Mahowald, and Junyi Jessy Li. [Is it JUST semantics? a case study of discourse particle understanding in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21704–21715. Association for Computational Linguistics.
- Regina Stodden, Laura Kallmeyer, Lea Kawaletz, and Heidrun Dorgeloh. 2023. Using masked language model probabilities of connectives for stance detection in english discourse. In *Proceedings of the 10th workshop on argument mining*, pages 11–18.
- Fatemeh Torabi Asr and Vera Demberg. 2013. [On the Information Conveyed by Discourse Markers](#). In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria. Association for Computational Linguistics.
- Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. [Surprise! Uniform Information Density Isn’t the Whole Story: Predicting Surprisal Contours in Long-form Discourse](#). ArXiv:2410.16062 [cs].
- Bonnie Webber. 2013. [What excludes an Alternative in Coherence Relations?](#) In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 276–287, Potsdam, Germany. Association for Computational Linguistics.
- Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. [Discourse Relations: A Structural and Presuppositional Account Using Lexicalised TAG](#). In *Proceedings of the 37th Annual Meeting of*

the Association for Computational Linguistics, pages 41–48, College Park, Maryland, USA. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. [What do RNN Language Models Learn about Filler–Gap Dependencies?](#) In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. [Prompting implicit discourse relation annotation](#). In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Amir Zeldes and Yang Liu. 2020. [A Neural Approach to Discourse Relation Signal Detection](#). *Dialogue & Discourse*, 11(2):1–33.

Zhi Min Zhou, Man Lan, Zheng-Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the SIGDIAL 2010 Conference*, pages 139–146.

A Dataset Construction

Candidate passages When selecting the passages that contain a discourse adverbial, its occurrence is not sufficient. Sometimes, the adverbial does not function as a discourse marker, and instead modifies part of a syntactic matrix clause. Thus, we select the sentence that starts with the adverbial immediately after the previous sentence (i.e., sentence-initial *otherwise*). We find this strategy works quite well due to its syntactic convention. Our search patterns for COCA and BNC are `._otherwise_`, where `_` indicates a blank space⁹. In total, we have extracted 294 passages for function annotation from 7,770 passages in COCA and 1,014 passages in BNC.

Dataset annotation All candidate passages are annotated by the researcher. One-fifth of the passages is additionally annotated by four native or near-native adult English speakers.

For each candidate passage, the researcher prepares a paraphrase for each function shown in Table 1. The paraphrase selection is in two steps. A participant first selects one of three paraphrases for CONSEQUENCE, ENUMERATION, and EXCEPTION. If CONSEQUENCE is chosen, the participant is asked to accept or reject the ARGUMENTATION paraphrase to further distinguish this subordinate function. We report inter-annotator agreement between researcher and participant in Table 9.

The annotation is completed on the Qualtrics XM Platform.

	Consq/ Enum/Excpt	Arg Yes/No	All
Participant 1	0.79	0.83	0.83
2	0.89	0.87	0.89
3	0.95	0.80	0.91
4	0.84	0.84	0.86
Average	0.87	0.83	0.87

Table 9: Inter-annotator agreement (Cohen’s Kappa) of *otherwise* functions implied by paraphrases between researcher and participant

We observe that the disagreements arise from participant bias toward an *otherwise* function or multiple interpretations of a passage. For example, of four instances where participants infer EXCEPTION and the researcher infers CONSEQUENCE,

⁹COCA requires blank space between tokens.

three instances come from Participant 1. The example below shows that multiple interpretations of a passage is possible:

(1) But the problem was that I wasn’t sure I could make it back to the hotel to catch my flight. Otherwise, I would have been game.

We believe this is one of the cases when both CONSEQUENCE and EXCEPTION may hold, as both paraphrases below are valid:

(1a) But the problem was that I wasn’t sure I could make it back to the hotel to catch my flight. **Because if I could have made it back in time**, I would have been game.

(1b) **Generally**, I would have been game. **An exception is** that I wasn’t sure I could make it back to the hotel to catch my flight.

B Candidate Discourse Markers and Their Continuation Acceptability Scores

For each candidate discourse marker, we provide the distribution of *continuation acceptability* scores from all models in Figure 2. There is no significant variation in the median, and this pattern is consistent across models.

C Surprisal Aggregates

Given a passage s in the order of main clause x , discourse marker d and *otherwise* clause y , a language model returns token-level surprisal for the continuation $s(y_t) = -\log p(y_t|y_{<t}, x, d)$. We then compare the predictive power of the following surprisal aggregates (Giulianelli et al., 2023) in inferring discourse functions:

Mean surprisal is the average of token-level surprisal over all tokens in y :

$$s_{\text{mean}}(y|x, d) = \frac{1}{N} \sum_{n=1}^N s(y_n)$$

Superlinear surprisal is the power sum of token-level surprisal, which indicates that a super-linear effect on y :

$$s_{\text{superlinear}}(y|x, d) = \sum_{n=1}^N [s(y_n)]^k$$

We experiment with $k = \{0.5, 0.75, \dots, 5\}$

Maximum surprisal is the maximum of token-level surprisal. It indicates that the most surprised token captures the overall surprisal of y :

$$s_{\max}(y|x, d) = \max [s(y_n)]$$

Surprisal variance is the variance of token-level surprisal from the mean surprisal.

$$s_{\text{variance}}(y|x, d) = \frac{1}{N-1} \sum_{n=2}^N [s(y_n) - s_{\text{mean}}(y)]^2$$

Surprisal Difference is the sum of differences between contiguous token-level surprisal:

$$s_{\text{difference}}(y|x, d) = \sum_{n=2}^N |s(y_n) - s(y_{n-1})|$$

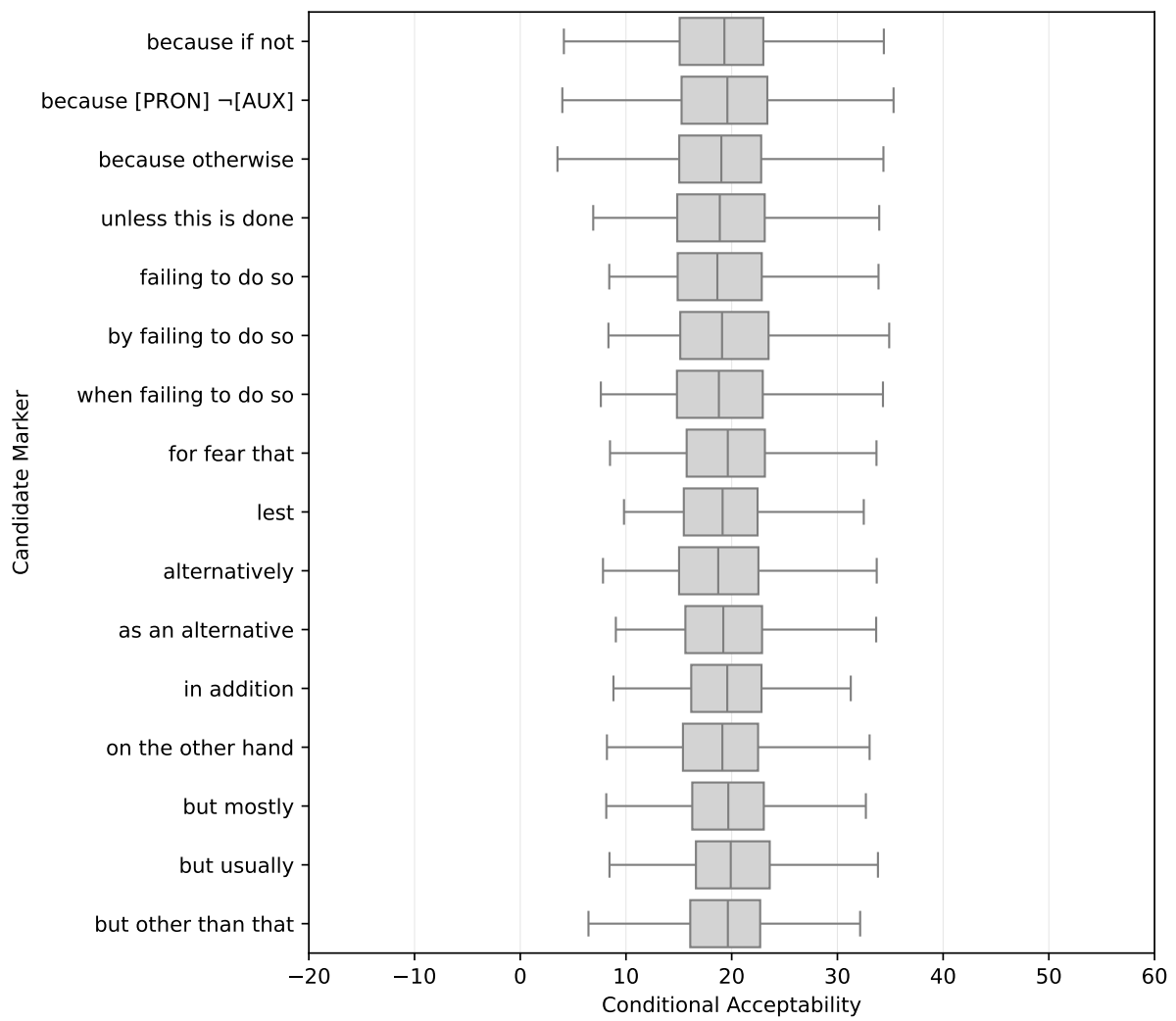


Figure 2: The distribution of *continuation acceptability* score of candidate markers

On the Role of Context for Discourse Relation Classification in Scientific Writing

Stephen Wan[♣] Wei Liu[◇] Michael Strube[◇]

[♣] CSIRO Data61, Australia

[◇] Heidelberg Institute for Theoretical Studies, Germany

stephen.wan@data61.csiro.au

{wei.liu, michael.strube}@h-its.org

Abstract

With the increasing use of generative Artificial Intelligence (AI) methods to support science workflows, we are interested in the use of discourse-level information to find supporting evidence for AI generated scientific claims. A first step towards this objective is to examine the task of inferring discourse structure in scientific writing. In this work, we present a preliminary investigation of pretrained language model (PLM) and Large Language Model (LLM) approaches for *Discourse Relation Classification* (DRC), focusing on scientific publications, an under-studied genre for this task. We examine how context can help with the DRC task, with our experiments showing that context, as defined by discourse structure, is generally helpful. We also present an analysis of which scientific discourse relation types might benefit most from context.

1 Introduction

Recent Artificial Intelligence (AI) advances coupled with the agentic AI approach have seen a burst of activity in the area of “AI for Science”, the application of AI techniques to help accelerate scientific discovery. Examples include usage of *Google’s Co-scientist* (Penadés et al., 2025), OpenAI’s *Deep Research*¹, NVidia’s foundation models for life sciences² and the agentic AI platform *Future House*³. Many of these tools offer an AI research assistant that helps complex research information needs, such as question answering and research planning.

Within these AI for Science applications, generative AI approaches based on Large Language Models (e.g., Brown et al., 2020) are used to generate answers (novel text) to complex questions, introducing the problem of addressing hallucina-

tions and lack of faithfulness (to source references) (Fang et al., 2024).

A popular approach to these problems is to show passages from the source material that supports the generated answer. This approach, sometimes referred to as “contextualising scientific claims”, was the focus of the *Context24* shared task (Chan et al., 2024). Interestingly, the leading contribution in the Context24 shared task demonstrated the utility of scientific discourse cues for detecting such justification material (Bölücü et al., 2024).⁴ This raises an interesting question: *can discourse information be further employed to help in providing supporting evidence for generative AI answers to scientific questions?* A necessary precursor to such an approach would be the ability to infer the discourse structure of a given paper. As a first step towards a study of this topic, in this paper, we focus here on studying the technical challenge of inferring the discourse relations between passages of scientific writing.

We focus on data from two discourse datasets for scientific text, SciDTB (Yang and Li, 2018) and CovDTB (Nishida and Matsumoto, 2021). Both datasets follow the approach that annotates discourse structure as dependency trees, introduced in the SciDTB approach (Yang and Li, 2018). To our knowledge, these are the largest discourse datasets currently available. An example of such a discourse tree is presented in Figure 1.

We present an example from the SciDTB dataset (Yang and Li, 2018) in Figure 2.⁵ The figure shows segmented elementary discourse units (EDUs) for the arguments of the relation. The ground truth relation between the two arguments was annotated as *condition*.⁶ This classification task is highly ambiguous. We note that, in this dataset, the word

¹<https://openai.com/index/introducing-deep-research>

²<https://www.nvidia.com/en-au/use-cases/biomolecular-foundation-models-for-discovery-in-life-science>

³<https://www.futurehouse.org/>

⁴Discourse cues are described as *common expressions* in the original paper.

⁵SciDTB dataset, document ID:P14-1131.

⁶That is, a conditionality for a given situation.

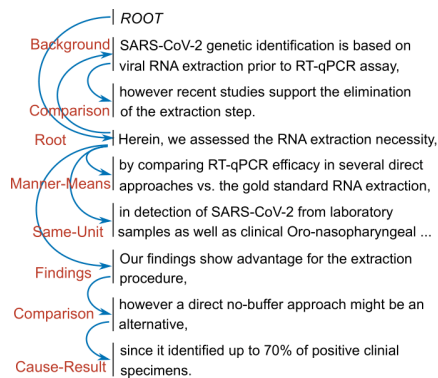


Figure 1: An example of a dependency discourse structure for an abstract for COVID19-related science article. Figure from (Nishida and Matsumoto, 2021)

EDU 1	because it can compute a single node similarity
EDU 2	(rel:condition) <i>without</i> having to compute the similarities of the entire graph .
context	(rel:elaboration) that is efficient ...

Figure 2: An example of how contextual information can help disambiguate which discourse relation holds between two text fragments.

"without" that begins *arg 2* is also associated with two other relations such as *contrast* or *manner*. This ambiguity can be alleviated with additional context. In the example, *arg 1* is annotated as being preceded by the fragment that has an *elaboration* relation and is about *efficiency*. Of the possible relations, this context potentially provides additional information that supports *conditionality* as a more favourable interpretation compared to *contrast* or *manner*.⁷

Inspired by examples such as this and also by recent work in discourse analytics that examines the role of context for DRC in other text genres (e.g., Zhang et al., 2021; see Section 2 for a comprehensive survey), we are further interested in examining the role discourse context for the DRC task. As the majority of these studies have focused on the Penn Discourse Treebanks (PDTB) (Webber et al., 2019) and given that scientific writing (and the discourse relations therein) as notably different from other genres (e.g., see Shi and Demberg, 2019), our goal is to study how context affects DRC for scientific writing. In particular, we are interested in examining whether context selection informed by discourse structure, which we refer to here as

⁷That is, in the example, *computation* is described as *efficient*, not because of an explicit comparison (contrast) nor an indication of how to perform a task (manner) but by virtue of conditions, in this case, *without* the described *expenses*.

discourse structured context, has advantages over other methods, such as adjacent text spans.

This paper makes several unique contributions. We show that structured context helps to improve DRC for scientific writing, as represented by the two datasets, finding that this approach benefits both pre-trained language model and large language model approaches. Furthermore, we present an error analysis to explore the situations context in which context is helpful, revealing some interesting correspondences between scientific discourse relations within the two datasets.

2 Related Work

2.1 Discourse Relation Classification

Discourse Relation Classification (DRC), the task of inferring the relation type that holds between two EDUs, has a rich history and is an actively researched area (Pitler et al., 2009; Gessler et al., 2021; Long and Webber, 2022; Zhou et al., 2022; Liu and Strube, 2023a). The DRC task is typically divided into two types, *explicit* and *implicit* (Pitler et al., 2009). In the former, the two text spans are connected by a discourse relation signalled by an observable discourse connective. The latter is termed implicit as there is an absence of such a connective. The implicit variant is regarded as much harder than the explicit variant (Pitler et al., 2009).

In recent years, to facilitate DRC research, shared tasks have been organised within the DIS-RPT series (Zeldes et al., 2021; Braud et al., 2023). These shared tasks had the benefit of broadening the datasets used to evaluate DRC approaches, which has tended to focus on the PDTB (Prasad et al., 2008). Notably, the two datasets used in this paper were introduced in the 2023 shared task (Braud et al., 2023), however, the submitted approaches at the time did not focus on analysing the use of context for these datasets.

2.2 Context in DRC

The position survey article of Atwell et al. (2021) notes that "most shallow discourse parsers use only the argument pairs to determine the discourse sense without considering ... context." There have, however, been some exceptions. The effect of using discursive context on DRC has been studied in the context of annotation quality and annotator confidence in (Atwell et al., 2022). This work examines the role of context in the PDTB 2 (Prasad et al.,

2008) and 3 (Webber et al., 2019) datasets and shows that annotation quality improves with context (in this case, preceding text), particularly for certain relationships. In that same work, Atwell et al. translate the insights from human annotations and context into modelling insights, exploring the use of the XLNet model for classification with context, that incorporates some modelling of when context is needed. Atwell et al. note the prior work of Scholman and Demberg (2017) in examining the linkage between pronoun use and the need for context within PDTB, again from the perspective of acquiring human annotations.

The sequential modelling of adjacent text spans (and their relationships) has been studied by Dai and Huang (2018), Shi and Demberg (2019) and Zhang et al. (2021). These works generally evaluate on PDTB data. The exception is Shi and Demberg (2019), who also evaluate on biomedical data, albeit by mapping biomedical relations to PDTB discourse relations. Shi and Demberg also argue that the Next Sentence Prediction (NSP) capability of the the BERT model is particularly useful for modelling discourse relations. This same approach is used by Gessler et al. (2021), who also use BERT specifically for its NSP capability but add features relating to the surrounding context. This included direction features and embeddings of the surrounding sentences (to the text spans being considered). Zhang et al. (2021) model discourse structure as a graph and use graph representations in a neural network to capture discourse context, showing benefits for discourse relation classification. These investigations provide alternative representations of context to our study, which does not use specific features or graph representations of context. Our study differs in that we use various methods to select context, including a consideration of the discourse dependency tree, and we prepend contextual text to the EDUs being judged.

While prior treatments of context have shown that it is useful for DRC, these studies generally focus on non-scientific writing, like the PDTB. We note that the approach of Gessler et al. (2021), and the following work of Metheniti et al. (2024), is evaluated on the DISRPT 2021 dataset. However, while this dataset includes data a range of genres, it does not include scientific articles. Indeed, Shi and Demberg (2019) note that the discourse relations are notably different in science literature compared to relations from the PDTB. We argue that this dif-

Dataset	Train:Dev:Test size	Genres
CovDTB	(2400):2400:2587	Science (Biomed)
SciDTB	6061:1935:1912	Science (NLP)

Table 1: Overview of the scientific discourse datasets studied in this work.

ference thus merits a dedicated study of the effects of context for science literature.

2.3 Other Discourse-level Analytics for Scientific Writing

Related to the task of inferring discourse relations for scientific writing are the tasks of argument zoning (Teufel et al., 2009), citation function classification and classification (Teufel et al., 2006; Wan et al., 2009), scientific argument mining (e.g., Lawrence and Reed, 2019; Accuosto and Saggion, 2020; Binder et al., 2022; Fergadis et al., 2021) and science communication sentence classification (Louis and Nenkova, 2013; August et al., 2020); all of which infer a discourse-level features relating to argumentation or scientific writing structure. For this body of work, we note that the use of context has been studied has been studied within the topic of sentence-level categorisation for scientific function (Kiepura et al., 2024). While these associated fields can inform our work, in this paper, we focus on discourse relations rather than argument zones, single sentence classification, or discourse-level relationships across citations.

3 Data

Here, we focus on data from prior work in discourse analysis that provides ground truth annotations for scientific discourse structure, namely the the Covid (Discourse) Dependency Treebank (CovDTB) (Nishida and Matsumoto, 2021), and the Science (Discourse) Dependency Treebank (SciDTB) (Yang and Li, 2018).

3.1 Discourse Dependency Representations

The SciDTB and CovDTB datasets use the dependency discourse structure (DDS) introduced in the SciDTB work (Yang and Li, 2018). Here, structures are directed acyclic graphs, specifically trees. An example of a DDS is shown in Figure 1. In this structure, nodes are EDUs and edges represent the labelled relations between EDUs. The DDS will be used to help select relevant contexts for relation classification. In the DDS, each directed edge (arrows) is a dependency. The figure shows name of

the relation as the text in red. The direction of the relation indicates the importance of the information, with the arrowhead indicating the less important information. By following the links back through the tree towards the root, one can select the relevant context for the classification task, which often is associated with information of greater importance.

3.2 Context Selection Schemes

A context selection scheme relies on two sub-steps: segmentation and filtering preceding text. For this study, for a given EDU in the datasets, there are several options to select context. The simplest filtering method is a *null* method, where no context is used. Alternatively, one can include a preceding text window of some size s . For example, we can also always select the previous sentence for a given argument. However, just because some text precedes an argument does not necessarily mean it is relevant context that impacts text understanding. Thus, we also explore a discourse-oriented method where a discourse structure is employed to identify other text (EDUs) that are linked via discourse relations (referred to here as *structured context*). This gives rise to the following schemes for context selection:

Default This baseline is a *null* context (that is, no context is used).

(ADn) Add-n This approach will add n sentence that precede the first argument.

(ORn) Oracle-n This scheme relies on the ground truth annotations to select the preceding context.⁸ We use the following algorithm. For a pair of arguments considered for DRC, we find the parent node of the *first argument* as defined by the dependency discourse tree. By chaining together the preceding context, in principle, we can vary the amount of context to include. The intuition is that the discourse context, generally represented by a chain of EDUs following a path to the root, indicates which context is important enough to extract.

4 Models

We focus on two general neural network language models approaches, both based on the Transformer

⁸For our motivating example of *contextualising claims*, in practice we would need an initial method to compute a discourse graph connecting EDUs. We leave this to further work but note that prior work explores this task (e.g., Jeon and Strube, 2020).

architecture (Vaswani et al., 2023). The first approach employs the RoBERTa model (Liu et al., 2019), an example of a non-auto-regressive (pre-trained) language model which one can finetune for classification. The second approach uses large language models to perform prompt-based inference, typically used for generative AI.

4.1 PLM-finetuning with RoBERTa

Our RoBERTa-based approach is based on an approach that jointly models discourse connective generation and DRC (Liu and Strube, 2023a). This RoBERTa-based model performs training that combines two tasks: (1) the generation of a discourse connective that would link two arguments; and (2) the classification of the relation given the three pieces of information (argument 1, argument 2, and a connective). As such, this model is extremely flexible and can be applied to both implicit and explicit DRC.

In this study, we generate variants of the datasets, subject to the preprocessing outlined in Section 3, that differ by the amount and type of contextual information that is inserted *before* the first argument (of the relation classification task). That is, context added as per the selection schemes above. The datasets are split into training and testing subsets to train and evaluate the RoBERTa model.

We experimented with the full joint model described here and a simpler version that focuses just on classification.⁹ The latter was found to perform the best and so we report only these results. We used the default training setup and parameters, following the documentation in Liu and Strube (2023b).¹⁰

4.2 LLM-based Inference

Prompt-based generative AI approaches using Large language models (LLMs) have been revolutionary in providing new baseline solutions for many tasks that apply across domains. A key feature of LLMs are the comprehensive training regimes that potentially captures different kinds of knowledge, including common sense knowledge and linguistic capability (e.g., Brown et al., 2020). In this work, we examine two classes of LLMs for this inference approach: *open* and *closed* weight approaches. For the former, we use Meta’s LLaMA

⁹This is achieved by setting the connective to a constant value for all data.

¹⁰For computing environment, see Appendix A.

```

Replace the MASK token (a discourse
relation) by selecting only one of the
following labels: [ label1, label2,
... , labeln] Examples: Passage
1: <arg1_{ex1}>, Passage 2: <arg2_{ex1}>,
connective: <connective_{ex1}> | label1 EOL
<further examples for remaining labels>
EOL
Passage 1: <arg1>, Passage 2: <arg2>,
connective: <connective> | [MASK]

```

Figure 3: GPT4 Prompt template for discourse relation classification

model (et al., 2024). For the latter, we use OpenAI’s GPT4 (OpenAI, 2024).

LLaMA 3.1 In this work, we use a locally hosted version of the LLaMa-3.1-70b-Instruct model, hosted on a server with 16 CPU cores, 128GB memory, and three A6000(48GB) which host the Llama 3.1 70B model.

GPT4 For the GPT4 model, we use OpenAI’s API with the *chatcompletion* endpoint, using the gpt-4-0613 model.

For both models, for classification, we use In-Context Learning (ICL), widely acclaimed as having the ability to surpass supervised machine learning on many NLP tasks as a so-called “few-shot learner” (Brown et al., 2020). This model helps us estimate how context impacts the current methods for LLM-based inference. For this work, temperature was set to zero.

An example of prompt template for GPT4 is presented in Figure 3. This frames the relation classification task as a MASK replacement generation task. The prompt issues an instruction to replace the [MASK] token with one of the class labels provided in list format. ICL “training” examples are provided, where we randomly sample one training data point from the data set for each class label. The two arguments, the connective, and the [MASK] token are then added. The Llama 3.1 prompt is similar.

5 Experiment Results

We conduct an empirical study using ground truth linguistic data to examine the role of discourse structure in inferring discourse relations. We deploy the described models in two conditions: a control condition without discourse context and an experiment condition that includes context.

In the case of PLM-finetuning, for each pair of conditions, we run 10 different trials (that is, neu-

ral network training and testing with 10 different random seeds). We report macro-F1 classification results averaged over the 10 trials. For significance testing, we used the Wilcoxon Signed Ranked Test (WSRT) (Wilcoxon, 1945) and corrected for multiple comparisons using Bonferroni correction.

We first examine if *any* use of context leads to an improvement in the discourse classification task performance. For this investigation, we use $n=1$ for the context selection schemes.

Approach	CovDTB	SciDTB
default	75.54 (1.11)	57.42 (0.65)
AD1	73.45 (2.26)	57.75 (1.08)
OR1	75.78 (1.52)	58.33 (0.77)†

Table 2: Classification with a fine-tuned RoBERTa model. Macro-F1 scores (averaged over 10 runs) with standard deviations in parentheses. Bolded values indicate improvements above the default. Daggers indicate statistical significance improvement using the Wilcoxon Signed Rank Test ($\dagger : \alpha = 0.05$).

5.1 Context and PLM Fine-Tuning

Table 2 shows the results of including context for the PLM-finetuning approach using the RoBERTa model. The table presents macro-F1 scores to give some indication of performance across the unbalanced dataset.¹¹ We find that context generally helps for DRC when using a fine-tuned PLM, particularly when context is defined using discourse structure (OR1). This improvement is statistically significant for the SciDTB dataset (WSRT $p < 0.05$). The AD1 context does not lead to strong performance improvements in comparison. However, we note that the AD1 text window variant of context also helps mildly for the SciDTB but not for CovDTB.

Across the two datasets, we observe that the performance results are higher in the CovDTB dataset compared to the SciDTB dataset. This could be due to conventions in scientific writing for biomedical literature which may be more homogenous than the data from NLP domain found in SciDTB.

¹¹Here we report results for $n = 1$ as our experiments showed that for larger values, adding more context confused the models. Similarly, while we explored variants of the context representations that additionally utilised the relation class (for the linked context), the results were comparable to reported the OR1 variant, from which we conclude that the extra information did not help.

Approach	CovDTB	SciDTB
default	32.07	22.06
AD1	26.19	19.28
OR1	49.07	52.61

Table 3: GPT4 model: Classification macro-F1 scores.

Approach	CovDTB	SciDTB
default	11.20	07.71
AD1	10.04	05.15
OR1	10.36	11.15

Table 4: Llama 3.1 model: Classification macro-F1 scores.

5.2 Context and LLM Prompt-based Inference

Table 3 presents the corresponding results for the GPT4 model.¹² The results indicate a poor performance by the LLM for DRC under the default setting (with no context), even when using in-context learning. Performance improves when discourse structure is used to provide context, as opposed to the adjacent text. Indeed, performance drops when the adjacent sentence is used as context. In the case of SciDTB, this brings performance closer to the PLM-finetuning result, within a margin of 5 F1 points. However, while the DRC performance on CovDTB increases with discourse structure context, the macro-F1 scores still remain far behind the fine-tuned PLM, by a margin of over 25 F1 points.

In Table 4, we see a similar story, although Llama 3.1 performance is much lower than GPT4. We suspect that this is primarily due to the size of the model; the Llama model used here has orders of magnitude fewer parameters than the GPT model. Again, using the adjacent sentence leads to a drop in performance. Consistent with other results, an improvement in DRC performance was observed for SciDTB when using context defined by discourse structure. Here, we failed to detect any improvement with the CovDTB dataset.

We note that, while our focus is on comparison against our default baseline and the relative difference in performance with and without context, the models described in this paper are competitive with the reported performance in the literature. We report these values for completeness in Table 5, which lists comparisons with literature, with the metrics generally reported by convention. With the RoBERTa model and the “oracle” use of the ground

¹²Given the cost of using the commercial GPT LLM, we report results on single trials for the datasets.

truth discourse annotations used here, the measured accuracy of 83.63 represents an estimate of an improvement over the prior state-of-the-art (SOTA) result for the CovDTB that could be obtained if we were to employ a fully automated version of the inference.

Approach	CovDTB	SciDTB
Performance	70.03 Acc.	75.30 Acc.
OR1 model	83.03 Acc.	74.81 Acc.

Table 5: A comparison with performance reported in the literature. **Bolded** values indicate where our best RoBERTa model surpasses previous results. Reported performance from: covdtb and scidtb (Liu et al., 2023).

To summarise, the results show positive trends for using discourse context for the DRC task. Generally, discourse context can help with the PLM-finetuning approach and LLM-inference. When applying the text window context (AD1) results are mixed; the method does not work all the time and can decrease performance. However, when using discourse structure to determine relevant context (OR1) we generally see improved performance, with stronger gains demonstrated with the SciDTB dataset. This indicates that not all preceding text is useful for classification and that indiscriminately adding more context (without filtering) can make performance worse.

5.3 A Reflection on Datasets

We speculate that one reason why we do not see a bigger effect from the inclusion of discourse context is that our datasets may be limited to relatively short length of the text data. Indeed, Yang and Li (2018) note a related issue when studying news articles: discourse relations do not cross paragraph boundaries further making structures shallow.¹³

In this regard, the CovDTB and SciDTB datasets, as examples of short text data (i.e., abstracts) may also have simpler discourse structures than longer texts. We further investigated the nature of the discourse structures and found that, in the case of the SciDTB dataset, the structures were generally short-distance dependencies: 61% of relationships are adjacent, with 10% of relations separated by a gap of 3-5 sentences. We posit that when considering longer documents, the effect of structured context in DRC may be more pronounced.

¹³This is presumably due to journalistic writing style.

6 Error Analysis: DRC for Scientific Discourse Relations

In the experiments presented above, we observed that providing context, particularly *structured* context generally helps with DRC. In this section, we perform an error analysis to better understand when context helps, analysing performance per relation type. Here, we focus on the fine-tuned PLM approach as it yielded the highest macro-F1 scores. For each of the 10 seed runs, we used predictions from the best models for the default (no context) and the OR1 (structured context) conditions. Cases where the OR1 prediction was correct and the default was not was considered a *win*. The converse case was considered a *loss*. Where both approaches agreed, this was considered a *tie*. Margins for wins and loss were averaged over the 10 runs.

Table 6 provides a list of the scientific discourse relation types in the *winning* outcome for both data sets that benefited (overall) from OR1 context and their average win margins. We can see that *elaboration*, *comparison*, *attribution*, and *temporal* relations were common to both datasets.

CovDTB	SciDTB
elaboration ($\Delta = 5.7$)	elab-addition ($\Delta = 1.5$)
<u>enablement</u> ($\Delta = 1.2$)	elab-aspect ($\Delta = 0.8$)
<u>cause-result</u> ($\Delta = 0.8$)	temporal ($\Delta = 0.77$)
<u>condition</u> ($\Delta = 0.4$)	bg-compare ($\Delta = 0.66$)
attribution ($\Delta = 0.4$)	<u>joint</u> ($\Delta = 0.44$)
comparison ($\Delta = 0.3$)	contrast ($\Delta = 0.33$)
temporal ($\Delta = 0.1$)	progression ($\Delta = 0.22$)
	exp-reason ($\Delta = 0.22$)
	elab-enum ($\Delta = 0.22$)
	comparison ($\Delta = 0.11$)
	attribution ($\Delta = 0.11$)

Table 6: Winning relations: Discourse relations who DRC performance improved with the inclusion of structured context. Δ = indicates the average win/loss margin. Bolded text indicates potential correspondences across datasets.

Table 7 presents the corresponding table for the *losing* outcome. Here we see that, both data sets have *background* relations in common for this outcome. If we assume *findings* and *results* are related relations, then we can consider this a further potential alignment.

Table 8 shows the relations that had an equal number of wins and losses. We present these for completeness. However, it may be the case that, for these datasets, there is insufficient data to assign these to either the winning or losing outcomes.

There were some differences between datasets

CovDTB	SciDTB
findings ($\Delta = 0.9$)	bg-goal ($\Delta = 0.44$)
background ($\Delta = 0.3$)	manner-means ($\Delta = 0.44$)
	<u>enablement</u> ($\Delta = 0.44$)
	<u>evaluation</u> ($\Delta = 0.22$)
	result ($\Delta = 0.22$)
	bg-general ($\Delta = 0.22$)
	<u>condition</u> ($\Delta = 0.11$)
	exp-evidence ($\Delta = 0.11$)

Table 7: Losing relations: Discourse relations who DRC performance suffered with the inclusion of structured context. Δ = indicates the average win/loss margin. Bolded text indicates potential correspondences across datasets.

CovDTB	SciDTB
textual-organisation	elab-definition
manner-means	elab-process-step
	cause

Table 8: Tied relations: Discourse relations who DRC performance remained the same with the inclusion of structured context.

for a subset of relations, which were placed in different outcomes (winning, losing). These included *enablement* and *condition*. In CovDTB, a single relation is used for *cause-result* which was in the winning outcome. For the SciDTB dataset, the *result* relation was in the losing outcome. Similarly, while most background relations were in the losing outcome for both datasets, *bg-compare* was in the winning outcome for SciDTB; though this could be because the winning outcome contained more comparison-related relations. We treat these divergences as interesting outcomes to investigate further, noting that some of these may be due to annotation differences between the datasets.

In Table 9, we present some examples of data as assigned to the winning and losing outcomes. For the winning outcomes, the high-level statement of the research activity as context may contribute positively to the DRC task. For the losing outcome, we note that in the SciDTB example, the high-level context may simply be too broad. For the CovDTB, we note that both findings and background relations tended to be at the beginning of the text and so no prior context exists, explaining why these relations are in the losing outcome.

To dive deeper into what might potentially explain the difference between winning and losing outcomes, we examined the first word of the second argument, checking for a match against a list of known discourse connectives.¹⁴ Here, we make

¹⁴This list was based connectives from PDTB data

Condition	Dataset	Relation	Example
winning	scidtb	comparison	Context: We propose a novel method of jointly embedding entities and words into the same continuous vector space . Arg1: that jointly embedding brings promising improvement in the accuracy of predicting facts , Arg2: compared to separately embedding knowledge graphs and text .
losing	scidtb	result	Context: We describe a search algorithm Arg1: Our results show Arg2: parsing results significantly improve
winning	covdtb	comparison	Context: Herein we discuss application of the Collaborative Cross (CC) panel of recombinant inbred strains Arg1: Although the focus of this chapter is on viral pathogenesis , Arg2: many of the methods are applicable to studies of other pathogens , as well as to case-control designs in genetically diverse populations .
losing	covdtb	findings	Context: ROOT (no context) Arg1: In this work , we demonstrate a design of meta - holography Arg2: that can achieve 2 28 different holographic frames and an extremely high frame rate in the visible range .

Table 9: Examples of discourse relations in the winning and losing outcomes for both the SciDTB and CovDTB datasets.

Relation Category	Percentage of connective matches	
	CovDTB	SciDTB
Losing Relations	7.8%	25.0%
Winning Relations	16.6%	28.2%

Table 10: Percentage of matches to a list of explicit connectives across the positive, neutral and negative relations.

the simplifying assumption that the discourse connective is found between the two arguments.

Table 10 shows the percentage of instances where, for relations in either the winning or losing outcome, the first word of the second argument was a known discourse connective. We observe that, for both datasets, winning outcomes exhibit a higher percentage of matches for connectives. We take the matches as a potential indicator of the higher proportion of explicitly marked discourse relations. This raises the potential hypothesis that perhaps context may be more beneficial for DRC of certain explicitly marked relations.

7 Future Work

Our preliminary investigation here on the role of context for DRC in scientific writing highlights two potential avenues for future research. Our error analysis suggests that structured context may potentially be more beneficial the DRC for certain explicit relations (for scientific writing). We intended to further investigate this.

We note that our investigation here is limited

sets (2 and 3) and the collated connectives from the DiscoGEM dataset (Scholman et al., 2022). URL: <https://github.com/merelscholman/DiscoGeM/tree/main/Appendix>

to dependency discourse structures and the representation of context as string concatenations. In subsequent work, we aim to explore different automatically inferred graph representations of text structure, particularly longer text documents.

Our experiments were also limited to two categories of LLM-based inference, namely In-Context Learning (ICL) for closed and open weight LLMs (or proprietary and so-called "open-source" LLMs). In the future, we intend to include LLMs that include some reasoning capability, such as the recent GPT-o1 and DeepSeek models, as well as techniques like chain of thought, to see if these inference methods help with DRC. In this work, we also used one example of a transformer network for PLM fine-tuning. In future work, we aim to experiment with the model of (Gessler et al., 2021) as an alternative competing transformer model.

Finally, returning to our motivating example, we intend to examine the role of discourse relations in identifying relevant supporting source material to validate generative AI output. We intend to conduct qualitative and quantitative user studies to better understand the potential for discourse information to help with these goals.

8 Conclusions

In this work, we showed that adding discourse context, particularly *structured context*, helps with Discourse Relation Classification for scientific writing. We demonstrated this using two dominant neural language modelling methods: finetuning using a pre-trained language model, and inference with large language models using in-context learning.

The analysis presented here focuses on two scientific discourse datasets, CovDTB and SciDTB, representing biomedical and computer science disciplines. We found that, for the science discourse relations represented in these datasets, context might help for specific relations, such as with *elaboration*, *attribution*, *comparison* and *temporal* relations.

Acknowledgements

This work was funded by the CSIRO Julius Career Award. We are also grateful to the Heidelberg Institute for Theoretical Studies (H-ITS) for supporting this project and providing facilities for conducting this research. We would like to further acknowledge the feedback from the CSIRO Language Technology team, the H-ITS NLP team, and the anonymous reviewers on previous versions of this paper.

Limitations

In this work, we focused on English prose language data from publicly available datasets. As such, our conclusions about discourse relations, connectives and the need for using context for discourse relation classification are limited to this language and the genres represented. We note that while we are interested in scientific writing in general, here we study data from just two science disciplines: computer science and biomedical articles about Covid. We note that we only generated a single set of results using prompt-based methods (with Llama 3.1 and GPT 4), using a temperature of zero, due to costs. It is possible that multiple trials of the approach may yield different results. Finally, we note that prompt engineering was limited. It may be possible that stronger performance gains may be obtained if further prompt engineering is employed. For further limitations, see our future work section.

Ethical Considerations

In this work, we use publicly available discourse-related datasets. Our analysis is focused discourse-related linguistic phenomena and is not focused on any individual or subgroup in the community. The work, while motivated by current trends in applied AI, is not immediately applicable in real-world usage.

References

- Pablo Accuosto and Horacio Saggion. 2020. [Mining arguments in scientific abstracts with discourse-level embeddings](#). *Data and Knowledge Engineering*, 129.
- Katherine Atwell, Remi Choi, Junyi Jessy Li, and Malihe Alikhani. 2022. [The role of context and uncertainty in shallow discourse parsing](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 797–811, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. [Where Are We in Discourse Relation Recognition?](#) Technical report.
- Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020. [Writing strategies for science communication: Data and computational analysis](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.
- Arne Binder, Leonhard Hennig, and Bhuvanesh Verma. 2022. [Full-text argumentation mining on scientific publications](#). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 54–66, Online. Association for Computational Linguistics.
- Necva Bölücü, Vincent Nguyen, Roelien C Timmer, Huichen Yang, Maciej Rybinski, Stephen Wan, and Sarvnaz Karimi. 2024. [CSIRO-LT at Context24: Contextualising Scientific Figures and Tables in Scientific Literature](#). Technical report.
- Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. [The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). *CoRR*, abs/2005.1.
- Chu Sern Joel Chan, Aakanksha Naik, Matthew Akamatsu, Hanna Bekele, Erin Bransom, Ian Campbell, and Jenna Sparks. 2024. [Overview of the Context24 Shared Task on Contextualizing Scientific Claims](#).

- In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 12–21, Bangkok, Thailand. Association for Computational Linguistics.
- Zeyu Dai and Ruihong Huang. 2018. [Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.
- Aaron Grattafiori et al. 2024. [The llama 3 herd of models](#).
- Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. [Understanding faithfulness and reasoning of large language models on plain biomedical summaries](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.
- Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. [Argumentation mining in scientific literature for sustainable development](#). In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. [DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sungho Jeon and Michael Strube. 2020. [Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online. Association for Computational Linguistics.
- Anna Kiepura, Yingqiang Gao, Jessica Lam, Nianlong Gu, and Richard H.r. Hahnloser. 2024. [SciPara: A new dataset for investigating paragraph discourse structure in scientific papers](#). In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 12–26, St. Julians, Malta. Association for Computational Linguistics.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Wei Liu and Michael Strube. 2023a. [Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu and Michael Strube. 2023b. [Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation](#).
- Wei Liu, Fan Yi, and Michael Strube. 2023. [HITS at DISRPT 2023: Discourse Segmentation, Connective Detection, and Relation Classification](#). *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (Disrpt)*:43–49.
- 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} Pre-training Approach](#). *CoRR*, abs/1907.1.
- Wanqiu Long and Bonnie Webber. 2022. [Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Annie Louis and Ani Nenkova. 2013. [What makes writing great? first experiments on article quality prediction in the science journalism domain](#). *Transactions of the Association for Computational Linguistics*, 1:341–352.
- Eleni Metheniti, Chloé Braud, and Philippe Muller. 2024. [Feature-augmented model for multilingual discourse relation classification](#). In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.
- Noriki Nishida and Yuji Matsumoto. 2021. [Out-of-Domain Discourse Dependency Parsing via Bootstrapping: {A}n Empirical Analysis on its Effectiveness and Limitation](#). *Transactions of the Association for Computational Linguistics*.
- OpenAI. 2024. [Gpt-4 technical report](#).
- José R Penadés, Juraj Gottweis, Lingchen He, Jonasz B Patkowski, Alexander Shurick, Wei-Hung Weng, Tao Tu, Anil Palepu, Artiom Myaskovsky, Annalisa Pawlosky, Vivek Natarajan, Alan Karthikesalingam, and Tiago R D Costa. 2025. [AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution](#).
- Emily Pitler, Annie Louis, and Ani Nenkova. 2009. [Automatic sense prediction for implicit discourse relations in text](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). Technical report, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.
- Merel Scholman and Vera Demberg. 2017. [Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task](#). In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. Association for Computational Linguistics.
- Merel C J Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. [DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations](#). In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France. European Language Resources Association (ELRA).
- Wei Shi and Vera Demberg. 2019. [Next sentence prediction helps implicit discourse relation classification within and across domains](#). 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 5790–5796, Hong Kong, China. Association for Computational Linguistics.
- Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. [Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics](#). In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.
- Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. [Automatic classification of citation function](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#).
- Stephen Wan, Cécile Paris, and Robert Dale. 2009. [Whetting the appetite of scientists: producing summaries tailored to the citation context](#). In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '09*, pages 59–68, New York, NY, USA. Association for Computing Machinery.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. [The Penn Discourse Treebank 3.0 Annotation Manual](#). Technical report.
- Frank Wilcoxon. 1945. [Individual Comparisons by Ranking Methods](#). *Biometrics Bulletin*, 1(6):80–83.
- An Yang and Sujian Li. 2018. [SciDTB: Discourse dependency TreeBank for scientific abstracts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.
- Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. [The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. [Context tracking network: Graph-based context modeling for implicit discourse relation recognition](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599, Online. Association for Computational Linguistics.
- Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. [Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix: Computing Environment

Experiments for training and evaluating the ConRel model (RoBERTa-based, 82M parameters) were conducted on a server with 1 node (4x NVIDIA A40; 2x Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz; 32GB RAM). Each of the 3 approaches tested was trained and evaluated with 5 datasets, over 10 trials. Each trial ranged from between 30 minutes to 1.5 hours, depending on the dataset. Estimated GPU time per approach is 36 hours. Experiments were also repeated at least twice to test for replicability. This results in approximately, 432 hours of GPU time (single jobs).

Zero-Shot Belief: A Hard Problem for LLMs

John Murzaku^{♦♠}, Owen Rambow^{♠♠}

[♦]Department of Computer Science [♠]Department of Linguistics

[♠]Institute for Advanced Computational Science

Stony Brook University

jmurzaku@cs.stonybrook.edu

Abstract

We present two LLM-based approaches to zero-shot source-and-target belief prediction on FactBank: a unified system that identifies events, sources, and belief labels in a single pass, and a hybrid approach that uses a fine-tuned DeBERTa tagger for event detection. We show that multiple open-sourced, closed-source, and reasoning-based LLMs struggle with the task. Using the hybrid approach, we achieve new state-of-the-art results on FactBank and offer a detailed error analysis. Our approach is then tested on the Italian belief corpus ModaFact.

1 Introduction

The term “belief” (interchangeably referred to as “event factuality” in NLP) refers to the extent an event mentioned by the author or by sources in a text is presented as being factual. While this task has received attention over the years, no zero-shot experiments have been performed. We show that this task remains a hard task for LLMs.

Our major contributions are as follows:

- (1) We present unified and hybrid zero-shot frameworks for the source-and-target belief prediction task (i.e., who has what belief towards what). We test our approach on various LLMs.
- (2) Our hybrid approach achieves new state-of-the-art results (SOTA) on the FactBank corpus, but the problem is far from solved.
- (3) We are the first to evaluate FactBank on Nested belief, revealing that LLMs perform particularly poorly on this task. We perform an error analysis showcasing where LLMs fail.
- (4) We validate the transferability of our approach by testing on the Italian ModaFact belief corpus.

This paper is organized as follows: we provide an overview of the belief detection task in Section 2. We follow by detailing our methodology in Section 4 and discuss our results and analysis for FactBank and ModaFact in Section 5.

2 Related Work

Corpora Many corpora explore the notion of belief on the sentence level. FactBank is one of the first corpora to do this, annotating source-and-target belief: both the belief presented by the author towards an event and the belief towards events by sources mentioned inside of the text (Saurí and Pustejovsky, 2009). Other corpora annotate only the author’s belief towards events: these corpora include LU (Diab et al., 2009), UW (Lee et al., 2015), LDCCB (Prabhakaran et al., 2015), MEANTIME (Minard et al., 2016), MegaVeridicality (White et al., 2018), UDS-IH2 (Rudinger et al., 2018), CommitmentBank (De Marneffe et al., 2019), and RP (Ross and Pavlick, 2019). Two recent corpora for event factuality are Maven-Fact (Li et al., 2024) which contains a large-scale corpus of event and supporting evidence annotations, and ModaFact (Rovera et al., 2025), which is an Italian author belief corpus that annotates in a similar style and inspiration as FactBank.

Methods Previous methods for author belief prediction mainly involve fine-tuning: Rudinger et al. (2018) fine-tune multi-task LSTMs; Pouran Ben Veysseh et al. (2019) fine-tune a graph convolutional network with BERT (Devlin et al., 2019) representations; Jiang and de Marneffe (2021); Murzaku et al. (2022) fine-tune RoBERTa (Liu, 2019) with span representations; Li et al. (2024) fine-tune RoBERTa and Flan-T5 (Chung et al., 2024), and also explore four LLMs predictions using few-shot learning; Rovera et al. (2025) fine-tune BERT, mT5-XXL (Xue et al., 2021), Aya23-8B (Aryabumi et al., 2024), and Minerva-3B (Orlando et al., 2024).

There has been much less focus on the complete source-and-target belief task: Murzaku et al. (2023); Murzaku and Rambow (2024) both fine-tune a Flan-T5 model, with the latter optimizing for the structure of belief represented as a tree.

3 Preliminaries

Consider this sentence: *Trurit Inc. said it is phasing out legacy routers.* This sentence reports on two events: a “said” event and a “phasing” event.

Author Belief The definition of author belief (also called event factuality) is how committed is the author of the text (the source) to the truth (or factuality) of an event. In this sentence, the author is presenting the “said” event as factual, i.e., they are committed to the “said” event having happened. On the other hand, the author is presenting the “phasing” event as having an unknown factuality; the author is not directly committing to the truth of the event, rather they are reporting on what “Trurit Inc.” said.

Nested In nested belief, we report the belief towards events according to nested sources inside of a text. The task can be split into three steps: (i) identifying the nested or attributed source in the text; (ii) linking the source to the events (i.e., which events does the source commit to); (iii) labelling the belief of the event according to that source. In our example, the source is “Trurit Inc.” Once the source is introduced (i), we then link the source to the events in the text (ii): in this case, Trurit Inc. is reportedly committing to the event “phasing”, and asserting it as true (iii). Since the source is reporting about this event, and directly committing to the event happening, it is therefore true in Trurit Inc.’s perspective as reported by the author, and unknown in the author’s perspective.

4 Methodology

We use the test set of the source-and-target (author and nested sources) projection of FactBank released by [Murzaku and Rambow \(2024\)](#). Further dataset details are in Appendix D.

4.1 Zero-Shot

Unified Our **Unified** approach provides a single end-to-end zero-shot prompt to the LLM with the input text, a high-level descriptions of the task, the three main steps in the annotation process in detail, special cases guidelines, and the output format. We end the prompt with a summary of the specific steps on how to produce the final answer in a chain-of-thought format (CoT) ([Wei et al., 2022](#)), which has proven to work well for author event factuality ([Li et al., 2024](#)). The three steps are: (1) Label all events according to the FactBank annotation guidelines, which we provide. (2) Identify all nested

sources in the text. (3) Assign factuality labels for each event, according to that source. We leave all model details, API parameters, and our exact prompts in Appendix A.

Hybrid For our **Hybrid** zero-shot approach, we first extract events in a sentence using a DeBERTa ([He et al., 2021](#)) based tagger. After extracting the events, we prompt an LLM with the sentence and the list of events. We then follow the exact steps (minus event detection, since we provide events) as our **Unified** prompt: we instruct the LLM to identify all nested sources, ask the LLM to assign factuality labels for the events, according to the identified sources, and finish with instructions for answering with CoT. See Appendix B for further details on our **Hybrid** experiments and our exact prompts.

Event Tagger The FactBank corpus has a complex definition of what exactly is an annotatable event. [Murzaku et al. \(2023\)](#) found that annotating FactBank events is non-trivial, even with a specialized, generative fine-tuned model achieving only 85.4% F1 on event identification. We therefore choose to fine-tune a DeBERTa model for event token detection, and then pass the events to our **Hybrid** prompts.

4.2 Models

We perform experiments on a variety of LLM types: open LLMs, specifically LLaMA-3.3-70B ([Meta, 2024](#)), DeepSeek-v3 ([Liu et al., 2024](#)), and DeepSeek-r1; closed LLMs, specifically GPT-4o ([OpenAI, 2024a](#)), newly released reasoning models o1 ([OpenAI, 2024b](#)) and o3-mini ([OpenAI, 2025](#)), and Claude 3.5 Sonnet ([Anthropic, 2025](#)); and reasoning LLMs, DeepSeek r1 (henceforth R1), o1, and o3-mini.

4.3 Evaluation: Metrics

We evaluate on three F1 metrics: Full where we perform an exact match evaluation on all generated (source, event, label) annotations; Author where we perform an evaluation on all generated annotations where the source is the author of the text; and Nest where we perform an exact match evaluation on all generated annotations where the source is a nested source.

4.4 Evaluation: FactBank Sources

FactBank has specific conventions about annotating sources. Consider the example “Trurit Inc. shares rose by 5% today”. FactBank annotates on the









Model	Unified			Hybrid			Δ Hyb.-SOTA		Δ Hyb.-Unif.		
	Full	Author	Nest	Full	Author	Nest	Full	Author	Full	Author	Nest
<i>Previous / Fine-Tuned SOTA (Murzaku and Rambow, 2024)</i>											
GPT-3 (Fine-tuned)	65.8	76.0	–	65.8	76.0	–	–	–	–	–	–
Flan-T5-XL	69.5	76.6	–	69.5	76.6	–	–	–	–	–	–
<i>Zero-Shot LLM Systems</i>											
GPT-4o	60.2	65.9	20.2	68.7	73.2	22.9	-0.8	-3.4	+8.5	+7.3	+2.7
o1 	65.0	73.2	18.9	70.3	78.9 [†]	19.2	+0.8	+2.3	+5.3	+5.7	+0.3
DeepSeek r1  	66.1	71.1	24.1	72.0 [†]	77.6	25.3 [†]	+2.5	+1.0	+5.9	+6.5	+1.2
o3-mini 	62.4	70.9	15.6	65.5	75.2	17.0	-4.0	-1.4	+3.1	+4.3	+1.4
Claude 3.5	63.2	69.7	19.7	70.4	77.6	21.4	+0.9	+1.0	+7.2	+7.9	+1.7
LLaMA 3.3 	53.1	60.4	14.4	58.8	66.0	19.9	-10.7	-10.6	+5.7	+5.6	+5.5
DeepSeek-v3 	56.3	61.4	17.1	60.5	65.3	18.2	-9.0	-11.3	+4.2	+3.9	+1.1

Table 1: **Unified** vs. **Hybrid** approaches with different LLMs. We report Micro F1 (Full), Author Micro F1 (Author), and Nested Micro F1 (Nest) scores (in %). Δ **Hyb.-SOTA** denotes the difference between the **Hybrid** result vs. the fine-tuned SOTA. The best scores are highlighted in **bold** and new state-of-the-art (SOTA) results are denoted by \dagger . Δ **Hyb.-Unif.** highlight the **Hybrid-Unified** difference for Full, Author, and Nest F1s.  indicates open models and  indicates reasoning models.

token level, and the source is “Inc.”. We do not wish to penalize LLMs for not knowing this conversion, and therefore propose a few-shot normalization technique for preprocessing. We perform all source normalization experiments with GPT-4o (OpenAI, 2024a). Exact prompts for our source normalization methods and a detailed ablation study are shown in Appendix E.

Our task setup is as follows: Given a predicted source, we prompt GPT-4o to transform the predicted source into a FactBank-compliant version.

5 Results and Analysis

5.1 FactBank

Main Results Our main results for FactBank are shown in Table 1. We compare all our results to the previous fine-tuned SOTA from Murzaku and Rambow (2024), evaluating on exact match F1 (Full) and author exact match F1 (Author) as described in Section 4.3. We add one more metric: nested exact match F1 (Nest), where we evaluate on nested sources only.

Our **Unified** zero-shot results (column **Unified**) achieve competitive performance compared to fully fine-tuned models, with R1 (66.1% for Full) and o1 (73.2% for Author). We outperform the fine-tuned GPT-3 model from Murzaku and Rambow (2024) on Full, but do not outperform the Flan-T5-XL system.

We achieve new SOTA on FactBank with our **Hybrid** systems. Our R1 **Hybrid** system achieves Full of 72.0%, outperforming the previous state of the art by 2.5% (column Δ vs. SOTA). Similar to the **Unified** results, o1 excels in Author, achieving 78.9% Author and outperforming the previous SOTA by 2.3%. We also note that GPT-4o and Claude-3.5 also achieve competitive performance, with Claude-3.5 outperforming the previous SOTA on Full and Author by 0.9% and 1.0% respectively. We hypothesize that these models excel due to CoT prompting.

Nest F1 We are the first to provide Nest F1 metrics on FactBank. Our top performing model is r1, which achieves a nested F1 of 25.3%. For reasoning models o1, o3-mini, and r1, we notice that going from **Unified** to **Hybrid** does not increase Nest F1 dramatically (0.3% for o1, 1.4% for o3-mini, and 1.2% for r1), showcasing the models’ lack of capabilities for nested belief predictions. We note that these results are low, and believe modelling of nested beliefs is essential future work and a challenging task for reasoning LLMs.

Zero vs. Hybrid We quantify the exact difference (in %) between our **Unified** and **Hybrid** models in Table 1 (column Δ (Hyb.-Unif.)). We see improvements in every model, with the greatest improvements occurring in GPT-4o and Claude-3.5 for Full and Author. On average, our **Hybrid** models

Model	Type	F1
DeBERTa	Fine-tuned	89.0
DeepSeek R1	Zero-shot	82.0
	Few-shot	76.4
GPT-4o	Zero-shot	78.2
	Few-shot	81.1
Claude 3.5	Zero-shot	83.3
	Few-shot	81.8

Table 2: Event detection performance (in % F1) of various language models. The fine-tuned DeBERTa model outperforms all major LLMs in zero-shot and few-shot settings.

outperform our **Unified** models by 5.7% for Full, 5.9% for Author, and 2.0% for Nest. Our results emphasize the need for a specialized event tagger and hybrid approach, allowing the LLM to focus on linking sources and tagging belief labels.

Event Detection We investigate how LLMs perform on event tagging. We show these results in Table 2. We compare three LLMs (r1, GPT-4o, and Claude-3.5) to the fine-tuned DeBERTa event tagger used in the **Hybrid** system. For our LLMs, we try two configurations: zero-shot and few-shot (5 examples). We find that a fine-tuned DeBERTa outperforms all LLMs in all settings, emphasizing that event detection is still a difficult task. We leave further experimental details and prompts in Appendix F.

Error Analysis We perform an error analysis on the top-performing model (R1, **Hybrid**) on nested beliefs (F1 of only 25.3%). We categorize errors as follows: **(1) Source** mismatch, often labeling the author instead of the nested source or failing to classify pronoun sources such as “it” correctly (123 errors); **(2) FN** (false negatives on events), where context-dependent event nouns or verbs are missed (e.g., “*acquisition*,” “*construction*”) (77 errors); **(3) FP** (false positives on events), over-predicting event nouns (73 errors); **(4) Label** errors, notably predicting *True* or *Probable* instead of *Unknown* for future/uncommitted events (e.g., “Mary offered to **buy** an apple”, where the **buy** event should be *Unknown*) (53 errors). We note that the FN errors are consistent with findings from prior FactBank studies: Murzaku et al. (2022) also found similar errors. More detailed results and analysis of our error analysis are in Appendix G.

Model	Method	Bel.+Pol.
mT5 XXL	Fine-tune	64.4
DeepSeek r1	Hybrid	63.6 [†]
o3-mini	Hybrid	62.6 [†]
GPT-4o	Hybrid	61.2 [†]
GPT-4o	Unified	42.9
o3-mini	Unified	40.8
DeepSeek r1	Unified	38.6

Table 3: Model performance on Belief+Polarity (Bel.+Pol.) F1. Rovera et al. (2025) mT5-XXL baseline is shown in **bold**. Results on Bel.+Pol. metric within 5% of the SOTA are marked with a [†].

5.2 Multilingual Belief

The ModaFact Italian corpus (Rovera et al., 2025) annotates the author’s belief, polarity, and modality towards events and temporal information. We only use the belief and polarity annotations and combine these to tags similar to those of FactBank (and perform an exact match evaluation on Belief+Polarity F1). This is different from how Rovera et al. (2025) evaluate, but they kindly shared their raw results so that we could apply our evaluation.

Results We perform our ModaFact experiments with three cost-effective models that performed well for FactBank: GPT-4o, o3-mini, and R1. Our results are shown in Table 3. Unlike our FactBank results, we fall short of the fine-tuned SOTA for our **Hybrid** system (by 0.8%). Similar to our FactBank results, **Hybrid** strongly outperforms **Unified** in all settings. Finally, we see that R1 and o3-mini (both reasoning models) come very close to the fine-tuned SOTA. GPT-4o also proves competitive, but falls short of the reasoning models by 2.4%. We note that while we do not beat the SOTA, the LLMs we use are not explicitly trained for multilingual data (in contrast with mT5-XXL). For example, R1 is specifically optimized for English and Chinese data (Guo et al., 2025). We hypothesize that future multilingual optimizations for these reasoning LLMs would in fact lead to a new SOTA for the ModaFact corpus.

6 Conclusion

We show that belief detection from text remains a challenging problem for LLMs. This is particularly true for nested beliefs, which the author ascribes to other sources. Our new SOTA system includes a distinct fine-tuned event detection component.

Limitations

While our model achieves a new state-of-the-art on the English only FactBank, our results, while still competitive, do not perform as well for the Italian ModaFact corpus. We acknowledge this as a shortcoming and aim to work towards broader multilingual generalization for this task.

We note that our LLM approach yields poor results on the nested F1 metric, indicating a large gap and potential for future improvement. We will explore improving these results in future work and believe this to be a gap for all major open, closed, and reasoning LLMs.

Finally, we note that our top performing LLM approach, while using the open DeepSeek r1 model, is reliant on API calls for the source normalization technique. We attempt to minimize costs by using GPT-4o, but note that we can (i) achieve better performance using a larger, reasoning model (more cost) or (ii) switch to an open model. We will explore both techniques.

Ethics Statement

We note that our paper is foundational research and we are not tied to any direct applications. We do not foresee any potential risks with our work. We do not perform any annotations or human evaluation as we use the already existing FactBank dataset and ModaFact dataset.

References

- Anthropic. 2025. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.
- 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.
- Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. **Committed belief annotation and tagging**. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. **He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics**. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.
- Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. **Event detection and factuality assessment with non-expert supervision**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.
- Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. **MAVEN-FACT: A large-scale event factuality detection dataset**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11140–11158, Miami, Florida, USA. Association for Computational Linguistics.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Meta. 2024. Llama-3.3. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.
- Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. **MEANTIME, the NewsReader multilingual event and time corpus**. In *Proceedings of the Tenth International Conference*

- on *Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).
- John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. **Towards generative event factuality prediction**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715, Toronto, Canada. Association for Computational Linguistics.
- John Murzaku and Owen Rambow. 2024. **BeLeaf: Belief prediction as tree generation**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 97–106, Mexico City, Mexico. Association for Computational Linguistics.
- John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. **Re-examining FactBank: Predicting the author's presentation of factuality**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- OpenAI. 2024a. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2024b. o1. <https://openai.com/o1/>.
- OpenAI. 2025. o3-mini. <https://openai.com/index/openai-o3-mini/>.
- OpenRouter. 2025. Openrouter api. <https://openrouter.ai/>.
- Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Edoardo Barba, Simone Conia, Sergio Orlandini, Giuseppe Fiameni, Roberto Navigli, et al. 2024. **Minerva llms: The first family of large language models trained from scratch on italian data**. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*.
- Amir Poursan Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. **Graph based neural networks for event factuality prediction using syntactic and semantic structures**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. **A new dataset and evaluation for belief/factuality**. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.
- Alexis Ross and Ellie Pavlick. 2019. **How well do NLI models capture verb veridicality?** 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 2230–2240, Hong Kong, China. Association for Computational Linguistics.
- Marco Rovera, Serena Cristoforetti, and Sara Tonelli. 2025. **ModaFact: Multi-paradigm evaluation for joint event modality and factuality detection**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6378–6396, Abu Dhabi, UAE. Association for Computational Linguistics.
- Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. **Neural models of factuality**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.
- Roser Saurí and James Pustejovsky. 2009. **Factbank: a corpus annotated with event factuality**. *Language resources and evaluation*, 43:227–268.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. **Chain-of-thought prompting elicits reasoning in large language models**. *Advances in neural information processing systems*, 35:24824–24837.
- Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. **Lexicosyntactic inference in neural models**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. **Transformers: State-of-the-art natural language processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. **mT5: A massively multilingual pre-trained text-to-text transformer**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Unified Experiments

Details For all **Unified** zero-shot experiments, we use a temperature of 0.0 where applicable (all models besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). To prompt all other models (LLaMA-3.3, DeepSeek-v3, DeepSeek-r1, and Claude-3.5-Sonnet), we use the OpenRouter API (OpenRouter, 2025). The open models are ran at full precision (henceforth why we used the OpenRouter API and external providers).

Prompt Our zero-shot **Unified** prompt is shown in Figure 1.

B Hybrid Experiments

Details For all **Hybrid** zero-shot experiments, we use a temperature of 0.0 where applicable (all models besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). To prompt all other models (LLaMA-3.3, DeepSeek-v3, DeepSeek-r1, and Claude-3.5-Sonnet), we use the OpenRouter API (OpenRouter, 2025).

Prompt Our **Hybrid** zero-shot prompt is shown in Figure 2.

C LLM Experiment Details

For all our FactBank experiments, we report a single run, especially due to cost. We note that o1 experiments cost up to \$75 per run on the FactBank test set. To minimize randomness, we set the temperature to 0.0 where applicable (besides o1 and o3-mini). For o1 and o3-mini, we use the default reasoning setting (Medium). For our ModaFact experiments, we report the average over all five folds. Due to API costs and performing five-fold cross validation, we limit all ModaFact experiments to GPT-4o, o3-mini, and DeepSeek r1, which are the most cost effective models.

D Dataset Details

FactBank We use the author and source-and-target projection of FactBank from Murzaku and Rambow (2024), who follow the article split from Murzaku et al. (2022). We use their provided code for data extraction and follow their exact article split. The release of the FactBank corpus that we use can be found at the Linguistic Data Consortium, catalog number LDC2009T23. The test set contains 280 sentences and 1,326 examples.

ModaFact We use all five-folds of the test set of the ModaFact corpus from Rovera et al. (2025), which is publicly available. All results we report are averages over the five folds. To get the events from ModaFact for our **Hybrid** zero-shot experiments, we use the author’s provided prediction files and inference script with mT5-XXL.

Fold	Sentences	Examples
Fold 1	646	2098
Fold 2	605	2097
Fold 3	606	2096
Fold 4	626	2094
Fold 5	601	2090

Table 4: Dataset details for the ModaFact test set.

E Source Normalization

We propose two source normalization prompts: a few shot source normalization prompt and an oracle source normalization prompt. For these prompts, we use GPT-4o, with temperature 0.0. We prompt GPT-4o using the OpenAI API. Our exact few shot source normalization prompt is shown in Figure 4. Our exact oracle normalization prompt is shown in Figure 3.

We perform an ablation analysis of our few-shot and oracle normalization techniques described in Section 4.4. We showcase these results for our top performing system (DeepSeek r1, **Hybrid**) in Table 5. Without any normalization, we achieve a Full F1 of 68.9% and Nest F1 of 17.5%. Our few shot normalization technique improves us 2.1% for Full F1, and more notably by 7.8% for Nest F1. Our oracle method, as expected (since we provide gold sources), performs even better than our few shot method, achieving a Full F1 of 72.7% and 27.1%. However, we choose to perform all experiments with our few shot normalization method instead of our oracle method to truly showcase LLMs capabilities for belief detection without any gold sources as input.

F Event Tagger

DeBERTa Tagger We use DeBERTa-large for token classification, setting the number of labels to 2 (O vs. EVENT). We use the following hyperparameters: Epochs: 5; Batch Size: 16; Learning Rate: 1e-4; Max Sequence Length: 128. We do not perform any hyperparameter optimization or tun-

Norm.	Full	Nest
None	68.9	17.5
Few Shot	72.0	25.3
Oracle	72.7	27.1

Table 5: Performance of DeepSeek r1 (Hybrid) under three source normalization settings. “None” denotes no normalization, “Few Shot” applies few-shot normalization, and “Oracle” uses ground-truth for normalization. Bold values indicate the best results for each test set.

Category	Count	Breakdown	Count
Source	123	Gold=AUTHOR	50
		Gold=“it”	13
FN	77	Missed Noun	38
		Missed Verb	30
Label	73	Pred:CT+ → Gold:UU	28
		Pred:PR+ → Gold:UU	22
FP	53	Predicted Noun	33
		Predicted Verb	10

Table 6: Error analysis for our **Hybrid** DeepSeek r1 system on nested predictions, showing counts of each error type relative to its category total count.

ing. The model is trained using the HuggingFace Transformers library (Wolf et al., 2020).

LLM Event Tagging We perform event tagging on multiple LLMs. We set the temperature to 0.0. We use the OpenRouter API (OpenRouter, 2025) for DeepSeek r1 and Claude 3.5, and the OpenAI API for GPT-4o. We do not perform experiments with o1 to avoid high costs. Our zero-shot event detection prompt is shown in Figure 5. Our few-shot event detection prompt is shown in Figure 6.

G Nested Error Analysis

We expand our error analysis on the nested sources. Table 6 shows our error counts and error types.

We specifically analyze the errors for nested beliefs, which is where all LLMs fail on (our top performing model achieving F1 of only 25.3%). We showcase the error category, and then the top two error types by count. We use the following labels: **Source** indicates a source mismatch error; **FN** indicates a false negative, where the LLM did not generate a certain event type; **Label** indicates a label error where the LLM had the source and event correct, but incorrectly labeled the event. **FP** indicates a false positive, where the LLM overpredicted (that is, it generated an event that was not

actually an event).

Our most notable error is **Source**, where 123 errors are made. The most common error is the model predicts the author is the source instead of the correct nested source. Another notable error is where the model does not classify the source as it, but rather predicts the name of the entity. Next, we see a repeating of similar errors that Murzaku et al. (2022) discovered. Specifically, event nouns can be hard to determine (e.g. nouns like “concerns”, “acquisition”, “construction”). Our **FN** and **FP** errors showcase that LLMs simultaneously overpredict event nouns, while also missing both event nouns and verbs. Finally, we notice two notable label flips for our **Label** error category: the LLM predicts CT+ (the event happened/is true) when the gold label is UU (unknown), and PR+ (possibly true) when the gold is UU. This is due to FactBank’s definitions of nested sources and future events: when a reporting of a future event happening (e.g. “Mary said it will happen”), the factuality of the event according to the source is UU (the source is not committing to the event; rather, the author is committing to it).

Our analysis emphasizes that despite our source normalization method and use of strong reasoning LLMs, there is much room for improvement. Our error analysis findings are further supported with similar errors that have been reported in previous works on FactBank (Murzaku et al., 2022).

H Code Release

We will release all of our code. We will provide the full pipelines, datasets, and model checkpoints where applicable.

Figure 1: **Instruction for our zero-shot Unified Belief Annotation.** The instruction for FactBank-style event factuality annotation consists of three parts: a brief task description, detailed step-by-step instructions, and the formatting structure. Our CoT instructions are shown in the end of the prompt (Step-by-Step Output).

You are an annotation assistant trained to process sentences according to a FactBank-style event factuality framework. Given a sentence (or short text), your task is to analyze and annotate events by:

- **Event Identification:** Finding and listing all event-denoting predicates (verbs, event nouns, state-denoting adjectives)
- **Source Analysis:** Identifying who is expressing or committing to each event
- **Factuality Assessment:** Determining how certain each source is about the events
- **Nested Attribution:** Managing multiple layers of reporting and belief
- **Special Cases:** Handling future events, negation, modality, and hedging

Follow these steps precisely for annotation:

STEP 1: Event Identification

- Find all event-denoting predicates in the text
- Each predicate must be a single token
- Include verbs, event nouns, and state-denoting adjectives

STEP 2: Source Identification

- Start with "AUTHOR" as the root source (text narrator)
- Identify source-introducing predicates (SIPs) like "said," "believed," "reported"
- For new sources (e.g., "Apple officials"), normalize to single-token labels (e.g., "officials")
- Format as "AUTHOR_<ShortLabel>" (e.g., "AUTHOR_officials")
- For nested sources, add additional levels with underscores (e.g., "AUTHOR_officials_spokesperson")
- Handle negated sources (e.g., "did not say") at the higher level

STEP 3: Factuality Labeling Assign one of these labels for each event-source pair:

- **true:** Certainly factual (e.g., "confirmed," "knew")
- **false:** Certainly counterfactual (e.g., "denied," "did not happen")
- **ptrue:** Probably true (e.g., "might," "could," "likely")
- **pfalse:** Probably false (e.g., "doubted")
- **unknown:** Non-committal or unspecified stance

Special Cases Guidelines

- Future/prospective events: Label as unknown unless probability indicated (then ptrue)
- Negative statements: Use false for explicit denials
- Modality/hedging: Use ptrue for "might," "could," "suspected"
- Uncommitted author: Use unknown for purely reported events

Your annotation should be formatted as a JSON-style list of dictionaries:

```
[
  {
    "source": "<source_label>", // e.g., "AUTHOR" or "AUTHOR_<source>"
    "event": "<event_token>", // exact predicate from text
    "label": "<factuality_value>" // true/false/ptrue/pfalse/unknown
  },
  ...
]
```

Step-by-Step Output Process:

- Walk through each event in the sentence
- Identify and explain all sources and their nesting
- Justify each factuality label from each source's viewpoint
- Produce the final JSON-style output

Figure 2: **Instruction for our Hybrid Belief Annotation.** The instruction for FactBank-style event factuality annotation consists of three parts: a brief task description, detailed step-by-step instructions, and the formatting structure. Our CoT instructions are shown in the end of the prompt (Step-by-Step Output).

You are an annotation assistant trained to process sentences according to a FactBank-style event factuality framework. Given a sentence (or short text) and a list of event predicates marked in that sentence, your task is to analyze and annotate events by:

- **Source Analysis:** Identifying who is expressing or committing to each event
- **Factuality Assessment:** Determining how certain each source is about the events
- **Nested Attribution:** Managing multiple layers of reporting and belief

Follow these steps precisely for annotation:

STEP 1: Source Identification

- Start with "AUTHOR" as the root source (text narrator)
- Identify source-introducing predicates (SIPs) like "said," "believed," "reported," "estimated," "argued"
- For new sources (e.g., "Apple officials"), normalize to single-token labels (e.g., "officials")
- Format as "AUTHOR_<ShortLabel>" (e.g., "AUTHOR_officials")
- For nested sources, add additional levels with underscores (e.g., "AUTHOR_officials_spokesperson")
- Handle negated sources (e.g., "did not say") at the higher level

STEP 2: Factuality Labeling Assign one of these labels for each event-source pair:

- **true:** Certainly factual (e.g., "confirmed," "knew")
- **false:** Certainly counterfactual (e.g., "denied," "did not happen")
- **ptrue:** Probably true (e.g., "might," "could," "likely")
- **pfalse:** Probably false (e.g., "doubted")
- **unknown:** Non-committal or unspecified stance

Special Cases Guidelines

- Future/prospective events: Label as unknown unless probability indicated (then ptrue)
- Negative statements: Use false for explicit denials
- Modality/hedging: Use ptrue for "might," "could," "suspected"
- Uncommitted author: Use unknown for purely reported events

Your annotation should be formatted as a JSON-style list of dictionaries:

```
[
  {
    "source": "<source_label>", // e.g., "AUTHOR" or "AUTHOR_<source>"
    "event": "<event_token>", // exact predicate from text
    "label": "<factuality_value>" // true/false/ptrue/pfalse/unknown
  },
  ...
]
```

Step-by-Step Output Process:

- Walk through each event in the sentence
- Identify and explain all sources and their nesting
- Justify each factuality label from each source's viewpoint
- Produce the final JSON-style output

Figure 3: **Oracle Source Normalization Prompt**

You are determining if two source names refer to the same entity. Consider company abbreviations, common variations, and parent/subsidiary relationships. Also consider the context of the sentence and entity coreference.

Answer only YES if these definitely refer to the same entity, NO if they are different or if you're unsure. Include a brief explanation of your reasoning.

Sentence: {Sentence}
 Predicted Source: {Predicted Source}
 Gold Source: {Gold Source}

Figure 4: Few Shot Source Normalization Prompt

You are a FactBank-style source normalization assistant.

Your task: **Identify and normalize the subject (speaker/thinker/etc.) of each source-introducing predicate (SIP)** in a sentence. The normalized form must be a short, single-token label following “AUTHOR_”. If nested sources appear (i.e., one speaker quotes another speaker), nest them by appending an underscore plus the new label.

Use these **rules and guidelines**:

1. Source-Introducing Predicates (SIPs):

- Common SIP verbs: “said,” “reported,” “believed,” “estimated,” “argued,” “announced,” “denied,” “claimed,” etc.
- If an entity is repeated (the same subject for multiple SIPs), reuse the same label.

2. Normalization:

- Reduce corporate entities to “Corp.” or “Inc.” instead of the full name. E.g.:
 - “Marathon Widget Corp.” → **AUTHOR_Corp.**
 - “Skyline Media Inc.” → **AUTHOR_Inc.**
- If it’s just “the company,” consider normalizing to **AUTHOR_company** *only* if no more specific corporate form (like “Corp.”) is available.
- For people:
 - “He,” “she,” “they” → **AUTHOR_he**, **AUTHOR_she**, **AUTHOR_they**
 - “Mr. Alvarez,” “Ms. Hurt,” or “Dr. Kim” → **AUTHOR_Alvarez**, **AUTHOR_Hurt**, **AUTHOR_Kim**
 - If the sentence says “I stated...” → **AUTHOR_I**
- For large institutions:
 - “Ministry of Defense” → **AUTHOR_ministry**
 - “Police Department” → **AUTHOR_police**
 - “officials” → **AUTHOR_officials**
 - “board” → **AUTHOR_board**
- If you have a nested quote, e.g., “AUTHOR_officials_spokesperson” if the spokesperson is quoting officials.

3. Polarity:

- Even if the SIP is negated, you still label that source. (e.g., “he denied...” is valid.)

4. Output:

- Output **only** the normalized label(s). If no new source is introduced, or if you’re uncertain, you can leave the text unchanged or indicate “No SIP found.”

Few-Shot Examples (*truncated for space; we use 10 few shot examples*)

- Sentence:** Alpha Widget Corp. said it is launching a new product line.
Predicted: AUTHOR_Alpha_Widget_Corp.
Corrected: AUTHOR_Corp.
- Sentence:** “I believe the results speak for themselves,” he announced.
Predicted: AUTHOR_he
Corrected: AUTHOR_I
(*Because “I” is the direct speaker—if the text clearly attributes the quote to the first person.*)
- Sentence:** In its quarterly filing, LRS Acquisition stated it expects higher revenue.
Predicted: AUTHOR_LRS
Corrected: AUTHOR_Acquisition
- Sentence:** A portfolio unit of Greenbank Corp. reported continued growth this year.
Predicted: AUTHOR_portfolio unit
Corrected: AUTHOR_unit
- Sentence:** The foreign minister declared that cooperation would improve global stability.
Predicted: AUTHOR_foreign minister
Corrected: AUTHOR_minister

Return:

- Return the final normalized label(s) if a new source arises from the SIP.
- If none or unclear, output “No SIP found” or leave the text as is.

Figure 5: FactBank Single-Token Event Identification Prompt

You are an expert at identifying single-token events in text following FactBank guidelines.
Find ALL single-token predicates that:

Criteria:

1. **Are ONLY ONE of:**

- Reporting verbs (communication)
- Cognitive verbs (mental states)
- Action verbs (physical/abstract actions)
- Event nouns (occurrences/happenings)
- State adjectives (temporary states)

2. **Must represent:**

- Something that happened/happens/will happen
- Something that can be assessed as true or false
- Something with a temporal dimension

3. **Critical Distinction for Nouns/Nominalizations:**

- **INCLUDE** only nouns that refer to specific instances of events with:
 - Concrete temporal bounds
 - Specific participants
 - Ability to be assessed as having occurred or not
- **DO NOT include** nouns that refer to:
 - General concepts or types of events
 - Abstract categories
 - Topics or subjects of discussion
 - Generic processes
 - Institutional practices

Key rules:

- Extract SINGLE tokens only
- Include all verbs from source-introducing predicates
- Include nested events
- Include events under modals or negation
- Include events in complement clauses

Do NOT include:

- Multi-word phrases
- Generic nouns
- Auxiliary verbs
- Articles, prepositions, or conjunctions
- References to event types without specific instances

Output Format:

Your output is a JSON-style list of dictionaries. Each dictionary has:

- "event": The exact event token or predicate from the sentence.

Output Example:

```
[  
  {"event": "EVENT1"},  
  {"event": "EVENT2"},  
  {"event": "EVENT3"}  
]
```


Figure 6: FactBank Few-Shot Single-Token Event Identification Prompt.

You are an expert at identifying single-token events in text following FactBank guidelines.
Find ALL single-token predicates that:

1. **Are ONLY ONE of:**

- Reporting verbs (communication)
- Cognitive verbs (mental states)
- Action verbs (physical/abstract actions)
- Event nouns (occurrences/happenings)
- State adjectives (temporary states)

2. **Must represent:**

- Something that happened/happens/will happen
- Something that can be assessed as true or false
- Something with a temporal dimension

3. **Critical Distinction for Nouns/Nominalizations:**

- **INCLUDE** only nouns that refer to specific instances of events with:
 - Concrete temporal bounds
 - Specific participants
 - Ability to be assessed as having occurred or not
- **DO NOT include** nouns that refer to:
 - General concepts or types of events
 - Abstract categories
 - Topics or subjects of discussion
 - Generic processes
 - Institutional practices

Key rules:

- Extract SINGLE tokens only
- Include all verbs from source-introducing predicates
- Include nested events
- Include events under modals or negation
- Include events in complement clauses

Do NOT include:

- Multi-word phrases
- Generic nouns
- Auxiliary verbs
- Articles, prepositions, or conjunctions
- References to event types without specific instances

Output Format:

Your output is a JSON-style list of dictionaries. Each dictionary has:

- "event": The exact event token or predicate from the sentence.

Examples: (we truncate the examples omit 3 examples here for brevity)

1. **Sentence:** In composite trading Friday on the New York Stock Exchange, BellSouth shares fell 87.5 cents.

Output:

```
[
  {"event": "trading"},
  {"event": "fell"},
]
```

2. **Sentence:** Many local residents denounced the bigotry.

Output:

```
[
  {"event": "denounced"},
  {"event": "bigotry"}
]
```

Probing the Limits of Multilingual Language Understanding: Low-Resource Language Proverbs as LLM Benchmark for AI Wisdom

Surendrabikram Thapa^{1,*}, Kritesh Rauniyar^{2,3,*}, Hariram Veeramani⁴,
Surabhi Adhikari⁵, Imran Razzak⁶, Usman Naseem⁷

¹Virginia Tech, USA, ²Delhi Technological University, India,

³IIMS College, Nepal, ⁴UCLA, USA, ⁵Columbia University, USA,

⁶Mohamed bin Zayed University of Artificial Intelligence, UAE,

⁷Macquarie University, Australia

Abstract

Understanding and interpreting culturally specific language remains a significant challenge for multilingual natural language processing (NLP) systems, particularly for less-resourced languages. To address this problem, this paper introduces PRONE¹, a novel dataset of 2,830 Nepali proverbs, and evaluates the performance of various language models (LMs) in two tasks: (i) identifying the correct meaning of a proverb from multiple choices, and (ii) categorizing proverbs into predefined thematic categories. The models, including both open-source and proprietary, were tested in zero-shot and few-shot settings with prompts in English and Nepali. While models like GPT-4o demonstrated promising results and achieved the highest performance among LMs, they still fall short of human-level accuracy in understanding and categorizing culturally nuanced content, highlighting the need for more inclusive NLP.

1 Introduction

Language is a powerful medium for conveying culture, traditions, and shared human experiences. Training language models (LMs) to learn multiple languages and contexts can significantly enhance their ability to understand diverse human perspectives and communicate across cultural boundaries (Li et al., 2024; Hu et al., 2020; Thapa et al., 2025). While this can enable more inclusive and globally aware AI systems, it also presents substantial challenges in accurately capturing the unique nuances, idioms, and culturally specific references that vary widely between languages and societies (Liu et al., 2025; Agarwal et al., 2025; Aleem et al., 2024; Tao et al., 2024; Myung et al., 2024; Pawar et al., 2025).

For instance, what is considered common knowledge in one culture may not hold the same rel-

evance in another. A phrase like 'watching the ball drop' immediately invokes the image of New Year's Eve in Times Square for those familiar with American culture. At the same time, it may mean nothing to someone from a different cultural background. Similarly, in Japan, a *Hanami* or 'flower-viewing party' carries deep cultural significance associated with cherry blossoms in spring. In contrast, it might simply be interpreted as a generic gathering in other parts of the world. Proverbs are another prime example of how deeply language is intertwined with culture. Unlike general phrases or idioms, proverbs frequently rely on metaphors, analogies, and references unique to their origin (Kordoni, 2018; Qiang et al., 2023; Verma and Vuppuluri, 2015; Abebe Fenta and Gebeyehu, 2023). They are not just linguistic expressions but also cultural artifacts that reflect the lived experiences and shared understanding of a community.

For example, the proverb 'शङ्कर सहायता गर्छन् त भयङ्करको के पिर' (If Shankar is helpful, then what is there to fear?), reflects a deeply rooted cultural belief in divine protection and faith. In Hinduism, Shankar (name for Lord Shiva) is revered as a powerful god, and the proverb suggests that if a divine force is on one's side, there is no need to worry about any dangers or challenges. For a language model unfamiliar with Hindu deities or the cultural context of Nepal, the significance of this proverb would likely be misunderstood or lost. For instance, the model might interpret 'Shankar' as a common proper name for a person rather than recognizing it as a reference to Lord Shiva. Thus, it is crucial to develop language models that are not only proficient in multiple languages but also attuned to the cultural contexts and nuances that shape the meaning of expressions, idioms, and proverbs. While recent advancements in multilingual NLP for cultural understanding have focused on major languages like Hindi, Chinese, and Span-

* These authors contributed equally to this work and are listed as joint first authors.

¹<https://github.com/therealthapa/prone>

ish (Hu et al., 2020; Kakwani et al., 2020; Baccells et al., 2025), there remains a considerable gap when it comes to less-resourced languages such as Nepali (Thapa et al., 2024; Rauniyar et al., 2023). To address this gap, we introduce **PRONE**, a novel dataset of 2,830 Nepali proverbs and evaluate the performance of large language models (LLMs) in interpreting and categorizing them accurately. Our contributions are:

- We introduce **PRONE**, a manually curated novel dataset of 2,830 **PRO**verbs in **NE**pali, reflecting diverse cultural expressions and wisdom unique to Nepali.
- We manually classify the proverbs into five broad categories, capturing key themes and contextual nuances.
- We benchmark the performance of LLMs on two specific tasks: **Task A**: Evaluating the ability of LLMs to correctly identify the meaning of a proverb from a set of options consisting of one correct and three incorrect choices. **Task B**: Assessing the capacity of LLMs to accurately categorize the proverbs into predefined categories.

By focusing on Nepali proverbs, our study supports the United Nations Sustainable Development Goal (SDG) of 'Leave No One Behind' by promoting linguistic inclusivity and cultural representation in AI.

2 Related Works

Prior research in figurative languages, such as metaphor detection, generation, and interpretation, has employed various approaches, including linguistic and visual embeddings, context-based analysis, and paraphrasing tasks, which are also relevant to understanding proverbs (Pramanick et al., 2018; Chakrabarty et al., 2021; Bizzoni and Lappin, 2018; Wachowiak and Gromann, 2023; Liu et al., 2022). Goren and Strapparava (2024) examine GPT-3.5's ability to detect word-level metaphors in proverbs using different prompting strategies. They expand the PROMETHEUS dataset (Özbal et al., 2016) with hypothetical contexts and test three prompting approaches. The results show that the model performs best with hypothetical context, followed by first providing the proverb's meaning. Similarly, there have been efforts to enhance language models' understanding

of cultural and linguistic nuances, such as the work by Wibowo et al. (2024), who developed COPAL-ID, a dataset tailored for commonsense reasoning in Indonesian. They experiment with different LLMs, including open-source models such as XLM-R (Conneau et al., 2020), BLOOMZ (Muennighoff et al., 2023b), and PolyLM (Wei et al., 2023), as well as proprietary models such as ChatGPT and GPT-4, to evaluate their ability to handle the cultural and linguistic nuances embedded in the COPAL-ID dataset. Their findings indicate that while proprietary models like GPT-4 achieve relatively higher accuracy, they still fail human-level performance in understanding local nuances.

Expanding on the theme of evaluating the understanding of language models of culturally nuanced language, Liu et al. (2024) investigated the abilities of various language models, such as BLOOMZ (Muennighoff et al., 2023b), LLaMA-2 (Touvron et al., 2023), XGLM (Lin et al., 2022), XLM-R (Conneau et al., 2020), and mT0 (Muennighoff et al., 2023a), in reasoning with proverbs and sayings across different cultures. They evaluated these models using culturally diverse proverbs in six languages (English, German, Russian, Bengali, Mandarin Chinese, and Indonesian). Their findings showed that while these models could memorize proverbs to some extent, they often struggled to understand them in conversational contexts, particularly when dealing with figurative language and cross-cultural translations. However, these studies have primarily focused on high-resource languages, leaving less-resourced languages like Nepali largely unexamined. Our work is the first to address this gap, introducing a novel dataset of 2,830 Nepali proverbs and evaluating the ability of LLMs to interpret and categorize them effectively.

3 Dataset

We created a dataset of 2,830 Nepali proverbs collected from various sources, including online databases, literature, and local cultural repositories. The primary collection relied on three subject matter experts (SMEs), each with at least a master's degree in fields related to Nepali language, literature, or culture. The collected proverbs were checked among the SMEs to filter out any proverbs that were deemed irrelevant.

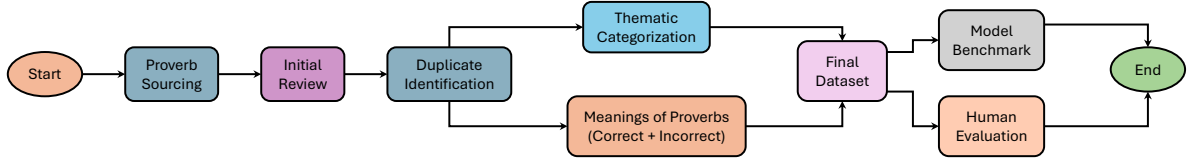


Figure 1: Overview of the End-to-End Pipeline for Annotating and Evaluating Nepali Proverbs

3.1 Deduplication

For deduplication, we used NepBERTa (Timilsina et al., 2022), a pre-trained language model, to generate embeddings, which capture the semantic meaning of each proverb. We then compute the cosine similarity between these embeddings to measure the similarity between pairs of proverbs. Using this approach, we identify semantically similar proverbs and treat them as near-duplicates. We manually visit the near-duplicates and remove if there are redundant proverbs.

3.2 Thematic Categorization

To categorize the proverbs, we manually annotated them into five categories: (i) Social Behavior and Relationships, (ii) Fate and Caution, (iii) Hard Work and Perseverance, (iv) Wisdom and Knowledge, and (v) Nature and Environment; using a rigorous annotation criterion (Appendix A). Each proverb was annotated by three annotators, and the final category was determined by majority agreement; in cases where all three annotators assigned different categories, the disagreement was resolved through a consensus Zoom meeting.

3.3 Final Dataset

For each proverb, as shown in Figure 1, we assigned one correct meaning and three plausible but incorrect meanings to test the interpretative capabilities of language models. The final dataset thus consists of proverbs categorized into five thematic groups (Table 1) and accompanied by multiple-choice options for their meanings.

Category	Proverbs
Social Behavior and Relationships	1274
Fate and Caution	1177
Hard Work and Perseverance	200
Wisdom and Knowledge	157
Nature and Environment	22
Total	2830

Table 1: Distribution of Nepali Proverbs.

4 Experimental Setup

4.1 Language Models

To evaluate the understanding and categorization of Nepali proverbs, we employed a range of language models, including open-source and proprietary ones. We conducted experiments in both zero-shot and few-shot settings for all models, prompting the models in both English and Nepali languages (prompts in Appendix B). The models evaluated included: **BERT-based LMs:** DistillBERT-Ne (Shrestha, 2023), RoBERTa-Ne (Chaudhary, 2023), NepBERTa (Timilsina et al., 2022), NepaliBERT (Ghimire, 2023), NepNewsBERT (Pudasaini, 2023). **Closed/ Proprietary Models:** GPT-3.5, GPT-4, GPT-4o (OpenAI, 2023), Gemini Pro 1.5, Gemini Flash 1.5, Mistral Medium (Mistral AI, 2024). **Open-sourced Models:** LLaMA-2 (7B) (Touvron et al., 2023), Mistral (7B) (Jiang et al., 2023), Gemma (7B) (Mesnard et al., 2024).

4.2 Evaluation Metrics

For Task A, we used accuracy to measure the proportion of correct selections by the LLMs from a set of options (one correct and three incorrect) as it directly reflects the models' ability to identify the correct meaning. Similarly, for Task B, we employed the F-score to evaluate the models' performance in categorizing proverbs into predefined categories, as it balances precision and recall, addressing the imbalanced distribution of categories.

5 Results and Discussion

Table 2 shows the performance of language models in Task A. Among all the models used, GPT-4o consistently performs the best across all the settings. The results show that across all model types, performance in the few-shot (FS) setting is consistently higher than in the zero-shot (ZS) setting, reflecting the benefit of additional context or

	Model	ZS-En	ZS-Ne	FS-En	FS-Ne
BERT Based LMs	DistillBERT-Ne	33.72	26.19	40.56	43.16
	RoBERTa-Ne	35.97	28.45	43.19	44.05
	NepaliBERT	36.41	29.71	45.87	45.87
	NepBERTa	38.67	32.84	45.34	47.64
	NepNewsBERT	40.41	35.76	50.23	50.18
Open Source	LLaMA-2	57.33	48.91	65.20	62.46
	Mistral	58.87	46.42	64.90	61.52
	Gemma	60.19	52.74	68.95	66.36
Closed/ Proprietary	GPT-3.5	14.88	15.62	22.14	25.09
	GPT-4	68.57	59.58	76.54	79.65
	GPT-4o	80.92	74.63	86.19	87.99
	Gemini Pro 1.5	79.93	75.02	83.72	83.75
	Gemini Flash 1.5	66.25	62.93	83.72	85.12
	Mistral Medium	17.14	53.57	24.39	50.99
Human Annotator	95.17				

Table 2: Accuracy of Different Language Models on Task A (Proverb Meaning Identification) Across Zero-Shot (ZS) and Fine-Tuned (FS) Settings in English (En) and Nepali (Ne).

examples. For example, the accuracy of BERT-based models improves from 26.19%-40.41% in the zero-shot setting to 43.16%-50.23% in the few-shot setting. Open-source models also show notable improvements with fine-tuning, where accuracy increases from 46.42%-60.19% in zero-shot to 61.52%-68.95% in few-shot. Similarly, closed/proprietary models such as GPT-4 and GPT-4o achieve much higher accuracy in few-shot settings, with GPT-4o reaching 87.99% compared to 74.63% in the zero-shot setting.

Table 3 presents the performance of various language models on Task B. The results indicate varying levels of performance across models and settings. BERT-based models show modest F-scores, ranging from 21.05% (DistillBERT-Ne, ZS-Ne) to 40.37% (NepNewsBERT, FS-En), with a slight improvement observed in the few-shot setting compared to zero-shot. Open-source models demonstrate moderate performance, with F-scores ranging from 35.74% (Mistral, ZS-Ne) to 49.87% (Gemma, ZS-En), indicating some capacity to handle the proverb categorization task. However, they do not reach the highest scores. Closed/proprietary models exhibit a wider range of F-scores, from as low as 6.68% (Gemini Pro 1.5, ZS-Ne) to as high as 84.52% (GPT-4o, ZS-En). Among these, GPT-4o consistently achieves the best performance, with the highest F-scores across all settings, particularly in the zero-shot English setting (84.52%) and the few-shot English setting (74.66%).

	Model	ZS-En	ZS-Ne	FS-En	FS-Ne
BERT Based LMs	DistillBERT-Ne	31.87	21.05	32.05	34.09
	RoBERTa-Ne	33.86	23.74	32.98	35.22
	NepaliBERT	35.91	24.56	36.71	36.73
	NepBERTa	36.24	26.73	38.56	39.67
	NepNewsBERT	38.17	28.42	40.37	40.02
Open Source	LLaMA-2	46.88	37.76	45.37	41.59
	Mistral	44.61	35.74	44.23	40.38
	Gemma	49.87	39.04	45.37	42.42
Closed/ Proprietary	GPT-3.5	26.02	9.18	42.73	31.23
	GPT-4	50.95	43.90	49.53	47.29
	GPT-4o	84.52	53.22	74.66	63.90
	Gemini Pro 1.5	47.40	6.68	31.77	20.50
	Gemini Flash 1.5	58.18	12.69	55.67	52.93
	Mistral Medium	28.67	11.57	23.50	34.51
Human Annotator	88.74				

Table 3: Performance (F-score) of Various Language Models on Task B (Proverb Categorization) in Zero-Shot (ZS) and Few-Shot (FS) Settings for English (En) and Nepali (Ne).

5.1 Human Evaluation

We also performed a human evaluation on both Tasks A and B to compare the performance of LLMs against human understanding. We employed a different set of three native Nepali speakers as annotators, each with at least a school-level education in Nepali. In Task A, human annotators achieved an accuracy of 95.17%, while in Task B, they obtained an F-score of 88.74%. These high-performance metrics indicate that these tasks are relatively straightforward for native speakers.

6 Conclusion

We evaluated various LLMs' abilities to understand and categorize Nepali proverbs using a novel dataset of 2,830 proverbs. While some models, such as GPT-4o, showed promising results, their performance still lags behind human annotators, who achieved the highest F-score of 95.17% and 88.74% in task A and task B, respectively. The gap highlights the need for further improvement in handling culturally specific content, particularly for less-resourced languages like Nepali. Future research should enhance models' understanding of diverse linguistic contexts to achieve more culturally inclusive NLP systems.

Limitations

While our study offers valuable insights into the performance of language models on the PRONE dataset, several limitations must be addressed.

First, the dataset, though substantial with 2,830 Nepali proverbs, may not encompass the full spectrum of cultural and contextual nuances inherent in Nepali language use. The limited scope of proverbs may restrict the models' ability to generalize across a broader range of culturally specific expressions. Second, despite promising results from models like GPT-4o, a noticeable gap remains compared to human annotators, highlighting challenges in achieving full cultural comprehension and accurate categorization. This indicates a need for enhanced training methods, possibly involving more diverse cultural data and improved model adaptation techniques. Additionally, our evaluation using English and Nepali prompts in zero-shot and few-shot settings may not fully capture the models' potential in varied real-world applications. Future work should explore alternative approaches, such as fine-tuning culturally rich datasets and developing hybrid models, to improve understanding and performance in less-resourced languages.

Ethics Statement

Data Collection and Privacy: The PRONE dataset of Nepali proverbs was created using publicly available sources, ensuring no personal or sensitive data was involved. We complied with all relevant data protection guidelines and model usage terms, focusing solely on non-commercial research. While the dataset aims to advance culturally inclusive NLP, we acknowledge the potential for biases in model outputs and caution against misuse that could reinforce cultural stereotypes. Comprehensive documentation is provided, but researchers should be aware of the dataset's limitations and apply it responsibly in diverse contexts.

Annotators Recruitment: The human annotators for this study were recruited at the local prevailing rate, ensuring fair compensation for their contributions. We adhered to ethical recruitment practices, and there were no ethical issues identified in this process. The annotators' native proficiency and cultural understanding were essential to the study, enhancing the quality and accuracy of the evaluations conducted.

References

Anduamlak Abebe Fenta and Seffi Gebeyehu. 2023. Automatic idiom identification model for amharic

language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1--9.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1--21.

Mahwish Aleem, Imama Zahoor, and Mustafa Naseem. 2024. Towards culturally adaptive large language models in mental health: Using chatgpt as a case study. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 240--247.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. [IberoBench: A benchmark for LLM evaluation in Iberian languages](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491--10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the workshop on figurative language processing*, pages 45--55.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. [MERMAID: Metaphor generation with symbolism and discriminative decoding](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250--4261, Online. Association for Computational Linguistics.

Amit Chaudhary. 2023. [roberta-base-ne: A roberta-based language model for nepali](#). Accessed: 2024-09-16.

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.

Rajan Ghimire. 2023. [Nepalibert: A bert-based language model for nepali](#). Accessed: 2024-09-16.

Gamze Goren and Carlo Strapparava. 2024. Context matters: Enhancing metaphor recognition in proverbs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3825--3830.

- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411--4421. PMLR.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlp suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948--4961.
- Valia Kordoni. 2018. Beyond multiword expressions: Processing idioms and metaphors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 15-16.
- Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Nanaman Goyal, Shrutvi Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019--9052.
- Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016--2039.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652--689.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437--4452.
- Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Mistral AI. 2024. [Mistral ai models: Getting started guide](#). Accessed: 2024-09-16.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023a. [Crosslingual generalization through multitask finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991--16111, Toronto, Canada. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023b. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991--16111.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunso Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104--78146.
- OpenAI. 2023. Gpt (generative pre-trained transformer). <https://openai.com/research/gpt>. Accessed: 2024-09-16.
- Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787--3793.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1--96.
- Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An lstm-crf based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67--75.
- Shushant Pudasaini. 2023. [Nepnewsbert: A bert-based language model for nepali news classification](#). Accessed: 2024-09-16.

- Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740--754.
- Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092--143115.
- Dipesh Shrestha. 2023. [Nepali-distilbert: A distilbert-based language model for nepali](#). Accessed: 2024-09-16.
- 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.
- Surendrabikram Thapa, Kritesh Rauniyar, Ehsan Barkhordar, Hariram Veeramani, and Usman Naseem. 2024. Which side are you on? investigating politico-economic bias in nepali language models. In *Annual Workshop of the Australasian Language Technology Association (22nd: 2024)*, pages 104--117. Association for Computational Linguistics (ACL).
- Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHPSAL 2025)*, pages 71--82.
- Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd conference of the Asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing*. Association for Computational Linguistics (ACL).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the international conference recent advances in natural language processing*, pages 681--687.
- Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018--1032.
- Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.
- Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. Copal-id: Indonesian language reasoning with local culture and nuances. 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 1404--1422.

A Annotation Details

To categorize the proverbs, we followed the following definitions and criteria:

1. Social Behavior and Relationships:

Proverbs that focus on human interactions, social conduct, community norms, and interpersonal relationships. This category includes proverbs that highlight themes such as trust, deceit, friendship, love, and societal roles.

Criteria: The proverb must relate to behaviors, expectations, or dynamics between individuals or groups within a social context.

Examples: "जस्तो कुकुर, उस्तै पुच्छर।" ('Like the dog, like its tail') — emphasizes consistent behavior or traits, and "छिमेकीको घरमा आगो लाग्दा आफ्नो घर सुरक्षित छैन।" ('When a neighbor's house is on fire, your own house is not safe') — reflects interdependence in community relations.

2. Fate and Caution:

Proverbs that deal with themes of destiny, luck, and the importance of caution or mindfulness in life. This category encompasses advice on being prudent, prepared, or aware of one's circumstances and external forces.

Criteria: The proverb must convey messages related to fate, destiny, or the necessity of being cautious or aware in various situations.

Examples: "चोक्टा पर्छु, ढुंगामा लाग्छु।" ('If I fall, I fall on a rock') — suggests the inevitability of misfortune, and "सर्पसँग दोस्ती गर्नु खतरा हुन्छ।" ('Friendship with a snake is dangerous') — emphasizes the need for caution in relationships.

Proverb	True Meaning	Incorrect Meaning	Note
श्राद्ध गर्न त सजिलो तर सिधा पुर्याउन गाह्रो	ठूलो कामभन्दा सानो काम गर्न गाह्रो हुन्छ । (It is easier to do big things when smaller things are in place.)	श्राद्ध गर्दा धेरै मानिसहरुलाई बोलाउनुपर्छ (You have to call a lot of people in funeral).	In Nepali culture, श्राद्ध (Shraddha) refers to an important ritual to honor deceased ancestors, where the ceremony may seem straightforward, but strict attention to every detail is vital and often more challenging.
कामकुरो एकातिर, कुम्लो बोकी ठिमीतिर।	एकातिरको गर्नुपर्ने काम छाडी अर्कातिर लाग्ने चाला (Neglecting task that needs to be done and instead tending toward something else.)	मानिसहरु सधैं आफ्नो बोझ अरुलाई दिन्छन् (People always tend to give their burden to others.)	Thimi is a town in Nepal, and in this context, it represents a place that is irrelevant to the original task or goal.
मियाँको मस्जिदसम्म	व्यक्तिको प्रयास वा महत्वाकांक्षा सीमित दायरामा मात्र सीमित छ। (Person's efforts or ambitions are limited to a narrow range or familiar routine)	धार्मिक कर्तव्य पूरा गर्न दौडधुप गर्ने (To actively run around to fulfill the religious duties)	The use of मियाँ (Miyān), a respectful term for a Muslim man, and मस्जिद (mosque), denotes a routine or habitual practice.
'शङ्कर सहायता गर्छन् त भयङ्करको के पिर'	'यदि भगवान् साथ दिन्छन् भने डरको कुनै कारण छैन।' (If Lord is on your side, there is no reason to fear.)	यदि साथीले मद्दत गर्छ भने डराउनु पर्दैन। (If a friend helps, there is no need to be afraid.)	"शङ्कर" (Shankar) can be both a common proper name for a person and a reference to Lord Shiva in Hinduism; thus, without context, the proverb could be mistakenly interpreted as referring to an ordinary person's help rather than invoking divine protection.

Table 4: Examples of Nepali Proverbs with Their True and Incorrect Meanings, Along with Notes on Potential Misinterpretations

3. **Hard Work and Perseverance:** Proverbs that highlight the value of diligence, effort, endurance, and resilience in overcoming challenges or achieving goals. These proverbs often carry motivational or inspirational messages.

Criteria: The proverb should focus on themes of hard work, persistence, or the rewards of sustained effort and commitment.

Examples: "हलो जोते मुरी फल्छ, घाँस खाए गोरु मर्छ।" (Plowing the field yields a harvest; eating the grass kills the ox) — underscores the benefits of hard work, and "धेरै मेहनत नगरी कुनै चीज प्राप्त हुँदैन।" (Without much effort, nothing is obtained) — stresses the necessity of perseverance.

4. **Wisdom and Knowledge:** Proverbs that offer guidance or insights about life, learning, and understanding. These proverbs often reflect collective wisdom, experience, or philosophical reflections on human behavior or morality.

Criteria: The proverb should convey a lesson or insight related to knowledge, learning, or the deeper understanding of life.

Examples: "ज्ञान नै शक्ति हो।" (Knowledge is power) — emphasizes the importance of wisdom, and "साँढेको आँसु दूध हुँदैन।" (A bull's

tears are not milk') — encourages recognizing reality and not being swayed by appearances.

5. **Nature and Environment:** Proverbs that use elements of nature (such as animals, plants, weather, or landscapes) to convey lessons or truths. These proverbs employ natural metaphors to illustrate human behavior, morality, or life lessons.

Criteria: The proverb must use imagery from the natural world to communicate its message or lesson.

Examples: "हावा खाएको जस्तै गर्छ, पानी परेको जस्तै भिजाउँछ।" (It moves like the wind, wets like the rain) uses elements of nature to describe inevitability or impact, and "ओखती जति तीतो हुन्छ, रोग त्यति नै राम्रो हुन्छ।" (The more bitter the medicine, the better the cure) draws on natural elements to illustrate a life lesson.

Using these definitions and criteria, we ensured that each proverb was categorized accurately, reflecting its central theme and underlying cultural context. This approach allowed us to create a well-defined dataset that can be effectively used to evaluate the performance of large language models in understanding culturally specific content.

B Prompt Templates

Example of Zero-shot Prompt in English for Meaning

Select the correct meaning for the given proverb among given options: Proverb: _____
Options: A. _____ B. _____ C. _____ D. _____. Only output the correct option as
'A', 'B', 'C' or 'D'. Explanations are not needed.

Example of Zero-shot Prompt in Nepali for Meaning

तल दिइएको उखानको सही उत्तर दिनुहोस् । कृपया 'A', 'B', 'C' वा 'D' मध्य सही विकल्पहरु मात्र उत्तर दिनुहोस् ।
व्याख्या नगर्नुहोस् । उखान: _____ विकल्पहरु: A. _____ B. _____ C. _____ D. _____.

Example of Few-shot Prompt in English for Meaning

Select the correct meaning for the given Nepali proverb among the given options. Only output the
correct option as 'A', 'B', 'C', or 'D'. Explanations are not needed. Example 1: Proverb: अवसर
चुकेपछि के काम, मौकामा नै काम गर्नुपर्छ। Options: A. अवसर चुकाउनु राम्रो हुन्छ, B. अवसर चुकाउनु भनेको
सफलताको संकेत हो, C. अवसर चुकाएपछि अर्को अवसर आउँदैन, D. मौकामा नै काम गर्नुपर्छ Correct Option:
D; Example 2: Proverb: 'भेडा भेडासँग, बाख्रा बाख्रासँग' Options: A. सबै प्राणी आफ्ना-आफ्ना जातिसँग
मिल्छन्।, B. भेडा र बाख्रा कहिल्यै सँगै बस्न सक्दैनन्, C. भेडा र बाख्राको सम्बन्ध झगडालु हुन्छ, D. भेडा र
बाख्राको बिचमा सधैं प्रतिस्पर्धा हुन्छ Correct Option: A; Example 3: Proverb: 'बाहिरका ठूला, भित्रका लुला'
Options: A. देख्नलाई मात्र निकै भएजस्ता तर मनका फितला ।, B. बाहिर राम्रो देखिने तर भित्र कमजोर, C. बाहिर
धनी तर भित्र गरिब, D. बाहिर बोल्न सक्ने तर भित्र डराउने, Correct Option: B; Example 4: Proverb: औषधी
र उपदेश मीठो हुन्छ । Options: A. जब तपाईं खराब अवस्थामा हुनुहुन्छ, प्रभावकारी सुझाव र समाधानहरु राम्रो
वा सजिलो नलाग्न सक्छ।, B. औषधी र उपदेश सधैं तीतो हुन्छ, C. औषधी र उपदेश सधैं बेकार हुन्छ, D. औषधी
र उपदेशले कहिल्यै काम गर्दैन Correct Option: A; Example 5: Proverb: पहिलो गाँसमै ढुङ्गा । Options: A.
नराम्रो शुरुवात हुनु ।, B. पहिलो गाँसमै ढुङ्गा भनेको खाना पकाउन नजान्नु हो, C. पहिलो गाँसमै ढुङ्गा भनेको सधैं
असफल हुनु हो, D. पहिलो गाँसमै ढुङ्गा भनेको ढुङ्गा खानु हो Correct Option: A; Now, select the correct
meaning for the given Nepali proverb. Proverb: _____ Options: A. _____ B. _____ C.
_____ D. _____.

Example of Few-shot Prompt in Nepali for Meaning

दिइएको नेपाली उखानको तल दिइएको विकल्पहरु मध्य सही अर्थ भएको विकल्प छान्नुहोस् । कृपया 'A', 'B', 'C'
वा 'D' मध्य सही विकल्पहरु मात्र उत्तर दिनुहोस् । व्याख्या नगर्नुहोस् । उदाहरण १: उखान: अवसर चुकेपछि के
काम, मौकामा नै काम गर्नुपर्छ । विकल्पहरु: A. अवसर चुकाउनु राम्रो हुन्छ, B. अवसर चुकाउनु भनेको सफलताको
संकेत हो, C. अवसर चुकाएपछि अर्को अवसर आउँदैन, D. मौकामा नै काम गर्नुपर्छ सही विकल्प: D; उदाहरण
२: उखान: 'भेडा भेडासँग, बाख्रा बाख्रासँग' विकल्पहरु: A. सबै प्राणी आफ्ना-आफ्ना जातिसँग मिल्छन्।, B. भेडा
र बाख्रा कहिल्यै सँगै बस्न सक्दैनन्, C. भेडा र बाख्राको सम्बन्ध झगडालु हुन्छ, D. भेडा र बाख्राको बिचमा सधैं
प्रतिस्पर्धा हुन्छ सही विकल्प: A; उदाहरण ३: उखान: 'बाहिरका ठूला, भित्रका लुला' विकल्पहरु: A. देख्नलाई
मात्र निकै भएजस्ता तर मनका फितला ।, B. बाहिर राम्रो देखिने तर भित्र कमजोर, C. बाहिर धनी तर भित्र गरिब,
D. बाहिर बोल्न सक्ने तर भित्र डराउने, सही विकल्प: B; उदाहरण ४: उखान: औषधी र उपदेश मीठो हुन्छ ।
विकल्पहरु: A. जब तपाईं खराब अवस्थामा हुनुहुन्छ, प्रभावकारी सुझाव र समाधानहरु राम्रो वा सजिलो नलाग्न
सक्छ।, B. औषधी र उपदेश सधैं तीतो हुन्छ, C. औषधी र उपदेश सधैं बेकार हुन्छ, D. औषधी र उपदेशले कहिल्यै
काम गर्दैन सही विकल्प: A; उदाहरण ५: उखान: पहिलो गाँसमै ढुङ्गा । विकल्पहरु: A. नराम्रो शुरुवात हुनु ।, B.
पहिलो गाँसमै ढुङ्गा भनेको खाना पकाउन नजान्नु हो, C. पहिलो गाँसमै ढुङ्गा भनेको सधैं असफल हुनु हो, D. पहिलो
गाँसमै ढुङ्गा भनेको ढुङ्गा खानु हो सही विकल्प: A; अब दिइएको उखानको सही विकल्प उत्तर दिनुहोस् । उखान:
_____ विकल्पहरु: A. _____ B. _____ C. _____ D. _____.

Example of Zero-shot Prompt in English for Proverb Category

Classify the following Nepali proverb into one of the five categories: {'Wisdom and Knowledge', 'Hard Work and Perseverance', 'Social Behavior and Relationships', 'Nature and Environment', 'Fate and Caution'} Proverb: _____. Provide only the category name that best fits the meaning of the given proverb. No explanation is needed.

Example of Zero-shot Prompt in Nepali for Proverb Category

तल दिएको नेपाली उखानलाई पाँचमध्ये कुनै एक वर्गमा वर्गीकृत गर्नुहोस्: 'ज्ञान र बुद्धि', 'मेहनत र धैर्यता', 'सामाजिक व्यवहार र सम्बन्धहरू', 'प्रकृति र वातावरण', 'भाग्य र सावधानी'. उखान: _____ उखानको अर्थसँग सबैभन्दा राम्रोसँग मिल्ने वर्गको नाम मात्र प्रदान गर्नुहोस्। कुनै स्पष्टीकरण आवश्यक छैन।

Example of Few-shot Prompt in English for Proverb Category

Classify the following Nepali proverb into one of the five categories: 'Wisdom and Knowledge', 'Hard Work and Perseverance', 'Social Behavior and Relationships', 'Nature and Environment', 'Fate and Caution'. Provide only the category name that best fits the meaning of the given proverb. Example 1: Proverb: पाप धुरीबाट कराउँछ Category: Fate and Caution ; Example 2: Proverb: आधा गाग्रो छचल्किन्छ Category: Wisdom and Knowledge; Example 3: Proverb: बाहिरका ठूला, भित्रका लुला Category: Social Behavior and Relationships; Example 4: Proverb: अल्छे तिघ्रो, स्वादे जिब्रो Category: Hard Work and Perseverance Example 5: Proverb: वनको चरो वनैमा रमाउँछ Category: Nature and Environment Now, classify the following proverb: Proverb: _____

Example of Few-shot Prompt in Nepali for Proverb Category

तल दिएको नेपाली उखानलाई पाँचमध्ये कुनै एक वर्गमा वर्गीकृत गर्नुहोस्: 'ज्ञान र बुद्धि', 'मेहनत र धैर्यता', 'सामाजिक व्यवहार र सम्बन्धहरू', 'प्रकृति र वातावरण', 'भाग्य र सावधानी'. उखानको अर्थसँग सबैभन्दा राम्रोसँग मिल्ने वर्गको नाम मात्र प्रदान गर्नुहोस्। उदाहरण १: उखान: पाप धुरीबाट कराउँछ वर्ग: भाग्य र सावधानी उदाहरण २: उखान: आधा गाग्रो छचल्किन्छ वर्ग: ज्ञान र बुद्धि उदाहरण ३: उखान: बाहिरका ठूला, भित्रका लुला वर्ग: सामाजिक व्यवहार र सम्बन्धहरू उदाहरण ४: उखान: अल्छे तिघ्रो, स्वादे जिब्रो वर्ग: मेहनत र धैर्यता उदाहरण ५: उखान: वनको चरो वनैमा रमाउँछ वर्ग: प्रकृति र वातावरण। अब, तलको उखानलाई वर्गीकृत गर्नुहोस्: उखान: _____

Measuring Sexism in US Elections: A Comparative Analysis of X Discourse from 2020 to 2024

Anna Fuchs[♣] Elisa Noltenius[♣] Caroline Weinzierl[♣]
Bolei Ma^{♣,♡} Anna-Carolina Haensch^{♣,♡,◇}

♣LMU Munich ♡Munich Center for Machine Learning
◇University of Maryland, College Park

{anna.fuchs,elisa.noltenius,caroline.weinzierl}@campus.lmu.de, {bolei.ma,c.haensch}@lmu.de

Abstract

Sexism continues to influence political campaigns, affecting public perceptions of candidates in a variety of ways. This paper examines sexist content on the social media platform X during the 2020 and 2024 US election campaigns, focusing on both male and female candidates. Two approaches, single-step and two-step categorization, were employed to classify tweets into different sexism categories. By comparing these approaches against a human-annotated subsample, we found that the single-step approach outperformed the two-step approach. Our analysis further reveals that sexist content increased over time, particularly between the 2020 and 2024 elections, indicating that female candidates face a greater volume of sexist tweets compared to their male counterparts. Compared to human annotations, GPT-4 struggled with detecting sexism, reaching an accuracy of about 51%. Given both the low agreement among the human annotators and the obtained accuracy of the model, our study emphasizes the challenges in detecting complex social phenomena such as sexism.

Disclaimer: This paper contains content that can be offensive or upsetting.

1 Introduction

Sexism is defined as prejudice, stereotyping, or discrimination based on sex, typically against women (Oxford English Dictionary, 2023). Despite progress toward gender equality, it remains prevalent in many areas of society, from workplaces and education to the media, shaping perceptions and limiting opportunities for women. One area where sexism is particularly prominent is politics, where women are underrepresented and often unfairly judged compared to their male counterparts (Fox and Lawless, 2004; Lovenduski, 2014). Female candidates often face scrutiny over their leadership and competence abilities simply because of their gender. The media further intensifies these biases

by focusing on their looks and personal lives instead of their political views. As social media plays an important role in shaping voters' opinions, it reinforces existing gender biases and gender-based criticism, particularly affecting female politicians (Tromble and Koole, 2020).

Detecting sexism and understanding the intentions behind it are essential steps in overcoming deeply embedded gender norms and biases, especially in contexts where women seek leadership positions, such as presidential candidacy in politics. However, sexist comments do not always exhibit obvious negative emotions (Becker and Wright, 2011). Sexism can be subtle, often unnoticed, making it challenging to identify since it is embedded in cultural and societal norms (Swim and Cohen, 1997). Therefore, it becomes crucial to investigate these implicit forms of sexism and their impact on individuals and society.

The 2020 and 2024 US election cycles present a unique opportunity for researchers to examine whether gender continues to influence the chances of presidential candidates. The 2020 presidential election featured Joe Biden and Donald Trump as primary candidates. In contrast, the 2024 election was remarkable for a candidate switch during the campaign, as Joe Biden announced his resignation on 21 July, with Kamala Harris subsequently launching her campaign (CNN Politics, 2024).

Our paper analyzes X (formerly Twitter) data using tweets sampled from three time frames over two election periods (2020 and 2024). These periods represent two different candidate scenarios: male vs. male and female vs. male. The selected tweets were chosen based on election-specific keywords, from the data source publicized by previous research (Balasubramanian et al., 2024).

We use GPT 4.0 (OpenAI, 2024) to categorize these tweets, setting up two different approaches. The first approach directly classifies tweets into non-sexist and more granular sexist categories. The

second approach involves a two-step process: initially identifying tweets as either non-sexist or sexist, followed by categorizing the sexist tweets into finer-grained categories. A consistent set of prompts is applied to compare the two approaches, while a small subsample dataset with manually annotated data is used as a reference for evaluating GPT’s annotation capabilities.

Using this set-up, we address the following research questions (RQs):

- **RQ1:** How do two-step and single-step GPT-4-based categorization approaches compare for identifying and classifying sexist tweets?
- **RQ2:** Have sexist patterns and categories of sexist content in US election-related discourses on X changed over the three election time frames?

2 Literature Review

This review offers an overview of sexism in political discourse, discussing approaches for classifying sexist content, with a focus on methods that use generative AI, and prompting techniques for the automated detection of sexist language.

Sexism in Politics. Literature on sexism in politics often focuses on the gender-biased representation of female politicians and the undermining of women’s leadership roles, highlighting how such biases influence public opinion and election outcomes. Systematic marginalization and societal structures within political institutions contribute to underrepresentation and limited political participation of women (Lovenduski, 2014). Despite similar qualifications, women express less political ambition due to lack of encouragement to run for office and a lower self-perception of qualifications (Fox and Lawless, 2004). The 2016 US presidential election between Hillary Clinton and Donald Trump served as a crucial case for studying gender dynamics in politics. Research shows that sexism played a substantial role in Hillary Clinton’s defeat, as women candidates face challenges and unequal evaluations compared to their male counterparts (Knuckey, 2019).

Sexism also shaped voter favorability: men showed a much stronger preference for Trump than women, while attitudes toward Clinton were similar between genders (Glick, 2019; Ratliff et al., 2019). Political sexism, defined as the belief that men are better suited emotionally for politics than women, strongly predicted support for Trump, especially

among white voters (Bracic et al., 2019). Hostile sexism, defined as having negative views towards individuals who defy traditional gender stereotypes, emerged as a key factor benefiting Trump’s candidacy, while benevolent sexism, which is positive in tone but yet connotes inferiority to men, increased support for Clinton without affecting Trump (Glick and Fiske, 2001; Ratliff et al., 2019). From a broader point of view, Falk (2010) examines nine female political campaigns, uncovering how media portrayals often frame female candidates as unviable or incompetent. Analyzing political sexism in social media, particularly using X data, has already been addressed by Tromble and Koole (2020). This study reports no clear differences in the tone of messages directed at male and female politicians across three countries, including the US.

Sexism Classification. Lots of research on sexism in social media has focused primarily on detecting misogyny and hateful language directed at women (Guest et al., 2021; Pamungkas et al., 2020). However, sexism often operates in more nuanced ways. To capture its complexity, researchers have developed various classification frameworks. A common method is a two-step approach: first identifying sexist tweets, then categorizing them into more granular categories (Jiang et al., 2022; Plaza et al., 2023). These finer-grained categories can be defined from multiple perspectives. According to ambivalent sexism theory (Glick and Fiske, 2001), sexism can be hostile (overtly negative) or benevolent (seemingly positive but reinforcing inferiority). Other studies classify the degree to which sexism manifests itself - blatant, subtle, or covert (Swim et al., 2004). Studies also explored multi-label classification, with varying granularity. For instance, Rodríguez-Sánchez et al. (2021) define five categories, while Parikh et al. (2019) define 23. Some approaches incorporate cultural perspectives (Jiang et al., 2022), focus on specific forms of harassment (Sharifirad et al., 2018), or distinguish sexism by target (individual or generic) (Jiang et al., 2022) and intention (Plaza et al., 2023). These different classification frameworks highlight the complexity of sexism and the need for approaches to successfully recognize its various forms.

Large Language Models and Prompt Design. Generative AI is emerging as a powerful tool for annotation and is being extensively researched as a substitute for human-annotated data, due to the human annotation challenges associated with the lat-

ter (Kern et al., 2023). Studies comparing human-annotated data with annotations using the ChatGPT show promising results in detecting hateful, offensive, and toxic (HOT) language (Li et al., 2024), with high accuracy. However, some highlight the persisting presence of additional bias in LLM annotations, given different contextual variations (Das et al., 2024; Okpala and Cheng, 2025). Huang et al. (2023) emphasize that hate speech detection is subjective and context-dependent, yet ChatGPT performs well even with identifying implicit hate speech. Maximizing the performance of an LLM relies on using qualitative prompts. Few-shot prompting, where the model is asked to perform a task with a few examples, generally performs better than zero-shot prompting, where no examples are given (Brown et al., 2020). Prompt engineering strategies, such as those introduced by White et al. (2025), offer adaptable structures for better results. Despite advancements, the use of LLMs as an annotation tool for sexism-related data remains sparsely researched. Given the widespread and evolving nature of sexism on social media, particularly in political discourse, further research is essential.

3 Research Design and Methodology

In this section, we provide an overview of the data used for this analysis, how we define the sexism categories, and the methodology we apply to answer our research questions.

3.1 Data

The data for this analysis consists of three distinct periods from US presidential election cycles:

- Biden vs. Trump 2020 (12 - 20 July 2020)
- Biden vs. Trump 2024 (12 - 20 July 2024)
- Harris vs. Trump 2024 (22 - 30 July 2024)

We will refer to these time frames as **BT2020**, **BT2024**, and **HT2024**, respectively, throughout the remainder of this paper.

The time frames BT2020 and BT2024 allow for a year-on-year comparison, providing insights into shifts in sexism in political discourse over time. The HT2024 time frame additionally allows for an analysis of sexism across different candidate scenarios, as it includes not only an election with two male candidates (Biden vs. Trump) but also a race featuring a female vs. male candidates (Harris vs. Trump). Including the two male vs. male candidacies aims to provide a clearer understanding

of whether sexist content has increased over time alone while keeping the candidates constant.

To extract tweets for these three different time periods, we made use of two public GitHub repositories from previous research capturing discourse on \mathbb{X} related to the US presidential elections (Chen et al., 2022; Balasubramanian et al., 2024).

Filtering for Relevant Tweets. Political discourse on social media covers a range of topics. To limit the tweets to more relevance regarding sexism, we filtered the tweets for specific keywords. The keywords used were: she, her, woman, women, men, man, female, girl, girls, lady, feminism, feminist, gender, sex, sexism, and sexist. This allows us to pre-filter the tweets for relevancy.

Data Retrieval for BT2020. Chen et al. (2022) provide a publicly available repository containing tweets from January 2020 to June 2021. These tweets were extracted using 227 different keywords and account references. The repository consists of several .txt files, organized by year, month, date, and hour, with each .txt file containing multiple tweet IDs. The .txt files covering our selected time frame, BT2020, were merged together, to then randomly select a sample of tweet IDs. To retrieve the actual tweet content corresponding to the IDs, access to the \mathbb{X} API is required. For this analysis, we used the Basic version of the \mathbb{X} API v2 (X Developer Platform, 2025) and extracted tweet texts and the creation date using the tweepy package in Python (Roesslein, 2020). Our access period for the Basic version spanned from January 20 to February 23, 2025. The sample size for this time frame was set to 15,000 since the Basic \mathbb{X} API version allows retrieval of up to 15,000 tweets per month. Of the sampled tweet IDs for which requests were sent via the \mathbb{X} API, we ultimately obtained 6,316 tweets for this analysis. Several factors contributed to this reduction in available data.

First, we restricted our data set to English-language tweets, meaning that any non-English tweets were automatically excluded. Additionally, a noteworthy number of tweet IDs belong to already deleted tweets, making it impossible to retrieve their content. Furthermore, the retrieved tweets included both original tweets and retweets. Due to a limitation of the \mathbb{X} API and the tweepy package, the full text of the retweets cannot be retrieved. Instead, only a truncated version is available, making such data unsuitable for this analysis. Since the \mathbb{X} API registers each request - regardless of whether

the tweet text is available, deleted, truncated, or not in English - this leads to a considerably lower number of tweets retrieved than originally anticipated. The implications of these limitations are discussed further in [section 7](#). The tweets were categorized into two groups, according to the keywords mentioned in the previous paragraph. All tweets containing a keyword (172) were used and a sample was chosen from tweets not containing the keywords, resulting in 431 tweets. The reason for the difference in sampled tweets arises from the piecewise approach to the OpenAI limit (see [subsection 3.3](#)).

Data Retrieval for BT2024 and HT2024. For the two election time frames in 2024 (BT2024 and HT2024) the data used for this analysis was previously extracted by [Balasubramanian et al. \(2024\)](#) using 44 different keywords. The corresponding public GitHub repository contained tweets from May until July 2024 and provided multiple `.csv.gz` files consisting of tweets related to the US election and information such as the tweet ID, text, url, date, number of retweets, view count, etc. For our analysis, we kept the tweet ID, the text, and the date of the tweet. After selecting the tweets that correspond to our two time frames, that is, July 12-20 and July 22-30, 2024, the tweets were filtered into two groups, according to previously mentioned keywords with relevance to sexism (see [subsection 3.1](#)), one group with tweets containing the keywords and the other group without containing them. For the BT2024 time frame, 3,000 tweets were randomly sampled per group, resulting in 6,000 tweets together. For the HT2024 time frame, 1,000-2,000 tweets were sampled per group, resulting in 3,000 tweets together. As for BT2020, the number in final categorized tweets per group differ slightly due to the piecewise approach to the OpenAI limit.

The final data used for the analysis consisted of 8,870 tweets, whose statistics are shown in [Table 1](#).

	BT2020	BT2024	HT2024
Total Number	431	5,630	2,809
With keywords	172	2,788	930
Without keywords	259	2,842	1,879

Table 1: Statistics of the final dataset.

3.2 Sexism Categories

We classify sexism into distinct categories using definitions similar to those of other studies ([Glick,](#)

[2019; Jiang et al., 2022; Rodríguez-Sánchez et al., 2021; Sharifirad et al., 2018; Swim et al., 2004](#)).

The sexism categories were defined as follows:

- **Sexist:** Tweets that discriminate, demean, or reinforce stereotypes based on gender, including offensive language, objectification, slurs, or preserving harmful gender roles. Tweets that discuss the topic of sexism but not in a way that is offensive towards people of certain genders.
- **Non-Sexist:** Tweets unrelated to gender bias, respectful or inclusive in tone, and free of gender-based stereotypes or discrimination.

For the finer-grained categories, the following were chosen:

- **Covert and Subtle Sexism:** Tweets that show unequal treatment that is not overtly hostile but reinforces systemic inequality. Masking sexism as a positive sentiment, depicting women as incompetent or unsuited for specific roles.
- **Discrediting:** Tweets that undermine women’s competence, achievements, or worth without meaningful critique, often dismissing them outright or marginalize women from decision-making and public discussions.
- **Objectification and Sexual Harassment:** Tweets that reduce women to their physical appearance, treating them as objects of desire rather than individuals with agency or intellect. Tweets that use sexualized language to intimidate women in the political sphere.
- **Remarks - Awareness and Advocacy:** Remarks or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These kind of tweets often aim to expose, discuss, or address sexism constructively.
- **Stereotyping:** Tweets that enforce traditional gender roles or suggest that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs.

These finer-grained categories were chosen because they capture types of sexism that are particularly relevant within the context of political discourse. The category *Remarks - Awareness and advocacy* was specifically included to analyze whether informative discussions about sexism increase over time and whether a female presidential candidate leads to more public discussions, awareness, and potentially more positive narratives about sexism in politics.

For a more detailed overview of the categories, including the complete definitions and corresponding examples, refer to the prompts in [Appendix A](#).

3.3 Methods

Classification Approaches. To classify tweets into defined sexism categories, we compare two classification approaches using GPT-4. The first approach follows a **single-step categorization**, in which GPT-4 directly categorizes each tweet as either *Non-Sexist* or into one of the finer-grained categories. The second approach consists of a **two-step categorization process**: First, GPT-4 classifies the tweets as either *Sexist* or *Non-Sexist*; all tweets that were classified as *Sexist* are further categorized into finer-grained categories.

To compare the two-step and single-step GPT-4-based classification approaches, we begin by addressing RQ1. As metrics for overall comparison of the prompting approaches for RQ1, we used accuracy and Cohen’s Kappa index; for category-wise comparison, we used recall and precision.

Tweet examples illustrating the alignment and difference between single-step and two-step categorization are provided in the [Appendix B](#).

Human Annotation. For the comparison of the two classification approaches, a subsample data set of 300 tweets was selected and manually annotated. To ensure that the annotated tweets represent different cases, the selected 300 tweets are composed as follows: 25% of the tweets were labeled as *Sexist* but classified into different finer-grained categories by both approaches. 25% of the tweets were labeled as *Sexist* by one approach but *Non-Sexist* by the other approach. 40% of the tweets were classified as the same finer-grained *Sexist* category by both approaches. 10% of the tweets were labeled *Non-Sexist* by both approaches. This selection guarantees the representation of cases where the two approaches differed or aligned. The 300 tweets were then manually annotated by three annotators. First, two annotators independently annotated all 300 tweets. For these two annotators, the agreement on the 300 selected tweets, which included both the *Sexist* category (subdivided into the five fine-grained categories) and the *Non-Sexist* category, resulted in a Cohen’s Kappa score of 0.394. This score is generally considered minimal agreement ([McHugh, 2012](#)). Because of this low agreement, a third annotator reviewed the annotations. If both of the first two annotators assigned the same category to a tweet, that category was retained. In cases where their categorization differed, the third annotator reviewed the tweet and either chose the more appropriate category or accepted both if either

categories were deemed valid. This serves as the final human annotation, used for the analysis. The purpose of this annotation was to determine which of the GPT-4-based approaches better aligned with human judgment, used as the ground truth in this analysis.

4 Results

In this section, the results obtained from the annotated tweets are presented. First, the prediction quality of the different categorization approaches is evaluated by comparing them to the human annotations (**RQ1**). Then, the change in the frequency of the sexism categories over time is analyzed (**RQ2**).

In total, 8,870 tweets were annotated by both single-step and two-step categorization: 430 tweets for the time frame BT2020, 5,630 tweets for BT2024, and 2,809 tweets for HT2024.

4.1 Comparison of Single- and Two-Step Categorization

When comparing the human annotation with GPT-4 categorization, a tweet is counted as correctly annotated by GPT-4 if the given category corresponds to one of the final human-annotated categories. In the following results, we refer to the human annotation as the ground truth.

Metric	Single-Step	Two-Step
Accuracy	0.510	0.503
Confidence Interval	[0.452, 0.568]	[0.445, 0.561]
Cohen’s Kappa	0.416	0.380

Table 2: Classification metrics for single- and two-step categorization, taking human-annotated data as the ground truth. The square brackets show the confidence interval: [lower bound, upper bound].

In [Table 2](#), the classification metrics chosen to compare the categorization approaches are depicted. The GPT-4 predictions for the single-step categorization have an accuracy of 51.0%, which means that 51.0% of the tweets were assigned to the correct category. Two-step categorization attained a similar accuracy of 50.3%. The accuracy confidence intervals for both approaches overlap, meaning there is no statistical significant difference between the two approaches. Cohen’s Kappa lies at 0.416 for the single-step process, which is considered weak agreement, and at 0.380 for the two-step process, which is considered minimal agreement ([McHugh, 2012](#)).

To get a better impression of how well the two approaches categorize the tweets, it is useful to

additionally look at classification metrics per finer-grained sexism category. Figure 1 depicts two confusion matrices, one for each categorization method, showing the agreement (in %) between the GPT-4 categorization and the human annotations. Darker fields indicate higher percentages and, therefore, higher agreement, while lighter fields represent lower agreement.

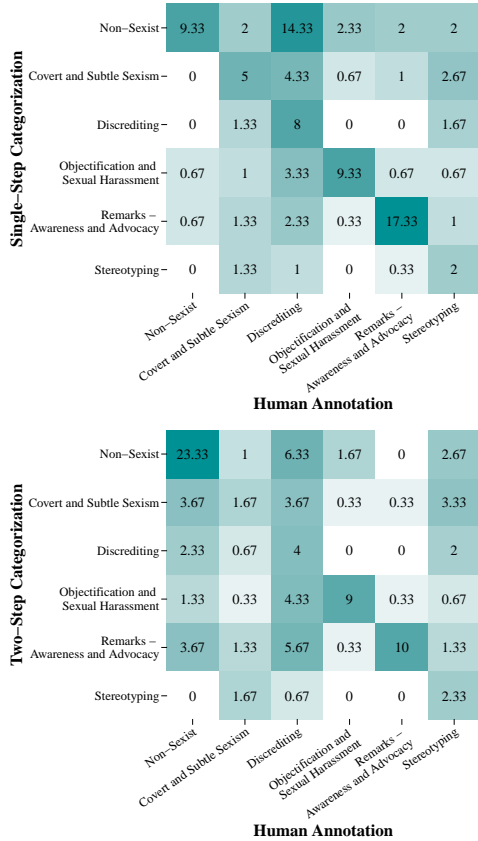


Figure 1: Confusion matrix of agreement between single-step (top) and two-step (bottom) categorization and human annotation

In the confusion matrices for single-step categorization (top) and two-step categorization (bottom), the off-diagonal elements for the single-step approach are slightly lighter, indicating fewer misclassifications. Additionally, the diagonal values for the single-step classification are mostly higher than those of their corresponding cells in the two-step matrix, suggesting that the single-step categorization achieves greater overall agreement with human annotations.

To further compare how the approaches performed, we looked at the precision and recall for each finer-grained category (see Table 3).

For single-step categorization, recall is highest for *Remarks - Awareness and Advocacy* (0.754), *Discrediting* (0.727), and *Objectification and Sex-*

ual Harassment (0.596). The remaining categories have a recall below 0.5. For the two-step categorization, the highest recall is for *Non-Sexist* (0.667), which in the single-step categorization has the lowest recall. The categories *Objectification and Sexual Harassment* and *Stereotyping* achieve a similar recall in the single-step and two-step categorization. However, all other categories have a recall below 0.5 in the two-step categorization, which is lower than for the single-step categorization.

The precision in single-step categorization is highest for *Non-Sexist* (0.875), followed by *Remarks - Awareness and Advocacy* (0.812) and *Objectification and Sexual Harassment* (0.737). The other categories have a precision below 0.5. In two-step categorization, precision is higher for *Remarks - Awareness and Advocacy* (0.938) but lower for *Non-Sexist* (0.680) compared to single-step categorization. The precision for *Objectification and Sexual Harassment* remains similar for both approaches. The other categories have a precision below 0.3. Table 3 also confirms the results seen in Table 2, where we assessed overall performance of the two approaches: better values are achieved for single-step categorization compared to two-step categorization.

Overall, single-step categorization outperformed two-step in most categories, as both recall and precision are higher. The accuracy is similar for both approaches, but Cohen’s Kappa is higher for single-step categorization. However, despite the higher Kappa for single-step categorization, it remains quite low, indicating only minimal agreement with human annotations. Consequently, these results should be interpreted with caution.

In the remainder of this section, where the category distribution is analyzed over time, and to answer **RQ2**, only the results for the single-step categorization are reported. All corresponding analyses for two-step categorization can be found in Appendix A.

4.2 Relative Frequencies of Categories

In Table 4, the relative frequencies of sexism categories are presented for different time frames, determined by single-step categorization. Since we have different numbers of tweets per time frame, the relative frequencies are assessed instead of the absolute. Table 4 shows that the relative frequency is highest for the *Non-Sexist* category across all three time frames: 90.72%, 85.22%, and 58.70% for BT2020, BT2024, and HT2024, respectively. When compar-

	Single-Step		Two-Step	
	Recall	Precision	Recall	Precision
Non-Sexist	0.292	0.875	0.667	0.680
Covert and Subtle Sexism	0.366	0.417	0.128	0.250
Discrediting	0.727	0.240	0.444	0.162
Objectification and Sexual Harassment	0.596	0.737	0.562	0.794
Remarks - Awareness and Advocacy	0.754	0.812	0.448	0.938
Stereotyping	0.429	0.200	0.500	0.189

Table 3: Classification metrics per sexism category for single- and two-step categorization.

	BT2020	BT2024	HT2024
Non-Sexist	90.72	85.22	58.70
Covert and Subtle Sexism	0.70	1.55	3.06
Discrediting	5.80	9.01	27.91
Objectification and Sexual Harassment	1.16	1.17	1.32
Remarks - Awareness and Advocacy	1.62	2.42	8.22
Stereotyping	0.00	0.64	0.78

Table 4: Relative frequency of sexism categories according to single-step categorization by time frame

ing BT2020 with BT2024 - the two election time frames where we had male vs. male candidates - single-step categorization suggests that *Non-Sexist* tweets decreased slightly (-5.50%). Whereas, when the election periods where two males were candidates, BT2020 and BT2024, are compared to the time frame HT2024 (female vs. male), we can see that *Non-Sexist* tweets became increasingly less prevalent (-32.02% and -26.52%, respectively).

The relative frequency of sexist tweets additionally increases when comparing male vs. male with female vs. male election periods, especially for the sexism categories *Covert and Subtle Sexism*, *Discrediting*, and *Remarks - Awareness and Advocacy*. The category *Discrediting* has the highest relative frequency (27.91%) for the election period HT2024 compared to the other categories and the election periods BT2020 and BT2024. In [Appendix C, Table 6](#) the additive and multiplicative changes between the three time frames are displayed.

When looking at the multiplicative change in relative frequency, we can observe the following. Comparing BT2020 with BT2024, single-step categorization suggests that sexist tweets became increasingly prevalent. *Covert and Subtle Sexism* had the largest relative increase, more than doubling in prevalence. *Discrediting* and *Remarks - Awareness and Advocacy* each increased by about 50%.

Comparing BT2020 to HT2024, these three categories (*Covert and Subtle Sexism*, *Discrediting*, and *Remarks - Awareness and Advocacy*) showed an even greater increase – up to 5 times as much. Meanwhile, *Non-Sexist* tweets decreased by 35%. The category *Objectification and Sexual Harassment* exhibited the least change in time frames. In

particular, in the first time frame, no tweets were classified as *Stereotyping*, though it is essential to consider that fewer tweets were classified in this period, which may have affected the results.

These results indicate that sexist tweets seem to have slightly increased between 2020 and 2024 and increase even more when a female is running for presidency. These results will be discussed in more detail in [section 6](#).

It is important to keep in mind that these interpretations are based on single-step categorization, which, as shown earlier in this section, has only limited reliability. In [Appendix C, Table 5](#) show the category distribution according to two-step categorization and its changes over time. However, it is crucial to note that the two-step approach performed comparatively poorly, making its distribution and observed changes over time less reliable.

In [Figure 2](#), the distribution of sexist categories over time is shown for the year 2024. The figure reveals that shifts in the distribution occurred suddenly rather than gradually, particularly when the presidential candidates changed and Harris replaced Biden. When looking at the shift for each sexism category, *Discrediting* and *Remarks - Awareness and Advocacy* have the steepest increase. This also reflects the results seen in [Table 4](#).

In [Appendix C](#) in [Figure 3](#) the same figures can be seen for two-step categorization. Also, in [Appendix C, Figure 4](#) the distribution for all categories (*Sexist* and *Non-Sexist*), according to single-step and two-step categorization, can be seen for each of the three time frames.

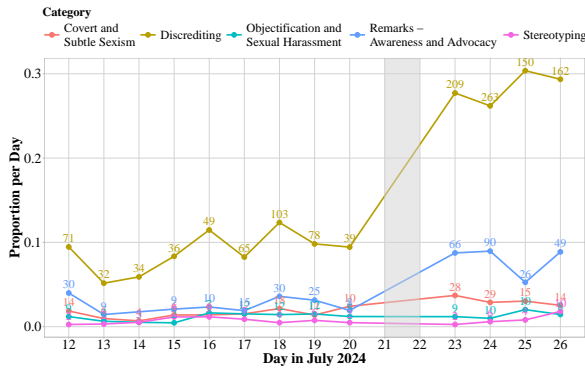


Figure 2: Distribution of sexist categories according to single-step categorization over time (gray area separating the two time frames BT2024 and HT2024)

5 Discussion

The results for **RQ1** show that single-step categorization outperforms the two-step approach. The findings indicate a higher Cohen’s Kappa, as well as better precision and recall for most sexist categories. However, the single-step categorization approach achieved weak agreement with human annotations, indicating that GPT-4 struggles to categorize finer instances of sexism. Although GPT-4 annotation has demonstrated promising results in the identification of hateful, offensive, and toxic language on social media (Li et al., 2024), these results contradict our findings. This could be due to the complexity of sexist language, making it harder for GPT-4 to detect, or the differences in methodology and the limitations presented in section 7. The minimal inter-annotator agreement additionally indicated that sexist content is complex and can be perceived differently by individuals. This shows that the challenge lies not only in the limitations of GPT-4 but also in the subjective nature of sexism classification itself. In contrast to Plaza et al. (2023) and Jiang et al. (2022), where classifying content as sexist or non-sexist before classifying it into finer categories proved more effective, our findings suggest the opposite.

The results for **RQ2** show a shift in sexist content across the three election periods. The relative frequency of sexist tweets increased from 2020 to 2024, with a particularly higher rise during the Harris vs. Trump 2024 election. These findings align with the previous research, showing that female candidates face increasing sexism in political discourse online (Bracic et al., 2019; Knuckey, 2019).

As seen in section 4, *Covert and Subtle Sexism* became increasingly more prevalent from 2020

to 2024, suggesting that sexist comments are becoming less explicit and more complex over time. The frequency of tweets that exhibited *Remarks - Awareness and Advocacy* regarding sexism particularly increased between the two male vs. male elections and HT2024. This indicates that a female candidate contributes to more discussions surrounding information about sexism or advocating for gender equality. The category *Discrediting* also had one of the highest relative frequency changes between elections with male vs. male candidates and female vs. male candidates. This increase in *Discrediting* tweets observed during HT2024 aligns with research by Falk (2010), which found that female politicians are often portrayed as less competent or natural compared to male candidates. However, the findings contradict Tromble and Koole (2020), who found no clear differences in the tone of messages directed at female and male politicians. This discrepancy could be explained due to the increasing polarization of US politics in recent years, especially on the platform X.

6 Conclusion

This research aimed to investigate sexist content in political discourse on social media during the 2020 and 2024 US election campaigns, comparing different time frames and candidate gender scenarios. Two approaches were used to detect sexism, and GPT-4’s role as a data annotation tool was evaluated. For **RQ1**, the results showed that the single-step categorization outperformed the two-step approach, but both had limited reliability and low agreement with human annotations. This highlights GPT-4’s limitations in sexism detection and the need for improved classification methods for social phenomena such as sexism. For **RQ2**, sexist discourse increased between 2020 and 2024, with a notable rise when Kamala Harris was a presidential candidate. These findings suggest female candidates continue to face gender-based discrimination in political discussions. At the same time, the challenges of detecting sexism are reflected both in the low human inter-annotator agreement and the model’s accuracy. This underlines the need for further research on capturing complex social phenomena such as sexism in computational research and emphasizes the importance of refining LLM-based sexism detection to support research on gender bias.

7 Limitations

This additional section points out several key limitations, which could potentially pave the way for future research.

Data Retrieval and API Constraints. A major limitation in the BT2020 timeframe is the availability and retrieval of tweets. Since data retrieval relied on tweet IDs from an existing dataset (Chen et al., 2022), many tweets were no longer accessible at the time of retrieval. Tweets that had been deleted by users or removed by the platform could not be retrieved, yet they still counted as requests due to the X API’s limitations. As controversial or highly offensive tweets may be more likely to be deleted, this introduces a potential bias. The BT2020 timeframe could underrepresent more extreme and offensive types of sexism, as tweets that provoked backlash or violated platform policies could have been removed. Additionally, the X API does not provide full-text access to retweets. Since retweets were included in the tweet ID dataset from Chen et al. (2022), when retrieving them, we obtained a truncated text, making them unsuitable for this analysis. Since the API does not allow pre-filtering based on whether a tweet is an original post or a retweet, extensive computational time was spent obtaining tweets that were not usable. The Basic X API version also limits the number of queries to 15 requests per 15 minutes, resulting in a long data collection period. Future research could explore alternative data retrieval methods, such as higher-level API access or pre-filtered data sets such as the data set for 2024 (Chen et al., 2022), to minimize data loss and computational time. Other research ideas for the future could expand the analysis beyond X. With the increasing role of platforms like TikTok, future research could use the TikTok API - which allows for quick keyword-based data collection without high computational time or major limitations - to reproduce this analysis. This would also enable researchers to examine sexist discourse across multiple social media platforms, providing a more comprehensive picture of sexism in online political discourse.

Annotation Bias. The annotation procedure potentially introduces a source of bias due to the limited number of annotators and their sociodemographic diversity. All three annotators in this study are white females with a shared social and cultural background, potentially influencing the

perception of sexist content. More diverse annotators, including individuals of different genders, ethnicities, and political perspectives, could provide a broader, less biased understanding of how to define sexist language in political debate. Additionally, the very low level of agreement between the first two annotators indicates that classifying sexism into fine-grained categories is a challenging and subjective task, even among individuals with similar backgrounds. As a result, the reported accuracy scores of GPT-4 should be interpreted with caution. Future research could focus on extending the annotation process in order to improve the classification reliability and strengthen the results of the research questions.

Keyword Discrepancies Between Time Frames. The 2020 data set was created using 227 keywords and account references (Chen et al., 2022), while the 2024 data sets are based only on 44 keywords (Balasubramanian et al., 2024). When comparing these, we found that only 10 keywords were identical in both data sets. Although some differences in the keywords are obvious, e.g., election-specific keywords such as "Trump2020" or "Harris2024," the overall difference in keyword quantity may have influenced the comparability of sexist content between the three time frames. A potential extension of this paper could be to reproduce the analysis by first generating a new list of keywords and extracting new tweets for each election time frame. This approach would address the issue of keyword discrepancies and also resolve the challenge of retrieving previously deleted tweets, as described in the Data Retrieval and API Constraints paragraph.

Platform Evolution. An important limitation when comparing 2020 and 2024 is the change in the social media platform X. Following Elon Musk’s acquisition of Twitter in October 2022, there were significant shifts in content moderation policies (Conger and Hirsch, 2022). While some previously suspended right-leaning accounts were restored, many left-leaning users left the platform (Barrie, 2023). As a result, the user base between 2020 and 2024 changed, which may have influenced the types of content shared and the tone of political discourse. This implies that the results should be interpreted within the context of X specifically, rather than as representative of the general population in the US. Future research could address this limitation by incorporating data from other platforms (e.g., TikTok or Reddit) or modeling changes in the

platform's user base over time. Despite this limitation, the results still indicate a notable increase in sexist content from the BT2024 time frame to the HT2024.

Candidate-Specific Factors. Sexist language is rarely isolated from other forms of marginalization. For instance, Kamala Harris is biracial, a stepmother, and a female candidate in a male-dominated office. This study centers on sexism and does not take other factors such as race, religion, or family structures into account. Consequently, some tweets labeled as sexist may be intersectional, while other tweets motivated by sexism but amplified by race or parental status could be under-captured. A better picture of online hostility might come from extending the taxonomy to include overlapping categories to control for other candidate-specific factors.

Contextual Differences Between Election Periods. Finally, when analyzing the results, the political context surrounding the 2020 and 2024 elections must be considered. The 2020 election period occurred during the COVID-19 pandemic. Although the 2020 election was dominated by online discussions and political discourse surrounding the pandemic, the 2024 election took place after the pandemic, which could lead to different discussion topics and a greater focus on other time-relevant topics. The presence or absence of major external events may have changed the way sexism manifested in online political discourse, making direct comparisons between time frames more complex. To account for this limitation, the tweets for both time frames were already filtered by specific keywords that could potentially be linked to sexism, as described in [subsection 3.3](#). One approach to further extend this analysis could focus on longitudinal tracking of sexist discourse beyond election cycles. Instead of focusing on a short 9-day election period, future research could analyze sexism in political discourse during a broader time period. This could help better understand whether the increase or decrease in sexist content in political discussions is temporary and event-driven or whether it indicates a broader societal trend.

References

Ashwin Balasubramanian, Vito Zou, Hitesh Narayana, Christina You, Luca Luceri, and Emilio Ferrara. 2024. [A public dataset tracking social media discourse](#)

[about the 2024 u.s. presidential election on twitter/x](#). Preprint, arXiv:2411.00376.

- Christopher Barrie. 2023. [Did the Musk takeover boost contentious actors on Twitter?](#) *Harvard Kennedy School (HKS) Misinformation Review*, 4(4).
- Julia C Becker and Stephen C Wright. 2011. [Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change](#). *Journal of personality and social psychology*, 101(1):62.
- Ana Bracic, Mackenzie Israel-Trummel, and Allyson F Shortle. 2019. [Is sexism for white people? gender stereotypes, race, and the 2016 presidential election](#). *Political Behavior*, 41(2):281–307.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Emily Chen, Ashok Dev, and Emilio Ferrara. 2022. [#election2020: the first public twitter dataset on the 2020 us presidential election](#). *Journal of Computational Social Science*, 5:1–18.
- CNN Politics. 2024. [Biden will not seek reelection; endorses harris](#). Accessed: 2025-03-15.
- Kate Conger and Lauren Hirsch. 2022. [Elon Musk Completes \\$44 Billion Deal to Own Twitter](#). *The New York Times*.
- Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, Aman Chadha, Mary Sandage, Laura-marie Pope, Gerry Dozier, and Cheryl Seals. 2024. [Investigating annotator bias in large language models for hate speech detection](#). In *Neurips Safe Generative AI Workshop 2024*.
- Erika Falk. 2010. *Women for president: Media bias in nine campaigns*. University of Illinois Press.
- Richard L Fox and Jennifer L Lawless. 2004. [Entering the arena? gender and the decision to run for office](#). *American journal of political science*, 48(2):264–280.
- Peter Glick. 2019. [Gender, sexism, and the election: did sexism help trump more than it hurt clinton?](#) *Politics, Groups, and Identities*, 7(3):713–723.
- Peter Glick and Susan T Fiske. 2001. [An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality](#). *American psychologist*, 56(2):109.

- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. [An expert annotated dataset for the detection of online misogyny](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. [Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech](#). In *Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion*, page 294–297, New York, NY, USA. Association for Computing Machinery.
- Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiega. 2022. [Swsr: A chinese dataset and lexicon for online sexism detection](#). *Online Social Networks and Media*, 27:100182.
- Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. [Annotation sensitivity: Training data collection methods affect model performance](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.
- Jonathan Knuckey. 2019. [“i just don’t think she has a presidential look”: Sexism and vote choice in the 2016 election](#). *Social Science Quarterly*, 100(1):342–358.
- Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. [“hot” chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media](#). *ACM Trans. Web*, 18(2).
- Joni Lovenduski. 2014. [The institutionalisation of sexism in politics](#). *Political Insight*, 5(2):16–19.
- Mary L. McHugh. 2012. [Interrater reliability: the kappa statistic](#). *Biochemia Medica*, 22(3):276–282.
- Ebuka Okpala and Long Cheng. 2025. [Large language model annotation bias in hate speech detection](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1389–1418.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Oxford English Dictionary. 2023. [sexism, n.²](#). Accessed: 2025-03-15.
- Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. [Misogyny detection in twitter: a multilingual and cross-domain study](#). *Information processing & management*, 57(6):102360.
- Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. [Multi-label categorization of accounts of sexism using a neural framework](#). 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 1642–1652, Hong Kong, China. Association for Computational Linguistics.
- Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. [Overview of exist 2023: sexism identification in social networks](#). In *European Conference on Information Retrieval*, pages 593–599. Springer.
- Kate A. Ratliff, Liz Redford, John Conway, and Colin Tucker Smith. 2019. [Engendering support: Hostile sexism predicts voting for donald trump over hillary clinton in the 2016 u.s. presidential election](#). *Group Processes & Intergroup Relations*, 22(4):578–593.
- Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. [Overview of exist 2021: sexism identification in social networks](#). *Procesamiento del Lenguaje Natural*, 67:195–207.
- Joshua Roesslein. 2020. [Tweepy: Twitter for python!](#) *GitHub*.
- Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. [Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114, Brussels, Belgium. Association for Computational Linguistics.
- Janet K Swim and Laurie L Cohen. 1997. [Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales](#). *Psychology of women quarterly*, 21(1):103–118.
- Janet K Swim, Robyn Mallett, and Charles Stangor. 2004. [Understanding subtle sexism: Detection and use of sexist language](#). *Sex roles*, 51:117–128.
- Rebekah Tromble and Karin Koole. 2020. [She belongs in the kitchen, not in congress? political engagement and sexism on twitter](#). *Journal of Applied Journalism & Media Studies*, 9(2):191–214.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2025. [A prompt pattern catalog to enhance prompt engineering with chatgpt](#). In *Proceedings of the 30th Conference on Pattern Languages of Programs, PLoP '23*, USA. The Hillside Group.
- X Developer Platform. 2025. [X API v2](#). Accessed: 2025-03-15.

A GPT-4 Prompts

A.1 Single-Step Approach

Single-Step Approach

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples: A tweet belongs to the category **Non-Sexist** if it is neutral or unrelated to sexism and it does not contain any discriminatory or gendered undertones. Examples for this category would be:

- "She gave a great speech on climate policy!"
- "We need more candidates like her!"
- "He would be better president."

A tweet belongs to the category **Remarks - Awareness and Advocacy** if it includes a remark or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These comments often aim to expose, discuss, or address sexism constructively. Examples:

- "We need more women in leadership roles—diversity strengthens democracy."
- "It's disheartening to see how women candidates are treated differently from men in this election."
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism."

A tweet belongs to the category **Discrediting** if it attempts to undermine women's competence, achievements, or worth without engaging in meaningful critique, often dismissing them outright (explicit negative sentiment targeting women, often demeaning their character, intentions, or abilities), or if it is an attempt to preserve male control or marginalize women from decision-making and public discussions. Examples:

- "What does she know about politics?"
- "She's only there to meet diversity quotas, not because she's actually qualified."
- "Politics is no place for a woman—leave it to the men."

A tweet belongs to the category **Stereotyping** if it enforces traditional gender roles or suggests that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs. Examples:

- "She's just another angry feminist, probably can't even cook."
- "Her place is to support men, not lead."
- "She's just overreacting, like all women do."

A tweet belongs to the category **Objectification and Sexual Harassment** if it reduces women to their physical appearance, treats them as objects of desire rather than individuals with agency or intellect, or uses sexualized language or harassment to intimidate or silence women in the political sphere. Examples:

- "Politics is for people with brains, not just pretty faces."
- "She needs to get laid, maybe she'll calm down."
- "She's asking for it, dressing like that in public."

A tweet belongs to the category **Covert and Subtle Sexism** if it shows unequal treatment that is not overtly hostile but perpetuates systemic inequality through normalized attitudes and actions, or if it masks sexism as positive sentiment, often portraying women as needing protection, incapable of independence, or inherently suited for specific roles. Examples:

- "She should smile more if she wants to connect with people."
- "It's surprising how well she handles tough questions for a woman."
- "We need women in politics to bring a softer touch."

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

A.2 Two-Step Approach

Two-Step Approach: First Step Prompt

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples: A tweet belongs to the category **Sexist** if it discriminates, demeans, or reinforces stereotype based on gender. This includes offensive language, objectification, gender-based slurs, or preserving harmful/negative gender roles. A tweet also belongs to this category if it discusses the topic of sexism, gender discrimination, or stereotypes but not in a way that is offensive towards people of certain gender. Examples for this category would be

- "It is insulting to women to have the obey-clause remain in the marriage service",
- "Girls shouldn't be allowed to be commentators for football games",
- "who asked you? Stupid bitch"

A tweet belongs to the category **Non-Sexist** if it is not related to sexism and do not contain any form of gender-based bias, discrimination, or stereotyping. The tweet is neutral, respectful, or positively inclusive in tone and content regarding gender. Examples for this category would be

- "We need more women in leadership roles—diversity strengthens democracy.",
- "It's disheartening to see how women candidates are treated differently from men in this election.",
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism."

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

Two-Step Approach: Second Step Prompt

I want you to perform a data annotation task. In your output, I want you to categorize a tweet into one of the following categories, considering the respective enclosed definition and three examples: A tweet belongs to the category **Remarks - Awareness and Advocacy** if it is/includes a remark or information highlighting sexism or advocating for gender equality in a way that is not offensive or derogatory. These kind of comments often aim to expose, discuss, or address sexism constructively. Examples for this category would be

- "We need more women in leadership roles—diversity strengthens democracy."
- "It's disheartening to see how women candidates are treated differently from men in this election."
- "She's breaking barriers and paving the way for future generations—proud of her resilience in the face of sexism."

A tweet belongs to the category **Discrediting** if it is/includes an attempt to undermine women's competence, achievements, or worth without engaging in meaningful critique, often dismissing them outright (explicit negative sentiment targeting women, often demeaning their character, intentions, or abilities), or if is an attempt to preserve male control or marginalize women from decision-making and public discussions. Examples for this category would be

- "What does she know about politics?"
- "She's only there to meet diversity quotas, not because she's actually qualified."
- "Politics is no place for a woman-leave it to the men."

A tweet belongs to the category **Stereotyping** if it is enforcing traditional gender roles or suggesting that women should occupy lower social, economic, or political statuses due to traditional or ideological beliefs. Examples for this category would be

- "She's just another angry feminist, probably can't even cook."
- "Her place is to support men, not lead."
- "She's just overreacting, like all women do."

A tweet belongs to the category **Objectification and Sexual Harassment** if it is reducing women to their physical appearance, treating them as objects of desire rather than individuals with agency or intellect, or if it using sexualized language or harassment to intimidate or silence women in the political sphere. Examples for this category would be

- "Politics is for people with brains, not just pretty faces."
- "She needs to get laid, maybe she'll calm down."
- "She's asking for it, dressing like that in public."

A tweet belongs to the category **Covert and Subtle Sexism** if it shows unequal treatment that is not overtly hostile but perpetuates systemic inequality through normalized attitudes and actions, or if it is masking sexism as positive sentiment, often portraying women as needing protection, incapable of independence, or inherently suited for specific roles. Examples for category would be

- "She should smile more if she wants to connect with people."
- "It's surprising how well she handles tough questions for a woman."
- "We need women in politics to bring a softer touch."

Your answer should only contain the name of the given category. Do not provide any other outputs or any explanation for your output.

B Alignment and Differences between Single-Step and Two-Step Approach

Alignment between Approaches

Tweet

The first female president should be honorable and should not have slept her way to the top.

Categorization for both Single-Step and Two-Step Approach

Discrediting

Difference between Approaches

Tweet

She has no skills. So maga it is

Categorization for Single-Step Approach

Discrediting

Categorization for Two-Step Approach

Non-Sexist

C Additional Results

	BT2020	BT2024	HT2024
Non-Sexist	98.144	95.702	88.501
Covert and Subtle Sexism	0.232	0.320	0.570
Discrediting	0.696	2.007	5.732
Objectification and Sexual Harassment	0.000	0.675	0.783
Remarks - Awareness and Advocacy	0.000	0.711	3.667
Stereotyping	0.928	0.586	0.748

Table 5: Relative frequency of sexism categories according to two-step categorization by time frame

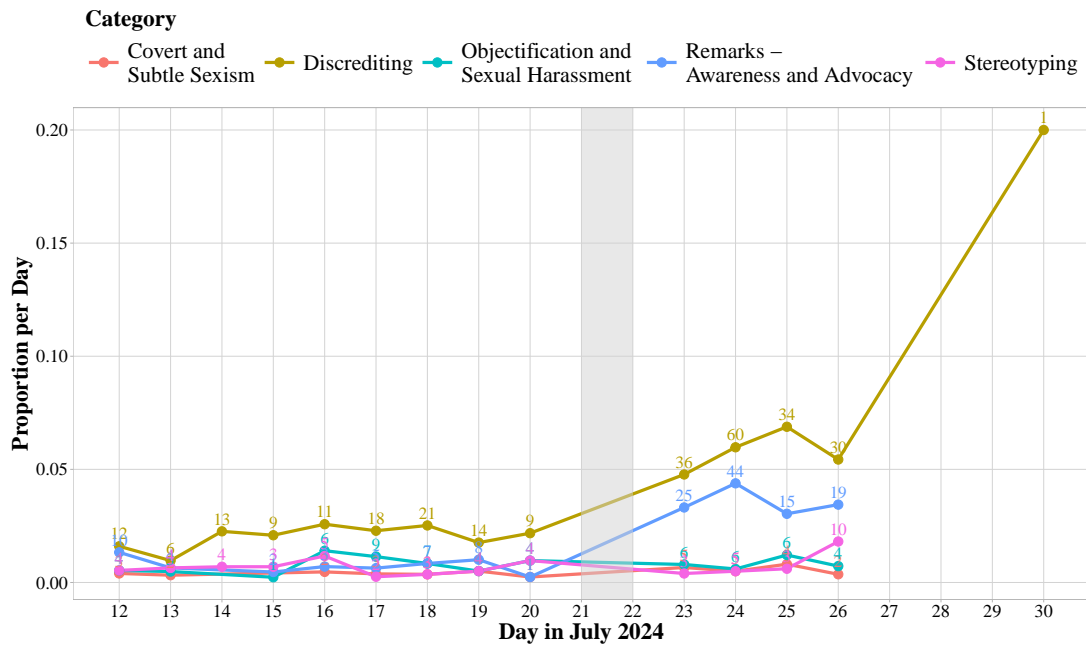


Figure 3: Distribution of sexist categories according to two-step prompting over time in 2024 (gray area separating the time frames BT2024 and HT2024)

Change from BT2020 to BT2024

	BT2020	BT2024	BT2020 to BT2024 (additive)	BT2020 to BT2024 (multiplicative)
Non-Sexist	90.72	85.22	-5.50	0.94
Covert and Subtle Sexism	0.70	1.55	+0.85	2.22
Discrediting	5.80	9.01	+3.21	1.55
Objectification and Sexual Harassment	1.16	2.42	+0.01	1.01
Remarks - Awareness and Advocacy	1.62	2.42	+0.80	1.49
Stereotyping	0.00	0.64	+0.64	Inf

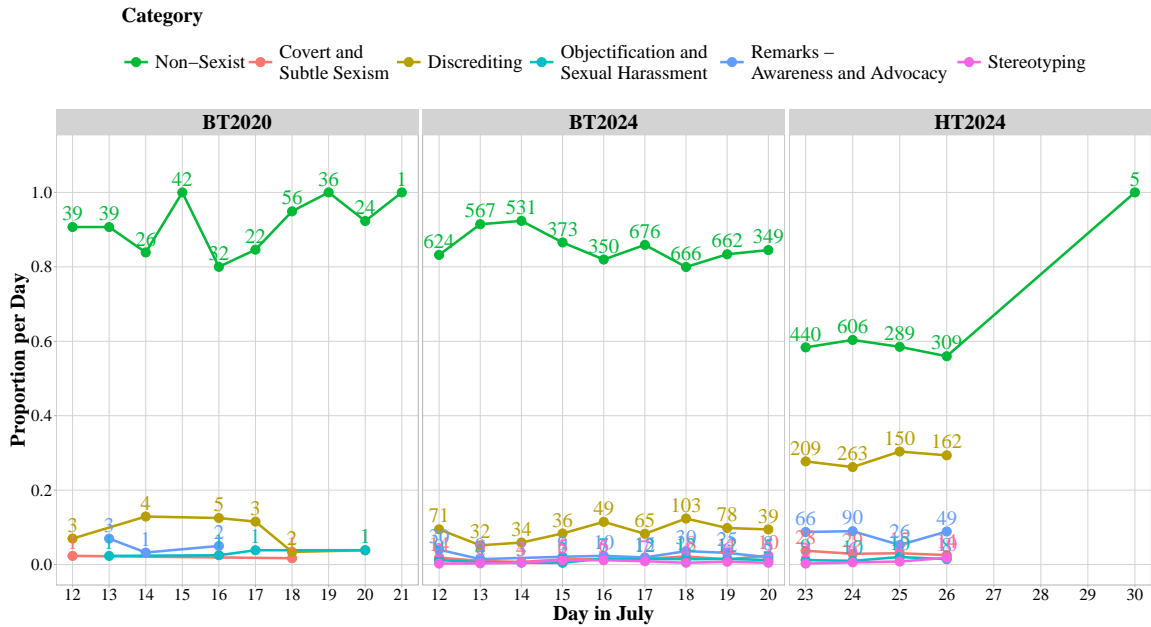
Change from BT2020 to HT2024

	BT2020	HT2024	BT2020 to HT2024 (additive)	BT2020 to HT2024 (multiplicative)
Non-Sexist	90.72	58.70	-32.02	0.68
Covert and Subtle Sexism	0.70	3.06	+2.36	4.40
Discrediting	5.80	27.91	+22.11	4.81
Objectification and Sexual Harassment	1.16	1.32	+0.16	1.14
Remarks - Awareness and Advocacy	1.62	8.22	+6.60	5.06
Stereotyping	0.00	0.78	+0.78	Inf

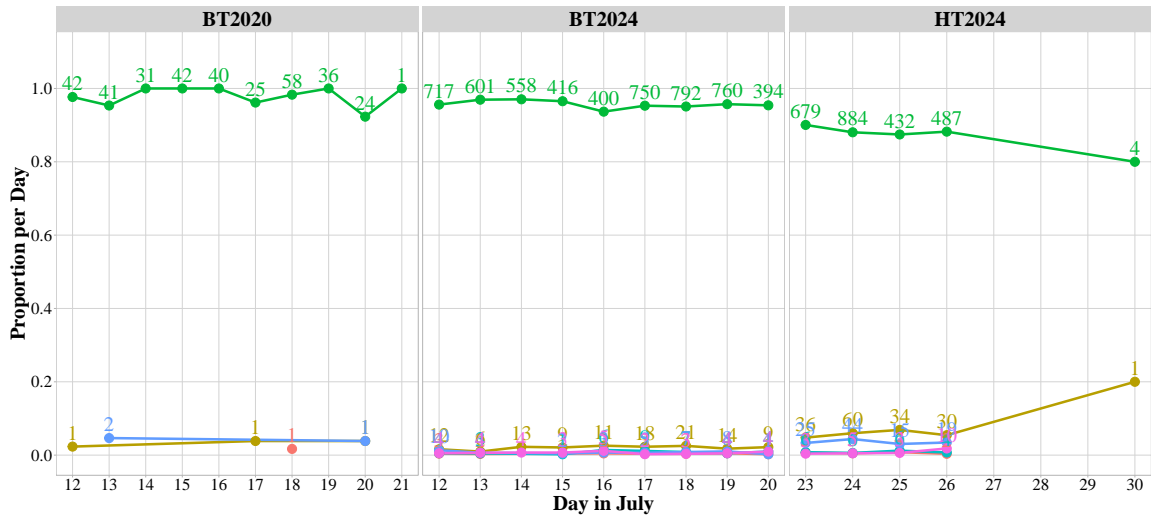
Change from BT2024 to HT2024

	BT2024	HT2024	BT2024 to HT2024 (additive)	BT2024 to HT2024 (multiplicative)
Non-Sexist	85.22	58.70	-26.52	0.69
Covert and Subtle Sexism	1.55	3.06	+1.51	1.98
Discrediting	9.01	27.91	+18.90	3.10
Objectification and Sexual Harassment	1.17	1.32	+0.15	1.12
Remarks - Awareness and Advocacy	2.42	8.22	+5.80	3.40
Stereotyping	0.64	0.78	+0.14	1.22

Table 6: Change in relative frequency of sexism categories according to single-step categorization by time frame



(a) Single-Step Categorization



(b) Two-Step Categorization

Figure 4: Distribution of all categories over time according to single-step (a) and two-step (b) categorization

Discourse Relation Recognition with Language Models Under Different Data Availability

Shuhaib Mehri Chuyuan Li Giuseppe Carenini

Department of Computer Science, University of British Columbia

V6T 1Z4, Vancouver, BC, Canada

shuhaib@student.ubc.ca, chuyuan.li@ubc.ca, carenini@cs.ubc.ca

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance across various NLP tasks, yet they continue to face challenges in discourse relation recognition (DRR). Current state-of-the-art methods for DRR primarily rely on smaller pre-trained language models (PLMs). In this study, we conduct a comprehensive analysis of different approaches using both PLMs and LLMs, evaluating their effectiveness for DRR at multiple granularities and under different data availability settings. Our findings indicate that no single approach consistently outperforms the others, and we offer a general comparison framework to guide the selection of the most appropriate model based on specific DRR requirements and data conditions.

1 Introduction

Discourse parsing automatically extracts the underlying discourse structure of a text, playing a pivotal role in various natural language processing (NLP) tasks. Its utility has been demonstrated in applications such as machine translation (Chen et al., 2020), summarization (Xu et al., 2020; Chen and Yang, 2021; Rennard et al., 2024), and question-answering (Jansen et al., 2014). Discourse parsing is particularly useful in scenarios that involve handling complex or large-scale text, such as in multi-document summarization (Chen et al., 2021; Li et al., 2020; Liu and Lapata, 2019).

A fundamental task in discourse parsing is discourse relation recognition (DRR), which aims to identify the relation sense between argument pairs. Typically, argument pairs are made up of text spans known as elementary discourse units (EDUs). When connectives are present between argument pairs (explicit DRR), training a simple classifier on the connectives can achieve a classification accuracy close to 95% (Xiang and Wang, 2023; Pitler and Nenkova, 2009; Varia et al., 2019). On the other hand, the task becomes more difficult when

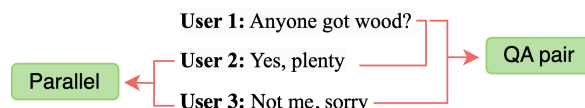


Figure 1: An example of discourse relation parsing in dialogue, taken from STAC corpus (Asher et al., 2016).

connectives are not present (implicit DRR), and current approaches for this task struggle to achieve an accuracy above 80% (Xiang et al., 2023; Zhou et al., 2022; Chan et al., 2023). To address this challenge, we explore relation recognition in dialogue discourse parsing (see Figure 1), where connectives play a less prominent role, alongside implicit discourse relation recognition (IDRR) in monologues. In dialogue discourse parsing, the argument pairs are made up of user utterances, and in IDRR, the argument pairs are made up EDUs.

Recent large language models (LLMs) (e.g., ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI et al., 2024)) have demonstrated remarkable performance on many NLP benchmarks, and display advanced reasoning and understanding capabilities. They also exhibit impressive abilities in zero-shot and few-shot settings (Wei et al., 2022), and can sometimes be competitive with prior state-of-the-art fine-tuning approaches (Brown et al., 2020). At the same time, many studies suggest that LLMs do not perform as well as small encoder-only models fine-tuned on specific-tasks (Qin et al., 2023; Lu et al., 2023).

This is the case for the DRR task on which LLMs seem to struggle with (Fan et al., 2024; Chan et al., 2024). Many of the current top-performing approaches rely on fine-tuning relatively smaller encoder-based pre-trained language models (PLMs) like RoBERTa (Zhou et al., 2022; Wu et al., 2023; Xiang et al., 2023, 2022; Li et al., 2023, 2024a,b).

In spite of these established approaches, it is still

unclear when it is more effective to use LLMs or PLMs for DRR. With this in mind, we conduct a comprehensive analysis of different approaches for the DRR task, focusing on comparing PLMs and LLMs under different data availability settings. For PLMs, we use the data for fine-tuning. For LLMs, we employ zero-shot prompting, in-context learning, and a new self-reflection technique we call confusion-matrix prompting. We explore these techniques using both monologues and dialogues with different relation types and granularities.

Confusion-matrix prompting is a novel technique that uses information from a confusion matrix to inform an LLM about the errors it tends to make, enabling it to self-reflect and adjust its predictions accordingly. This is inspired by the many studies that have shown how LLMs benefit from self-reflecting on and improving their initial generation (Madaan et al., 2023; Fernando et al., 2023; Welleck et al., 2023; Shinn et al., 2023), as well as learning from their mistakes (Zhang et al., 2024).

Our work advances the understanding of fine-tuning PLMs and various prompting techniques with LLMs in the context of DRR, across different dataset sizes and multiple relation sense granularities. Key takeaways include: (1) Zero-shot prompting leverages inherent knowledge embedded in LLMs and performs better than other techniques when there is little available data; (2) Confusion-matrix prompting achieves optimal performance when there is insufficient data for fine-tuning, but enough to surpass zero-shot performance; (3) Fine-tuned PLMs excel in scenarios with increased data, and is robust across various datasets regardless of complexity or number of relation senses.

2 Methodology

To simulate different data availability settings, we extract seven subsets of training datasets, each with different sizes. For each subset, we randomly select a certain number of examples for each relation sense. We start with a single example per relation sense, and increment the number up to 250 examples per relation sense. When a specific relation sense does not enough examples available, we randomly select the remaining examples from other relation senses to satisfy the target example count.

Next, we employ fine-tuning and prompting techniques that leverage these subsets for the DRR task, assessing how each performs across different data volumes.

- **Fine-tuning (FT).** We fine-tune an encoder-only PLM to encode the representation of argument pairs and predict a relation sense. Our representation of argument pairs follows the template from Zhou et al. (2022): Arg1: <Arg1>. Arg2: <Arg2>. In summary, the discourse relation between Arg1 and Arg2 is
- **Zero-shot (ZS).** Without using any annotated data, we frame the problem as a zero-shot fill-in-the-blank prompt to a LLM. Our prompt follows the same format as FT.
- **In-context learning (ICL).** Input-label pairs from the dataset are incorporated directly into the LLM’s prompt to leverage in-context learning. Typically, in-context learning approaches manually select the input-label pairs to ensure high-quality examples. However, in our approach, we use randomly selected pairs from the dataset to maintain consistency with our other techniques. Due to the limited input context length of the early GPT-3.5 version, we cannot include examples for all data availability settings, particularly for larger numbers of examples per relation sense, e.g., >25 examples per relation sense for PDTB top-level experiment.
- **Confusion-Matrix Prompting (CMP).** Using the dataset, we collect zero-shot performance of the LLM in the form of a confusion matrix, recording the model’s predictions against the true labels. This confusion matrix allows us to determine how often the model correctly predicts a relation, and how often it confuses it with another relation. During inference, we first let the model make its initial prediction. Based on this prediction we formulate a follow-up prompt using the confusion matrix, informing the model of its prediction accuracy and common mistakes. We provided this prompt as a follow-up, giving the model a chance to self-reflect and correct it’s initial prediction (see Appendix A for an example).

3 Experimental Setup

Monologue Data: The Penn Discourse Treebank 3.0 (PDTB 3.0). PDTB 3.0 is an annotated corpus of discourse relations that come from Wall Street Journal articles (Webber et al., 2019). It uses 3 hierarchies of relation senses, and contains both implicit and explicit relation types. For our

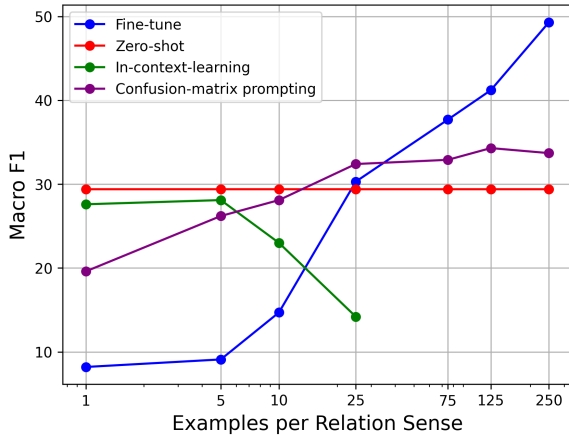


Figure 2: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB top-level. The number of ICL examples is restricted to up to 25 due to the input context length of GPT-3.5.¹

experiments, we use the four top-level and twenty second-level implicit relation senses, with a test set of 1,538 examples.

Dialogue Data: STAC. STAC is a corpus of multi-party dialogues collected from an online game called *The Settlers of Catan* (Asher et al., 2016). The dialogues are annotated in the style of Segmented Discourse Representation Theory (SDRT), which uses sixteen relation senses (Asher and Lascarides, 2003). The test set consists of 1,128 examples.

Implementation Details. The fine-tuning experiments were conducted using RoBERTa-base (Liu et al., 2019). We selected this lightweight model for its strong performance on the DRR task. We employ a learning rate of $1e - 5$ and trained the model for 20 epochs with early stopping based on performance on the development set. To ensure robustness of our results, we repeat each experiment over 10 random seeds and report the average score.

ZS, ICL and CMP experiments were done using GPT-3.5 Turbo. We report the average of our results over 5 random seeds. Additionally, preliminary experiments were performed on Mistral 7B (Jiang et al., 2023), and indicated a similar trend of improvement in performance.

4 Results and Analysis

The results of our experiments are displayed in Table 1 and illustrated in Figures 2, 3, and 4 for

¹Logarithmic scale is used for the x-axis

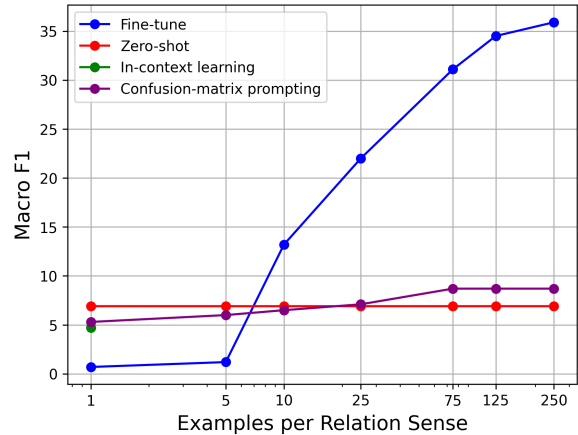


Figure 3: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB second-level. ICL examples is restricted to 1 example per relation sense.¹

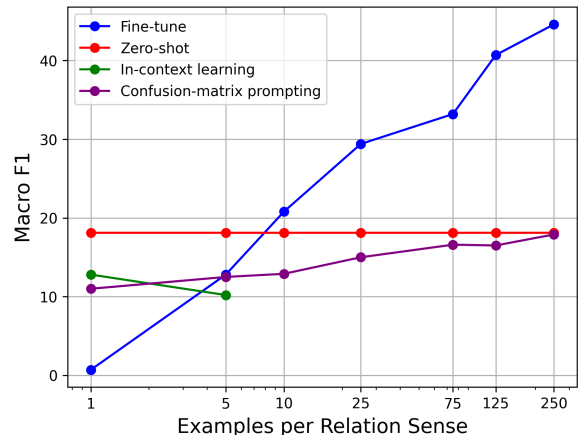


Figure 4: Comparisons of Macro F1 scores of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on STAC. ICL examples is restricted to up to 5 examples per relation sense.¹

PDTB top-level, second-level, and STAC corpus, respectively. Our analysis primarily uses Macro F1 scores, though similar trends are observed for accuracy (relevant figures are included in Appendix B).

In general, ZS is consistently the better technique for lesser amounts of data. As the number of examples per relation sense increases, fine-tuning (FT) demonstrates constant improvement and soon surpasses ZS, underlining the benefits of the technique. While CMP starts out with lower performance, it has shown to improve and eventually surpass ZS at higher data volumes. ICL, on the other hand, exhibits underwhelming performance across all datasets.

The observed trends indicate that ZS, which relies on inherent discourse knowledge embedded in

Number of Training Examples Per Relation Sense (always zero for ZS)														
	1		5		10		25		75		125		250	
Technique	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
PDTB Top-Level (4 relation senses)														
FT	20.9	8.2	22.7	9.1	22.9	14.7	34.5	30.3	39.9	37.7	43.1	41.2	51.7	49.3
ZS	42.8	29.4	42.8	29.4	42.8	29.4	42.8	29.4	42.8	29.4	42.8	29.4	42.8	29.4
ICL	40.8	27.6	39.6	28.1	33.1	23.0	18.5	14.2	-	-	-	-	-	-
CMP	22.5	19.6	33.1	26.2	35.4	28.1	39.5	32.4	42.0	32.9	42.1	34.3	42.3	33.7
PDTB Second-Level (20 relation senses)														
FT	5.6	0.7	6.2	1.2	15.0	13.2	25.4	22.0	38.4	31.1	42.6	34.5	50.1	35.9
ZS	17.8	6.9	17.8	6.9	17.8	6.9	17.8	6.9	17.8	6.9	17.8	6.9	17.8	6.9
ICL	9.1	4.7	-	-	-	-	-	-	-	-	-	-	-	-
CMP	11.5	5.3	13.9	6.0	13.4	6.5	11.7	7.1	19.3	8.7	19.5	8.7	20.5	8.7
STAC (16 relation senses)														
FT	4.3	0.7	20.0	12.8	31.0	20.8	39.9	29.4	46.5	33.2	53.0	40.7	59.1	44.6
ZS	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1	24.9	18.1
ICL	20.6	12.8	17.2	10.2	-	-	-	-	-	-	-	-	-	-
CMP	16.2	11.0	14.9	12.5	16.2	12.9	19.3	15.0	21.6	16.6	22.1	16.5	26.8	17.9

Table 1: Accuracy (Acc) and Macro $F1$ (F1) scores of FT, ZS, ICL, and CMP techniques on different numbers of examples per relation sense. The best results for each technique are bolded. The - values indicate that we were unable to experiments due to input length limitations.

the model, is the highest performing technique for DRR in low data availability scenarios. When considering ZS performance across the datasets, the performance diminishes for more complex problems where there are greater numbers of relation senses. ZS achieves higher performance on top-level PDTB, with 4 relations senses, and lower performance on STAC, with 16 relation senses, and even lower performance in second-level PDTB, with 20 relation senses. This increased difficulty highlights the limitations of relying solely on pre-trained knowledge.

FT scales very well with the data and always emerges as the most effective technique as data availability increases. Notably, the accuracy and F1 scores achieved by FT are relatively consistent across the different datasets. Unlike in ZS, we do not see a similar drop in performance as the number of relation senses increases. From this, we can gather that a more complex task does not proportionally impact the performance of PLMs the same way it does for LLMs.

CMP begins to outperform ZS as dataset sizes increase, showing that it is optimal when the amount of data is insufficient for fine-tuning, or if fine-tuning is not a viable option. In scenarios involving smaller datasets, the volatility of the confusion matrix is less representative of model performance, often causing a drop in performance. However, as we use larger datasets, the confusion matrix provides

a more accurate depiction of the model’s errors and overall performance. This allows CMP to help the LLM learn from its past performance and start outperforming ZS.

Furthermore, CMP proves to be more effective in the more complex datasets with larger numbers of relation senses. This effectiveness is attributed to the technique being beneficial when there are more potential mistakes that the LLM can make.

The results observed from ICL gives poor results, which is likely due to the random selection of examples and context length limitations. It never outperformed ZS, and the performance decreases as the datasets get larger, as if adding more data into the prompt makes it more difficult for the LLM to effectively process.

5 Conclusion

In order to identify the optimal techniques for DRR under different data availability settings, we perform an analysis on how these techniques perform with varying amounts of data. The techniques we explore include fine-tuning for PLMs, and various prompting techniques with LLMs. In our experiments, we find that in low data availability scenarios, zero-shot prompting performs best. CMP achieves the best performance when there is more data available, but not enough for effective fine-tuning. When we have more data, fine-tuning PLMs dominates, and performance is not affected

by more complex relation sense granularities. Unexpectedly, ICL is always dominated by ZS.

In future work, we plan to further investigate the trade-off between PLMs and LLMs for discourse processing tasks. We would like to extend this work by conducting further experiments on more powerful LLMs, more specific ICL techniques such as similarity-based selection, as well as more complex tasks such as discourse parsing. Additionally, we would like to explore self-reflection learning techniques for LLMs as we have found quite promising results. The methodology and experimental framework we have designed and implemented¹ will be critical in facilitating these further investigations by us and other researchers.

6 Limitations

In our experiments, we consider GPT-3.5 Turbo and RoBERTa-base as representatives for LLMs and PLMs, respectively. While these models serve as good representatives, exploring more powerful models would further strengthen our study. Furthermore, due to the lack of annotated data, our experiments were limited to two English datasets: PDTB 3.0 and STAC. These limitations highlight areas for future research to provide a more comprehensive understanding of discourse relation recognition

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. [Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. [Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. [DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. [Structure-aware abstractive conversation summarization via discourse and action graphs](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. [Modeling discourse structure for document-level neural machine translation](#). In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. [SgSum:transforming multi-document summarization into sub-graph selection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4063–4074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. [Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study](#). *Preprint*, arXiv:2305.08391.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. [Promptbreeder: Self-referential self-improvement via prompt evolution](#). *Preprint*, arXiv:2309.16797.

¹The code will be made public after publication

- Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. [Discourse complements lexical semantics for non-factoid answer reranking](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Chuyuan Li, Maxime Amblard, and Chlo   Braud. 2023. [A semi-supervised dialogue discourse parsing pipeline](#). In *Journ  es Scientifiques du GDR Lift (LIFT 2023)*.
- Chuyuan Li, Chlo   Braud, Maxime Amblard, and Giuseppe Carenini. 2024a. [Discourse relation prediction and discourse parsing in dialogues with minimal supervision](#). In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics.
- Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024b. [Dialogue discourse parsing as generation: A sequence-to-sequence LLM-based approach](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics.
- Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. [Leveraging graph to improve abstractive multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019. [Hierarchical transformers for multi-document summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.
- 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.
- Yuxiang Lu, Yu Hong, Zhipang Wang, and Guodong Zhou. 2023. [Enhancing reasoning capabilities by instruction learning and chain-of-thoughts for implicit discourse relation recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5634–5640, Singapore. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- OpenAI. 2022. [Chatgpt: Optimizing language models for dialogue](#). *OpenAI*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim  n Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

- Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Emily Pitler and Ani Nenkova. 2009. [Using syntax to disambiguate explicit discourse connectives in text](#). In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is ChatGPT a general-purpose natural language processing task solver?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.
- Virgile Rennard, Guokan Shang, Michalis Vazirgiannis, and Julie Hunter. 2024. [Leveraging discourse structure for extractive meeting summarization](#). *Preprint*, arXiv:2405.11055.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflection: language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.
- Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. [Discourse relation prediction: Revisiting word pairs with convolutional networks](#). In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#). *Preprint*, arXiv:2109.01652.
- Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. [Generating sequences by learning to self-correct](#). In *The Eleventh International Conference on Learning Representations*.
- Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. [Connective prediction for implicit discourse relation recognition via knowledge distillation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.
- Wei Xiang, Chao Liang, and Bang Wang. 2023. [TEPrompt: Task enlightenment prompt learning for implicit discourse relation recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12403–12414, Toronto, Canada. Association for Computational Linguistics.
- Wei Xiang and Bang Wang. 2023. [A survey of implicit discourse relation recognition](#). *ACM Comput. Surv.*, 55(12).
- Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022. [ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Discourse-aware neural extractive text summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.
- Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zheng, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. 2024. [In-context principle learning from mistakes](#). *Preprint*, arXiv:2402.05403.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. [Prompt-based connective prediction method for fine-grained implicit discourse relation recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Confusion-Matrix Prompting Examples

User: Arg1: Coupons and a newsletter will be mailed. Arg2: And the sponsor will be able to gather a list of desirable potential customers.

In summary, the discourse relation between Arg1 and Arg2 is

Model: Expansion

User: Arg1: Coupons and a newsletter will be mailed. Arg2: And the sponsor will be able to gather a list of desirable potential customers.

The initial prediction for the discourse relation between Arg1 and Arg2 was Expansion. 27% of the time when Expansion was predicted, the correct answer was Expansion. 27% of the time when Expansion was predicted, the correct answer was Contingency. 24% of the time when Expansion was predicted, the correct answer was Temporal. 20% of the time when Expansion was predicted, the correct answer was Comparison. Considering this information, what is the relation sense?

Model: Contingency

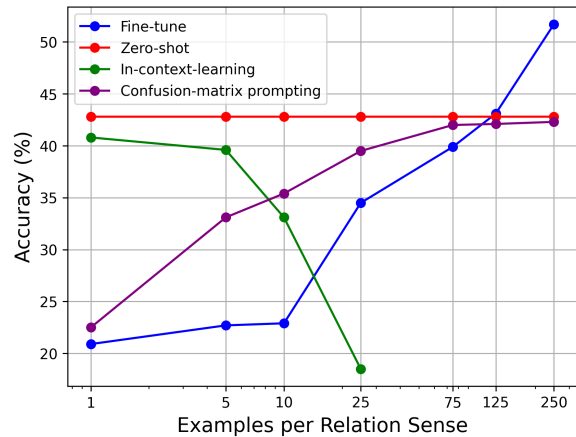


Figure 5: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB top-level. ¹

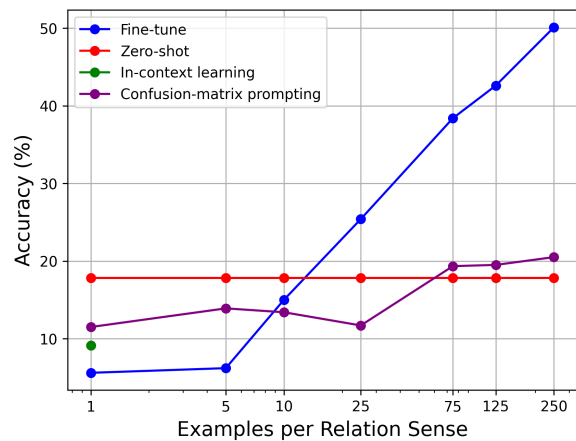


Figure 6: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on PDTB second-level. ¹

B Accuracy Comparisons of Techniques

¹Logarithmic scale is used for the x-axis

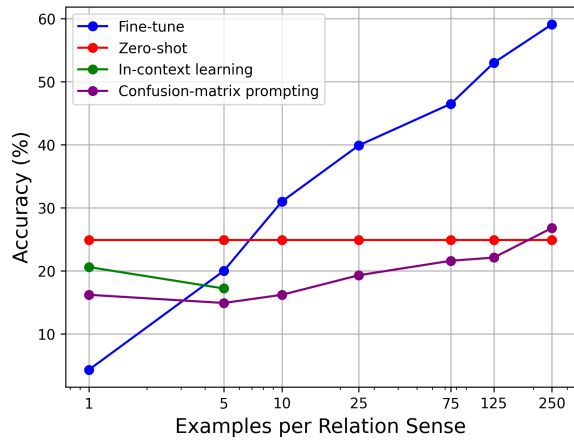


Figure 7: Comparisons of accuracy of FT, ZS, ICL, and CMP techniques across increasing numbers of examples per relation sense on STAC. ¹

EmbiText: Embracing Ambiguity by Annotation, Recognition and Generation of Pronominal Reference with Event-Entity Ambiguity

Amna

IT University of Copenhagen
amnasheikh1@gmail.com

Christian Hardmeier

IT University of Copenhagen
chrha@itu.dk

Abstract

Consider the example “*The bird sang the nursery rhyme beautifully. It made everyone in the room smile*”. The pronoun ‘it’ here refers either to the bird or to the event of singing. This example is inherently ambiguous. It cannot be meaningfully disambiguated as an event or entity reference, as both readings result in the same text meaning. This study introduces a new dataset **EMBITEXT** to preserve ambiguity in the language by navigating through the ambiguity surrounding the pronominal reference to the entity or event. Oftentimes, ambiguity does not necessarily need to be resolved but is modelled carefully. Furthermore, this study explores the capacity of LLMs (Llama, Mistral, Gemini, Claude AI) to embrace ambiguity in generating text that exhibit referential ambiguity via an In-Context learning approach. To evaluate of the dataset, RoBERTa was fine-tuned on this data to model ambiguity while simultaneously distinguishing between entity or event references. Results demonstrate EmbiText’s capacity to advance the ongoing NLP research by modelling linguistic ambiguity in computational environments instead of fully disambiguating it, thereby retaining diverse interpretations where resolution may alter meaning.

1 Introduction

Ambiguity in language represents multiple plausible meanings, interpretations, and contexts of words, phrases, or sentences. The occurrence of ambiguity in language is inherently natural and can be a stylistic choice or a result of poetic expression, but also occur unintentionally. Humans tend to navigate around these ambiguities naturally at most times as compared to computers and machines, although in extreme cases ambiguity may lead to profound confusion for even humans. Researchers and engineers are leveraging Artificial Intelligence (AI) to navigate through this ambiguity along with its contextual understanding with a level of naturalness akin to human understanding of language.

A simple pronoun could refer to any entity, an event, or in some cases may refer to both. An entity is typically a noun denoting an object inside the discourse realm, whereas an event is a verbal phrase describing an action that has occurred. Anaphora resolution research has traditionally focused on complete disambiguation of pronominal references, as seen in corpora like OntoNotes by Weischedel et al. (2010) and GUM Corpus by Zeldes et al. (2025), yet some ambiguities are difficult to resolve or resolving them lead to discrepancies in coreference annotation. Lapshinova-Koltunski et al. (2019) This calls for a dataset, specifically curated and robust to include ambiguous examples and their potential antecedents enabling the NLP models to both detect ambiguity and quantify uncertainty.

Examples of ambiguous cases this study primarily focuses on:

- *The bird mimicked the nursery rhymes beautifully. It made everyone in the room smile.*
- *The volcano erupted violently. It created a huge crater.*
- *The garden was blooming with flowers, which made me feel refreshed.*
- *The fireworks display wonderfully lit up the sky on time. This added colours to the ceremony.*

This study underscores the linguistic fundamentals of ambiguity, reflecting on semantics, linguistic theories, and syntax to interpret how multiple readings can be extracted from pronoun reference.

This study addresses these research questions:

- How can the ambiguity inherent in pronominal references between entities and event be identified, annotated, and modeled in Natural Language?

- Are LLMs capable of embracing ambiguity in natural language rather than resolving it?

We answer these questions by developing a curated dataset that models the ambiguity in pronominal references between entities and events. Data is annotated and then evaluated by fine-tuning an LLM. The text examples exhibit ambiguity surrounding pronoun reference. Additionally, LLMs are prompted to generate ambiguous examples and quantify the likelihood of pronouns referring to entities and events. This study aims to contribute to the literature in Natural Language Processing by exploring the complexities of natural language and leveraging AI to preserve linguistic ambiguity.

2 Background

The literature review explores types of ambiguities in natural language, i.e. syntactic, discourse, anaphoric, semantic, and lexical ambiguities such as in a study by [Anjali and Anto \(2014\)](#). Considerable focus has been on investigating ambiguities in different settings. [Duzi \(2013\)](#) claims that ambiguity is not only prominent in informal conversations but also evidently exists in formal discussions and arguments, particularly focusing on philosophical approaches. [Chukwu \(2015\)](#) discovers and resolves ambiguities and incorporates admissible ambiguity into literary writing to understand word order and context.

Researchers have developed corpora for coreference resolution. [Yuan et al. \(2023\)](#) introduced a corpus of sentence pairs with ambiguous and unambiguous referents to compare human and model sensitivity to ambiguity, focusing primarily on disambiguation. In contrast, EmbiText models graded ambiguity, thereby prioritizing preservation over resolution of ambiguity. Datasets like LitBank by [Bamman et al. \(2020\)](#), LegalCore by [Wei et al. \(2025\)](#), and PreCO by [Chen et al. \(2018\)](#) are aimed at entity-level coreference and yield efficient error analysis, but ignore combined representation of entities and events. [Emami et al. \(2019\)](#) introduces a context-driven coreference corpus by eliminating gender and number cues. KnowRef focuses on disambiguation, unlike EmbiText, which embraces ambiguity.

[Loáiciga et al. \(2017\)](#) investigates the ambiguous nature of the pronoun "it" by applying the maximum Entropy classifier to differentiate between anaphoric, event-referential, and pleonastic uses of "it", with a focus on a single type of pronoun "it"

and silver-standard data. [Loáiciga et al. \(2020\)](#) subsequently introduced cross-lingual signals-related disambiguation system for event-based ambiguities with exclusive focus on the English pronoun 'it' and reliance on silver standard data. This narrows the scope and limits applicability to a wider range of contexts. [Bevacqua et al. \(2021\)](#) investigated linguistic patterns in event-entity coreference across five languages using the story continuation task, while focusing on disambiguation using a psycholinguistic approach rather than representing ambiguity as linguistic characteristic with computational models. [Joshi et al. \(2019\)](#) proposed efficient BERT-based system for entity coreference which struggled with encoding relations between entities. [Le et al. \(2022\)](#) proposed extremely accurate scientific coreference resolution with In Context learning, but is restricted by prompt-based capacity and cross-domain generalization. These studies underscore the need for extensive research, which our study aims to accomplish by curating EmbiText, annotated data to navigate through ambiguity in pronominal reference and leveraging LLMs to provide insights about the linguistic phenomena related to pronoun reference. Unlike traditional disambiguation, this study coherently embraces ambiguity by introducing data with inherently ambiguous cases.

3 Methodology

This section outlines the complete study pipeline, from data acquisition and processing to model training. Provided that the focus of the study is on pronoun reference, the selected pronouns are: 'It', 'This', 'That and 'Which'.

3.1 Data Acquisition

Georgetown University Multilayer Corpus (GUM) by [Zeldes et al. \(2025\)](#) was chosen for experimental analysis as it contains real-world examples from various domains such as academic, art, literature, interviews, etc. Datasets from the coreference section of GUM were selected due to their relevance to the focus of this study, containing a total of 14158 text examples. About 4860 text examples containing the selected pronouns were extracted. Each example was subjected to extensive auditing to examine whether the pronoun referred to an antecedent and to check for potential ambiguity. Examples containing dummy pronouns e.g. "*Basing letters on objects (pictographs) is an easy way to start a writing system. Try this with a group of friends. It's*

much more fun when there are other people that can understand your language” and informal dialogues i.e. *”But, but I remember, like I went there with this person, it’s kind of funny*” were excluded. The *’it’* here functions as a syntactic placeholder without an antecedent. In total, 249 examples were sampled from the corpus, containing a mix of ambiguous and unambiguous examples. We conducted a text generation experiment by using Decoder-only LLMs with an In-Context Few Shot learning approach to assess their capability in generating examples containing ambiguous pronoun references. Mistral AI (Mistral-7B-Instruct-v0.1) by (Jiang et al., 2023), Gemini 2.0 by (Google DeepMind, 2023), Llama 3.0 (llama-3-8b-instruct) by (Grattafiori et al., 2024) and Anthropic’s Claude AI (Claude 3 Haiku) (Anthropic, 2024) were selected, fine-tuned to generate texts identical to the requirement of this research. The hyperparameters included: temperature set to 0.7, do_sample set to true, and max_new_tokens set to 700 for Mistral and 1000 for Claude. This setting ensured a controlled level of linguistic creativity, diversity, and randomness in the generated examples while adhering to the task-specific prompts. With extensive prompt engineering (see 7), 120 examples were generated; 30 examples from each model.

The data examples generated from LLMs (Gemini, Mistral AI, Claude, and Llama) along with the shortlisted data examples from GUM corpus were integrated into a composite dataset for annotations.

3.1.1 Annotations

To identify ambiguity surrounding pronominal references to entities or events, text examples were annotated. Dataset was annotated using Label Studio by Tkachenko et al. (2025). Each text example was critically examined and classified as ambiguous when it contained both entity and event references or as unambiguous when it contained only one reference type. To ensure an unbiased and systematic annotation process, both authors of this study independently annotated the examples. This strategy was used to mitigate individual bias. The custom labeling setup involved rating examples on an 11-point scale ranging from 100% entity-leaning to 100% event-leaning. Annotators labeled and rated each example: Figure 1 illustrates the star icon used for annotations. Annotators used their contextual understanding and linguistic understanding while following annotation guidelines.

We initially calculated the inter-annotator agree-

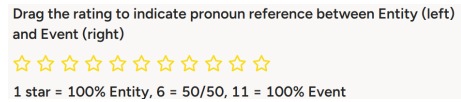


Figure 1: Star rating icon; the first four stars starting from the left represent pronoun reference leaning towards entity while the first star represents 100% entity reference, the three stars in the middle denote ambiguity, the last four stars represent pronoun reference leaning towards event, with the last and eleventh star indicating 100% event.

ment while correcting for chance agreement by using Cohen’s Kappa by Cohen (1960) and ordinal Krippendorff’s alpha by krippendorff (2004). This was followed by annotators adopting an adjudication approach. Both annotators jointly reviewed all cases of annotation disagreements and resolved them through systematic discussion, resulting in consensus for each case. This process led to a revised annotated dataset, devoid of inter-annotator disagreements and was subsequently used as final training and test data for the model.

An example of an ambiguous case is as follows: *”The dog barked at the mailman. It startled the children*”. Here, either the dog (entity) or ‘the barking of dog’ (event), startled the children, hence entity and event probability both receive values of 0.5. Probabilities are categorized for a simple annotation process; 0.1-0.39 represents entity-leaning while remaining 0.69-0.9 represent event-leaning and vice-versa, and 0.4-0.6 represent Ambiguous cases. These five labels were later condensed into three labels: entity leaning, event-leaning and ambiguous by removing ‘entity’ and ‘event’ labels. Three-label scheme categorizes probabilities in ranges of: 0-0.39 for entity-leaning and the remaining 0.69-1 for event-leaning. Refer to appendix 7 for an example. This approach introduced simplicity in labeling examples and subsequently helped mediate inter-annotator disagreement. In view of computation, merging exact entity-event categories into entity-event-leaning categories reduces sparsity. An overview of example categories in dataset is illustrated in figure 1.

Category	Number of Text Examples
Entity Leaning	127
Ambiguous	69
Event Leaning	53
Total	249

Table 1: Distribution of annotated examples across categories in the EmbiText.

3.1.2 Model Training

Transformer-based RoBERTa by Liu et al. (2019), was fine-tuned on EmbiText to test its interpretability when the pronominal reference potentially leads to multiple interpretations. The model then predicts the probability of the entity or an event, and the complementary probability is calculated as 1 minus the predicted probability, e.g. $p(\text{entity}) = 1 - p(\text{event})$ and vice versa. For instance, if the entity prediction is 0.72, the corresponding event prediction is 0.28.

We used the HuggingFace Transformers framework to fine-tune the model. The best model checkpoint was chosen on the basis of the validation loss. Appendices 7 and 7 show the input features and hyperparameter configuration. This hyperparameter configuration is widely used as the RoBERTa fine-tuning setting for small datasets, e.g. Wolf et al. (2020), resulting in stable convergence without overfitting.

The model uses sigmoid activation function to provide a probabilistic illustration of whether a pronoun refers to an entity or an event. Probability tokens are the targets for the model to compute loss using Mean Squared error (MSE) to project the difference between predictions and ground-truth probabilities by penalizing large deviations to train the model on necessary fine-grained contextual cues. Probabilistic outcomes enable the model to project ambiguity and degrees of entity/event-leaning instead of predicting binary choices. The output was post-processed using a threshold function to map probabilities to categories of: 1) Entity-leaning, 2) Ambiguous, 3) Event-leaning.

To evaluate the system’s performance, we compared it with an instruction-tuned baseline language model, Flan-T5, encoder-decoder-based system by Chung et al. (2022). It has shown promising performance across zero- and few-shot prompts setup. We conducted few-shot prompting by including a random sample of training examples. The model was responsible for generating output in the form of probability estimates for pronominal references to entities or events. The generated probabilities were categorized using the same thresholding method used to fine-tune RoBERTa.

4 Results

Our results demonstrate that EmbiText effectively embraces pronominal ambiguity, supporting its relevance for human-computer interactions. The over-

all results highlight the reasonable quality of the annotations of the data for this experiment. The annotators reviewed their annotated examples and disagreed on approximately 15% of the total examples. Cohen’s Kappa evaluation on the initial five-label scheme resulted in a fair agreement according to (Landis and Koch, 1977) but constrained consensus between both annotators with a value of 0.24. After reducing the labels to three, Cohen’s value increased to 0.36, highlighting improvement and fair agreement (Landis and Koch, 1977). Similarly, the ordinal Krippendorff’s alpha value improved from 0.31 to 0.46. Despite the improvement, the score still hints at low inter-annotator agreement, reflecting the subjectivity and difficult nature of differentiating between ambiguous and entity-event-leaning pronominal references. The initial five-label scheme in Appendix 7, shows strong agreement on ”ambiguous” cases but prominent disagreement on cases between ”Entity”, ”Entity leaning” and ”Ambiguous”. Contrastingly, as observed in figure 2, three-label scheme clarifies that the disagreement primarily is prominent between entity-leaning and the customized labeling setup involved rating examples on an 11-point scale ranging from 100% entity-leaning to 100% event-leaning. Annotator 1 labeled more examples as ambiguous, while Annotator 2 leaned towards entity-specific labels. See Appendix 7 for cases of inter-annotator disagreement and their resolution. This enabled the annotators to resolve the discrepancies through systematic discussions of each conflicted example until a common ground was established. Subsequently, the reconciled annotations were used as the final data to train and evaluate the model.

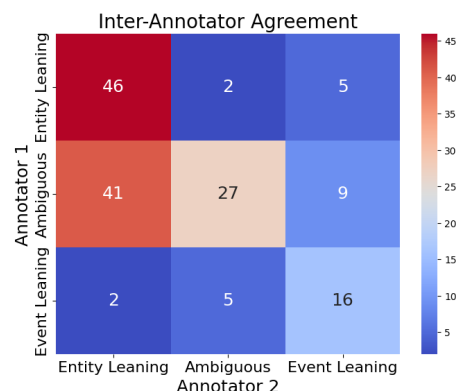


Figure 2: Confusion Matrix denoting inter-annotator agreement using three-labels scheme.

Findings from text generation experiment demonstrate that LLMs are capable of generating text ex-

amples that are ambiguous in nature. Figure 3 illustrates the superiority in performance of Llama, followed by mistral AI, however 60% of its output displayed a negative tone, emphasizing disasters, death and destruction despite the inclusion of positive and neutral examples in the prompt. This suggests that the output represents a subset of the distribution of possible examples. For example: "The tsunami hit the shore with huge waves. It caused widespread destruction and loss of life". Claude and Gemini demonstrated underwhelming performance.

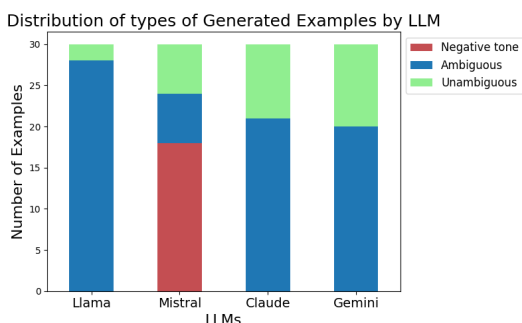


Figure 3: Overview of generated examples from LLMs.

Fine-tuning RoBERTa on this data showcased a lower rate of prediction errors: MSE of 0.019 (entity) and 0.1328 (event), RMSE of 0.3029 (entity), 0.3644 (event), and MAE of 0.2573 (entity) and 0.3119 (event), indicating more accurate results for entity references and overall reflecting a low value of average prediction error and deviation of predictions from ground truth values. Table 2 presents the comparison of our system (Fine-tunes RoBERTa) with baseline (Flan-T5). Our model yielded a lower error rates (MSE and RMSE and more accurate performance as compared to Flan-T5. This detail suggests that model succeeded in predicting probabilities values closer to true probability values.

Metric	Flan-T5 (few-shot)	RoBERTa (fine-tuned)
MSE	0.1063	0.0847
RMSE	0.3260	0.2911
Accuracy	0.314	0.353
Macro F1	0.096	0.205

Table 2: Comparison of the baseline (Flan-T5, few-shot) and our system (fine-tuned RoBERTa) on the test set, predicting entity probabilities.

5 Discussion and Conclusion

Our results corroborate that the curated EmbiText dataset and fine-tuned LLMs efficiently model nat-

ural ambiguity, especially in cases in common language where resolution is challenging. The proposed dataset demonstrate linguistic relevance and careful annotation approaches with systematic reconciliation of inter-annotator disagreements to mitigate bias and subjectivity. During annotation, some ambiguous examples featured event spans that elaborated on entities rather than individual actions i.e. "Tomorrow, when this image is shared with the world, it will be a historic moment for science and technology". The baseline comparison revealed that fine-tuned RoBERTa produced an improved probability calibration with a balanced distribution of categories, while Flan-T5, despite strong predictions, reflected a slight bias toward *Ambiguous* category.

Data	it	that	which	this
LLM-generated examples	108	0	0	0
GUM Corpus	28	13	7	11
Total	136	13	7	11

Table 3: Counts of selected pronouns across data examples.

The examples resulting from text generation experiments reflect the ambiguity found in natural language, where pronouns can refer to multiple antecedents and visualize multiple contextual interpretations. The results from the fine-tuned RoBERTa configuration suggest that the curated dataset accommodates referential ambiguity while distinguishing between entity and event references rather than resolving it. LLM generated text only included the pronoun "it", despite the prompts including other pronouns, suggesting further prompt refinement as seen in Figure 3. This demonstrates the ability of modern AI systems to interpret syntactic and semantic ambiguity in ways that project human-like sensitivity to multiple contexts through prompt engineering and fine-tuning. Although LLM performance is not equivalent to human cognition, the results support the second goal of this study: modeling ambiguity as a linguistic feature rather than resolving it. This study focuses on embracing ambiguity as a feature that uncovers deeper textual interpretations rather than a flaw, something previous research had neglected. This study contributes to applications involving human-computer interaction, i.e. customer service bots, dialogue systems, and assistive technologies.

6 Limitations

Our focus is primarily on the ambiguity arising from the pronominal reference between entity-event in English-specific text examples. Despite a fair agreement value and consensus-based resolution of disagreements, perception of ambiguity remains subjective, reflecting the level of difficulty of this task. Limited data size and label imbalance can cause differences between entity and event results. Additionally, the baseline Flan-T5 model is an instruction-tuned sequence-to-sequence model which makes probability prediction and classification tasks less direct as compared with encoder-only RoBERTa. Future direction should expand data size to enhance model generalizability, apply resampling techniques (i.e. SMOTE), involve cross-lingual analysis, enhanced prompt engineering techniques for text generation, employ other pretrained models as baseline models to compare with, multiple annotators (4-6) and multiple evaluators to evaluate generated examples for robust assessment and test different model architectures.

References

- M. K. Anjali and P. Babu Anto. 2014. [Ambiguities in natural language processing](#). *International Journal of Innovative Research in Computer and Communication Engineering*, 2:392–394. Accessed: 2025-05-05.
- Anthropic. 2024. [Claude 3 haiku \(version: claude-3-haiku-20240307\)](#). Large language model developed by Anthropic.
- David Bamman, Olivia Lewke, and Anya Mansoor. 2020. [An annotated dataset of coreference in english literature](#). *Preprint*, arXiv:1912.01140. Accessed: 2025-07-20.
- Luca Bevacqua, Sharid Loáiciga, Hannah Rohde, and Christian Hardmeier. 2021. [Event and entity coreference across five languages: Effects of context and referring expression](#). *Dialogue & Discourse*, 12(2):192–226. Accessed: 2025-09-15.
- Hong Chen, Zhenhua Fan, Hao Lu, Alan L. Yuille, and Shu Rong. 2018. [Preco: A large-scale dataset in preschool vocabulary for coreference resolution](#). *Preprint*, arXiv:1810.09807. Accessed: 2025-07-20.
- Ephraim Chukwu. 2015. [Understanding linguistic ambiguities for the effective use of english](#). *Awka Journal of English Language and Literary Studies*, 6. Accessed: 2025-05-09.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416. Accessed: 2025-09-18.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37 – 46. Accessed: 2025-09-18.
- Marie Duzi. 2013. [Ambiguities in natural language and ontological proofs.](#), pages 179–218. Accessed: 2025-05-05.
- Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. [The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics. Accessed: 2025-09-15.
- Google DeepMind. 2023. [Gemini](#). Accessed: 2025-04-17.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783. Accessed: 2025-06-02.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L el io Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825. Accessed: 2025-04-21.
- Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. [BERT for coreference resolution: Baselines and analysis](#). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Accessed: 2025-05-05.
- klaus krippendorff. 2004. [Measuring the reliability of qualitative text analysis data](#). *Departmental Papers (ASC)*, 38. Accessed: 2025-09-18.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174. Accessed: 2025-09-15.
- Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. [Cross-lingual incongruences in the annotation of coreference](#). In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora*

- and Coreference, pages 26–34, Minneapolis, USA. Association for Computational Linguistics. Accessed: 2025-09-18.
- Nghia T. Le, Fan Bai, and Alan Ritter. 2022. Few-shot anaphora resolution in scientific protocols via mixtures of in-context experts. *Preprint*, arXiv:2210.03690. Accessed: 2025-05-10.
- 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. Accessed: 2025-05-25.
- Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun ‘it’. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark. Association for Computational Linguistics. Accessed: 2025-05-28.
- Sharid Loáiciga, Christian Hardmeier, and Asad Sayeed. 2020. Exploiting cross-lingual hints to discover event pronouns. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 99–103, Marseille, France. European Language Resources Association. Accessed: 2025-05-28.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2025. *Label Studio: Data labeling software*. Open source software available from <https://github.com/HumanSignal/label-studio>.
- Kangda Wei, Xi Shi, Jonathan Tong, Sai Ramana Reddy, Anandhavelu Natarajan, Rajiv Jain, Aparna Garimella, and Ruihong Huang. 2025. *Legalcore: A dataset for event coreference resolution in legal documents*. *Preprint*, arXiv:2502.12509.
- Ralph Weischedel, Mitch Marcus, Martha Palmer, Eduard Hovy, Robert Belvin, Sameer Pradhan, and Lance Ramshaw. 2010. Ontonotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation*. Springer, New York, NY. Accessed: 2025-07-27.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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. 2020. *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. Accessed: 2025-09-05.
- Yuewei Yuan, Chaitanya Malaviya, and Mark Yatskar. 2023. *Ambicoref: Evaluating human and model sensitivity to ambiguous coreference*. *Preprint*, arXiv:2302.00762. Accessed: 2025-07-20.
- Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. *eRST: A signaled graph theory of discourse relations and organization*. *Computational Linguistics*, 51(1):23–72. Accessed: 2025-02-10.

7 Ethical Considerations

AI tools were used to assist with the polishing and enhancing the writing of this paper i.e: > - checking grammar, > - Improving and shortening text. AI assistive tools such as OpenAI’s ChatGpt and Copilot were leveraged in: > - Debugging Synthetic Text generation task. > - Debugging and refining code for data preprocessing and visualization, model training loop and evaluation practices. Importantly, all experimental designs, data annotation, model evaluation and interpretations were conducted independently by the author. The disclosure of usage of AI tools is in the interests of transparency in the research process and academic integrity.

A. Input Features

Train Dataset	Test Dataset
text_example	text_example
pronoun	pronoun
entity_candidate	entity_candidate
event_candidate	event_candidate
entity_prob	
event_prob	

Table 4: Overview of train and test set features.

B. Hyperparameters

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Number of Epochs	10
Early Stopping Patience	2
Dropout Rate	0.3
Weight Decay	0.01
Gradient Clipping	0.01

Table 5: Hyperparameter configuration for RoBERTa fine-tuning.

C. Prompt for Ambiguous text examples Generation using LLMs

prompt = "" Generate 30 ambiguous sentences where a pronoun could refer to either an entity or an event. Here are some examples: 'The volcano erupted violently. It created a huge crater.' 'The bird sang perfectly. It made everyone in the room very happy.' 'Garden was blooming with flowers, which made me feel refreshed.' Now, generate more examples: ""

D. Annotation Example



E. Confusion Matrix for Five-Label scheme

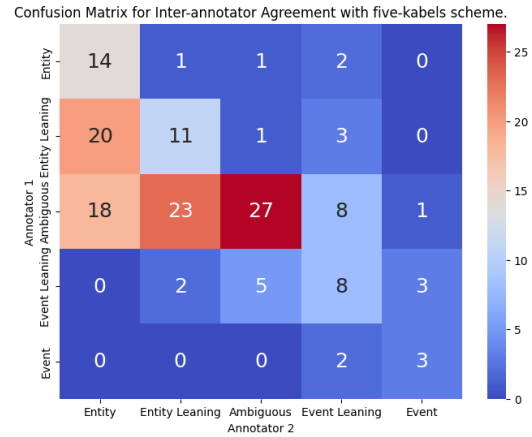


Figure 4: Confusion Matrix denoting inter-annotator agreement using five-labels scheme.

F. Annotation Disagreement and Resolution

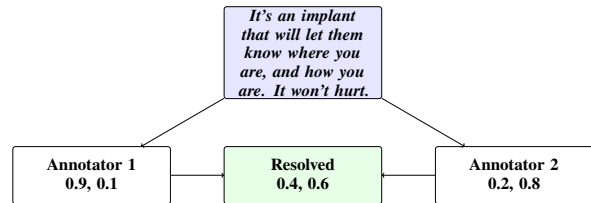


Figure 5: Annotation disagreement and resolution: unified probability distribution.

G. Annotation Disagreement and Resolution 2.0

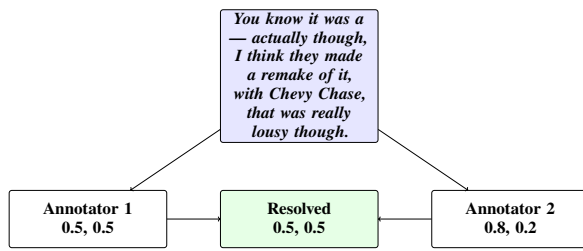
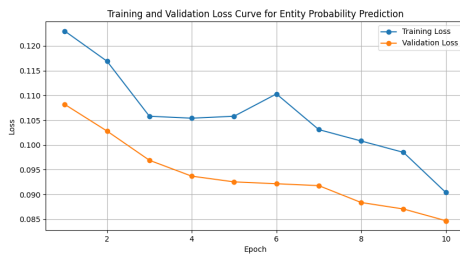
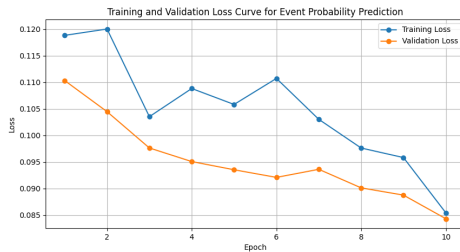


Figure 6: Annotation disagreement and resolution: unified probability distribution.

H. Loss Curve Comparison



(a) Loss curve for our model predicting entity probabilities.



(b) Loss curve for our model predicting event probabilities.

Figure 7: Training loss curves for the model predicting entities and events.

I. Metrics for Event Probabilities

Metric	Flan-T5 (few-shot)	RoBERTa (fine-tuned)
MSE	0.1063	0.0843
RMSE	0.3260	0.2903
Accuracy	0.314	0.372
Macro F1	0.096	0.214

Table 6: Comparison of the baseline (Flan-T5, few-shot) and our system (fine-tuned RoBERTa) on the test set, predicting event probabilities.

Human and LLM-based Assessment of Teaching Acts in Expert-led Explanatory Dialogues

Aliki Anagnostopoulou* Nils Feldhus^{1,3,4} Yi-Sheng Hsu^{*,7}
Milad Alshomary⁵ Henning Wachsmuth⁶ Daniel Sonntag^{1,2}

¹German Research Center for Artificial Intelligence (DFKI) ²Applied Artificial Intelligence, Oldenburg University

³Technische Universität Berlin ⁴BIFOLD – Berlin Institute for the Foundations of Learning and Data

⁵Data Science Institute, Columbia University ⁶Leibniz University Hannover, Institute of Artificial Intelligence

⁷Computer Science Institute, Ruhr West University of Applied Sciences

Corresponding authors: feldhus@tu-berlin.de h.wachsmuth@ai.uni-hannover.de

Abstract

Understanding the strategies that make expert-led explanations effective is a core challenge in didactics and a key goal for explainable AI. To study this computationally, we introduce ReWIRED, a large corpus of explanatory dialogues annotated by education experts with fine-grained, span-level teaching acts across five levels of explainee knowledge. We use this resource to assess the capabilities of modern language models, finding that while few-shot LLMs struggle to label these acts, fine-tuning is a highly effective methodology. Moving beyond structural annotation, we propose and validate a suite of didactic quality metrics. We demonstrate that a prompt-based evaluation using an LLM as a “judge” is required to capture how the functional quality of an explanation aligns with the learner’s expertise – a nuance missed by simpler static metrics. Together, our dataset, modeling insights, and evaluation framework provide a comprehensive methodology to bridge pedagogical principles with computational discourse analysis.

1 Introduction

Effective teaching is a masterclass in communication, where an expert dynamically adapts their language and strategy to guide a learner toward understanding. This process unfolds as a complex, structured dialogue, yet the specific discourse mechanisms that make an explanation effective, especially when tailored to different audiences, are not well understood from a computational perspective. While insights from education and psychology define what constitutes good teaching (Miller, 2019; Kulgemeyer, 2018), we lack the fine-grained datasets and evaluation frameworks needed to model these principles in natural language.

This paper addresses that gap through a multi-faceted approach, as illustrated in Figure 1. First,

*Work done while at DFKI.

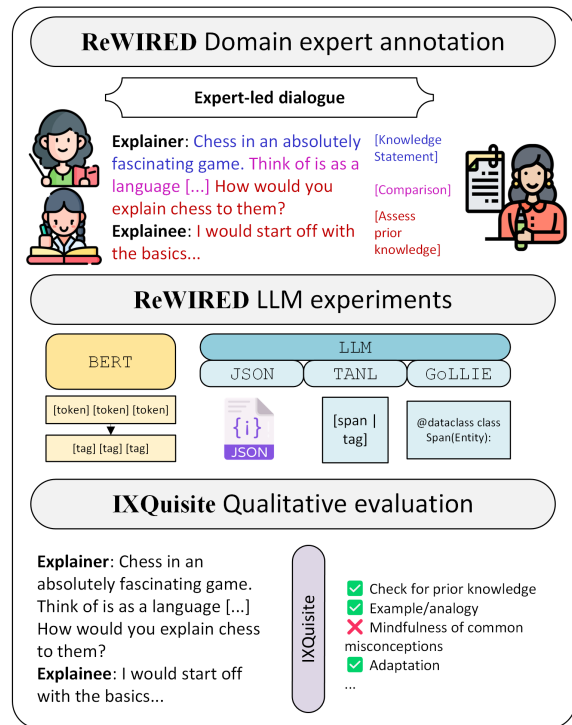


Figure 1: Our workflow: We begin by having education experts create span-level annotations of teaching acts in explanatory dialogues. We then experiment with various LLMs to automate this annotation. Finally, we conduct a qualitative evaluation, using both human experts and LLMs, to assess the quality of the explanations based on didactic principles.

we introduce ReWIRED, a new corpus resource that significantly extends the WIRED dataset (Wachsmuth and Alshomary, 2022). Our contribution lies in a new layer of **span-level annotations of teaching acts**, provided by education domain experts, across dialogues tailored to five distinct knowledge levels (from child to colleague). This provides an empirical foundation for studying pedagogical discourse structure (§3).

Second, we explore the feasibility of automating the detection of these acts. We evaluate a range of language models and prompting techniques, reveal-

ing that while few-shot LLMs struggle with this nuanced task, models fine-tuned on our data—even smaller ones—can achieve near-perfect accuracy. This establishes a robust methodology for analyzing instructional dialogues at scale (§4).

Finally, we move from structural annotation to quality assessment. We employ and extend IXQUISITE, a suite of metrics grounded in didactics, to evaluate explanation quality. We validate these metrics with our expert annotators and demonstrate that a prompt-based evaluation using LLMs as “judges” is significantly more effective at capturing the functional quality of instructional discourse than traditional static methods. This provides a new paradigm for evaluating pedagogically-aware systems (§5).

Together, these contributions – a richly annotated corpus, a validated modeling approach, and a nuanced evaluation framework – provide a comprehensive methodology for bridging educational theory with computational discourse analysis, paving the way for AI systems that can generate more effective, human-like explanatory dialogues ¹.

2 Background and related work

Instructional explanations are intended to transfer knowledge by introducing a new cognitive framework for understanding a concept or performing a task, bridging the gap between a knowledgeable individual and someone lacking that understanding. In science education, such explanations are considered both a fundamental activity and a goal of scientific practice, aimed at systematically addressing “how” and “why” questions (Kulgemeier, 2018). The authors highlight the separation of two interpretations for the term *explanation*: One is an explanation seen as activity, whose goal is to “engender understanding” between an explanation holder and an explainee; the other is a more philosophical understanding explanation, as that which connects *explanans* and *explanandum* (Zhu and Rudzicz, 2023). Although most studies concerning explainability have focused on the latter, we focus on its execution as a social, dialogical practice (Miller, 2019). In this view, the sequence of communicative acts, the choice of examples, and the adaptation to the learner are all crucial elements of the dialogue’s discourse structure.

Modeling Pedagogical Strategies with Anno-

¹The dataset, code, and test suite are available at <https://github.com/nfeInlp/InstruX>.

tation Schemata. To analyze this structure computationally, we draw from established *teaching models* from education science (Oser and Baeriswyl, 2002; Krabbe et al., 2015). These models are not just abstract theories; they provide a blueprint for effective instructional sequences. For instance, a common pattern is to first assess prior knowledge, then introduce a concept, provide an example, and finally test for understanding. We operationalize these pedagogical principles as a set of nine span-level *teaching acts* (Table 1). This approach treats teaching strategies as a form of domain-specific discourse annotation, allowing us to model the underlying functional structure of the dialogue beyond surface-level linguistics.

Corpora for Educational Dialogue and Explanation Quality. Several corpora have paved the way for analyzing educational dialogues. Datasets like CIMA (Stasaski et al., 2020), TSCG-2 (Caines et al., 2022), and NCTE (Demszky and Hill, 2023) capture teacher-student interactions, but often focus on general dialogue moves rather than the specific pedagogical functions within an explanation. The work closest to ours is the WIRED corpus (Wachsmuth and Alshomary, 2022) and its analysis by Alshomary et al. (2024), which includes annotations for high-level explanation and dialogue moves. Our work significantly extends this by: (1) doubling the dataset size; (2) providing more granular, **span-level** annotations of teaching acts rather than turn-level classifications; and (3) using **domain experts** in education for annotation, increasing the validity of the labels. This finer granularity is crucial for understanding how different teaching strategies are woven together within a single conversational turn.

Recent work has also leveraged LLMs in education, for tasks like assessing student answers (Carpenter et al., 2024) or cognitive engagement (McClure et al., 2024), and in human-AI tutoring systems (Wang et al., 2024; Jurenka et al., 2024). Evaluating the quality of these interactions remains a challenge. While some metrics focus on general dialogue quality (Mehri and Eskénazi, 2020) or textual features (McNamara et al., 2014), they often miss the pedagogical dimension. Inspired by the approach of Rooein et al. (2024), who use both static and LLM-prompted metrics for readability, we adopt and expand a suite of quality metrics to specifically assess instructional explanations, connecting discourse phenomena to didactic principles. This addresses the challenge noted by Xu et al.

Teaching Act	T. Mdl.
T01: <i>Assess Prior Knowledge</i> Checking what the student knows before starting a lesson	CB, UT
T02: <i>Lesson Proposal</i> Proposing the steps that will be taken during the lesson	UT
T03: <i>Active Experience</i> Providing the student with puzzle/question to explore; (Student:) Interacting with a mental concept	CB, UT
T04: <i>Reflection</i> Finding gaps in knowledge or inconsistencies; Asking questions about the experience or concept	PS
T05: <i>Knowledge Statement</i> Stating the concept(s) being taught via rules or facts	PS
T06: <i>Comparison</i> Considering similarities and differences between the main concept and other related topics or facts	UT
T07: <i>Generalization</i> Exploring how the concept applies to new scenarios, experiences and situations outside of the lesson topic	CB, PS
T08: <i>Test Understanding</i> Finding out if the concept previously established was received correctly and is properly understood	CB
T09: <i>Engagement Management</i> Maintaining the classroom context to facilitate effective teaching, creating rapport between teacher and student	

Table 1: Teaching acts in the ReWIRED dataset (with descriptions and their connection to a teaching model from didactics: Teaching as problem solving (**PS**), teaching as concept building (**CB**) (Krabbe et al., 2015), and unified teaching choreographies (**UT**) (Oser and Baeriswyl, 2002).

(2024) that LLMs excel at simple evaluation but struggle with complex teaching practices without proper guidance.

3 The ReWIRED dataset

To study instructional strategies in explanatory dialogues, we introduce ReWIRED, a new corpus resource featuring a novel layer of expert-provided, span-level annotations. We build upon and significantly extend an existing dataset of instructional dialogues, enriching it with annotations grounded in pedagogical theory to facilitate fine-grained discourse analysis.

3.1 Source data: Explanation dialogues

Our starting point is the WIRED corpus (Wachsmuth and Alshomary, 2022), which contains transcripts from the *5-Levels* video series². These videos provide a unique setting for discourse analysis, as they feature a domain expert explaining a complex STEM topic to five different explainees of progressively higher expertise: (1) a child, (2) a teenager, (3) an undergraduate, (4) a graduate student, and (5) a colleague (a fellow expert).

²<https://www.wired.com/video/series/5-levels>

#	Topic	#	Topic
1	Music harmony	14	Memory
2	Blockchain	15	Zero-knowledge proofs
3	Virtual reality	16	Black holes
4	Connectome	17	Quantum computing
5	Black holes	18	Quantum sensing
6	Lasers	19	Fractals
7	Sleep science	20	Internet
8	Dimensions	21	Moravecs Paradox
9	Gravity	22	Infinity
10	Computer hacking	23	Algorithms
11	Nanotechnology	24	Nuclear fusion
12	Origami	25	Time
13	Machine learning	26	Chess

Table 2: Topics in ReWIRED. 14-26 (yellow) are transcripts that were not part of the original WIRED dataset (Wachsmuth and Alshomary, 2022). The topic “black holes” is explained in two different videos, resulting in the duplicate (5, 16). Chess (26) applies distinctive knowledge levels (novice, intermediate, FIDE master, Grandmaster, and AI expert), as educational background doesn’t imply a player’s capability.

We expanded this resource by transcribing and incorporating 13 additional topics released after the original corpus’ publication, effectively doubling the dataset size. ReWIRED now comprises 130 dialogues across the 26 topics shown in Table 2. This expansion broadens the dataset’s scope and enriches the variety of linguistic phenomena available for analysis.

3.2 Annotation of Teaching Acts

The primary contribution of our work is a new layer of annotation. We argue that to model how instruction is delivered, we need annotations that are more granular than turn-level labels. Pedagogical strategies are often embedded within a single utterance or can overlap. Therefore, we adopt a **span-labeling** approach to precisely identify segments corresponding to nine distinct *teaching acts*, as defined in Table 1. This annotation scheme allows us to capture the fine-grained, and often nested, discourse structure of instructional explanations.

Annotation Process and Quality. To ensure the validity of our annotations, we recruited four annotators, all of whom hold a Master of Education degree or equivalent and have practical in-classroom teaching experience. Annotators were onboarded through a detailed process that included a written guide with definitions and examples for each act (see Appendix E and A), and a screencast

T01	12120	273	3794	1427	4559	1326	1501	550	2057
T02	362	1374	768	252	770	101	225	143	427
T03	58	34	14100	1700	1679	685	970	161	572
T04	1	0	329	3630	196	0	0	31	19
T05	1146	238	7350	5445	62280	2169	1965	1021	3411
T06	44	0	1213	2320	2770	1038	0	15	518
T07	126	22	2359	3370	3399	545	3069	126	401
T08	104	74	1970	2025	1870	748	702	3288	750
T09	605	106	2199	2546	2441	462	556	255	15096
	T01	T02	T03	T04	T05	T06	T07	T08	T09
κ	0.76	0.23	0.61	0.27	0.99	0.09	0.28	0.32	0.83

Figure 2: ReWIRED inter-annotator agreement for teaching acts on token level. For better visibility, we scale-adjust the colors by $\text{np.log1p}(\dots)^3$. Each cell shows the number of tokens for which annotators (dis)agreed on a label in a pairwise comparison. The bottom row with green and red highlights show the Fleiss’ κ per teaching act.

demonstrating the annotation tool (LABEL STUDIO (Tkachenko et al., 2020-2024)) and walking through ambiguous cases.

The full dataset was split, with each half annotated by two experts. The task proved to be challenging, reflecting the inherent subjectivity of interpreting pedagogical intent. This is visible in Figure 3, which shows how two experts can reasonably apply different labels to the same text. The resulting inter-annotator agreement is Fleiss’ $\kappa = 0.44$. While this value indicates moderate agreement, it is not unexpected for a complex discourse annotation task and highlights that human label variation can itself be an informative signal about the ambiguity of the underlying phenomena (Plank, 2022). To create a reliable gold standard, we introduced the pre-existing non-expert annotations from Feldhus et al. (2024) as a third opinion and consolidated all three label sets to adjudicate disagreements.

The final distribution of teaching acts across the five knowledge levels is shown in Figure 4. This newly annotated corpus provides a unique resource for studying how discourse strategies in explanations are adapted to listeners with varying levels of prior knowledge.

4 Experiments: Sequence-labeling acts

Having established a richly annotated dataset, a critical next step is to assess the feasibility of automating the detection of teaching acts. Automating this

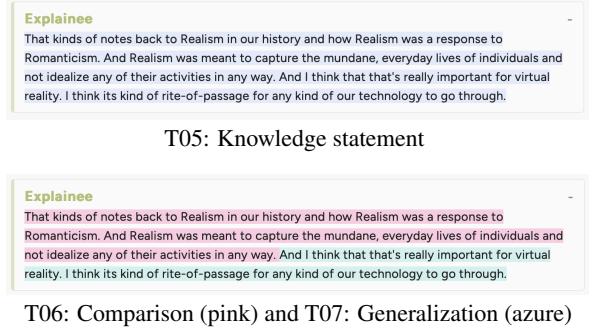


Figure 3: An example of a turn given labeled as different teaching acts by the two expert annotators.

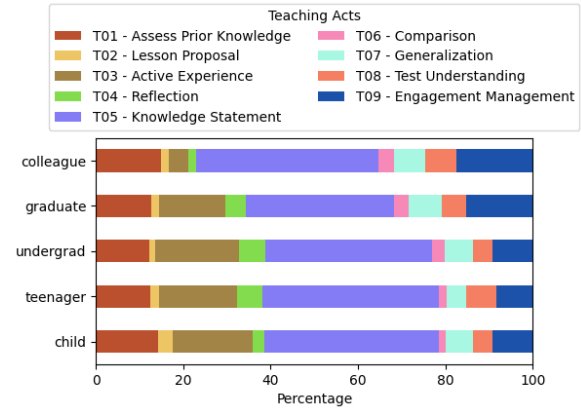


Figure 4: Distribution of teaching acts in ReWIRED across the five knowledge levels.

process is a prerequisite for analyzing instructional discourse at scale or for developing real-time assistive technologies. We therefore conduct a series of experiments to evaluate how well modern language models can perform this complex, span-level sequence labeling task.

We frame the task as structured prediction on the ReWIRED dialogues. Our evaluation compares three distinct approaches: a fine-tuned baseline model, large language models (LLMs) in a few-shot setting, and a fine-tuned LLM.

Models and Setups. As a strong baseline, we fine-tune BERT-base (Devlin et al., 2019) for token-level classification using 5-fold cross-validation, following the setup of Wachsmuth and Alshomary (2022). We then evaluate large proprietary LLMs—GPT-4o (OpenAI, 2023) and two versions of Gemini 1.5 (Reid et al., 2024)—using few-shot prompting. Finally, to directly compare the effect of fine-tuning on a modern architecture, we fine-tune GPT-4o-mini using the same 5-fold cross-validation setup. Further details on model implementation are in Appendix C.

Prompting for Structured Prediction. For the LLM experiments, we test three different prompt-

Teaching acts	T01	T02	T03	T04	T05	T06	T07	T08	T09	Macro- F_1	Span Al.
BERT FT	80.68 %	72.15 %	87.93 %	83.07 %	90.18 %	81.57 %	83.75 %	82.53 %	80.31 %	84.17 %	–
GPT-4o JSON	35.69 %	49.38 %	39.80 %	34.60 %	66.36 %	38.76 %	39.34 %	29.19 %	42.72 %	41.76 %	36.75 %
GPT-4o TANL	66.69 %	70.39 %	63.61 %	80.22 %	84.91 %	75.10 %	75.29 %	61.96 %	70.26 %	72.05 %	68.21 %
GPT-4o GoLLIE	71.39 %	67.26 %	72.83 %	78.99 %	82.70 %	79.11 %	78.05 %	71.66 %	67.07 %	74.34 %	73.54 %
Gemini 1.5 F TANL	53.39 %	71.65 %	77.76 %	85.86 %	86.13 %	81.88 %	83.73 %	63.04 %	74.83 %	75.36 %	74.09 %
Gemini 1.5 F GoLLIE	46.17 %	45.95 %	59.33 %	69.39 %	72.82 %	64.41 %	65.47 %	47.84 %	49.89 %	57.92 %	58.80 %
Gemini 1.5 P TANL	67.11 %	74.00 %	79.97 %	79.45 %	87.18 %	81.35 %	82.03 %	53.70 %	77.51 %	75.71 %	69.81 %
Gemini 1.5 P GoLLIE	46.25 %	30.56 %	53.60 %	63.00 %	70.56 %	47.44 %	49.23 %	24.88 %	48.60 %	48.23 %	49.53 %
GPT-4o-mini FT TANL	93.64 %	97.98 %	95.23 %	99.30 %	98.90 %	99.03 %	98.64 %	97.00 %	97.28 %	97.44 %	94.63 %
GPT-4o-mini FT GoLLIE	98.54 %	98.57 %	99.11 %	98.87 %	99.56 %	98.14 %	100.0 %	99.67 %	98.91 %	99.04 %	95.49 %

Table 3: Language models evaluated on the tasks of sequence-labeling teaching acts within dialogue turns from our ReWIRED dataset. Percentages under each of the acts show micro- F_1 scores in a 3-shot or fine-tuning (FT) setting. Span Alignment (last column) refers to how well the spans extracted by LLMs align with human-annotated spans.

ing paradigms designed to elicit structured, span-level output:

- **JSON**: Requesting a list of JSON objects, each containing a text span and its predicted label (Wu et al., 2024).
- **TANL**: An inline tagging format where predictions are structured as [span | label] directly in the text (Paolini et al., 2021).
- **GoLLIE**: Generating Python-like code where spans and labels are assigned to data structures, guided by a schema provided in the prompt (Sainz et al., 2024).

GPT-4o-mini is fine-tuned with 5-fold cross-validation (same setup as BERT, but with DPO, learning rate multiplier = 1.8, epochs = 3). Details and examples of these prompts are provided in Appendix D.

4.1 Results and discussion

Our experimental results, presented in Table 3, reveal several key insights into modeling domain-specific discourse acts.

Few-shot LLMs struggle with structured output and complex acts. Without fine-tuning, LLMs find the task challenging. The **JSON** format proved particularly unreliable, frequently producing malformed output that complicated post-processing and led to poor performance. While providing few-shot examples improved output consistency, the overall results remained low. Switching to more constrained output formats like **TANL** and **GoLLIE** yielded substantial improvements, nearly doubling the Macro- F_1 for GPT-4o. This highlights that for complex structured prediction, the choice of output format is critical. Even so, performance varied substantially across models and prompting

styles, with TANL emerging as the best few-shot approach, but still lagging behind the exceptional performance of fine-tuning.

Fine-tuning is essential for high performance.

The fine-tuned BERT baseline handily outperformed all few-shot LLM configurations across nearly every teaching act. This underscores the difficulty of the task and suggests that successfully capturing nuanced, domain-specific discourse phenomena requires task-specific adaptation.

This conclusion is further reinforced by our final experiment: the fine-tuned GPT-4o-mini achieves near-perfect scores, with a Macro- F_1 of up to 99.04% and a span alignment of 95.49%. Rather than suggesting the task is trivial, this result demonstrates that **fine-tuning is the most effective and reliable paradigm for this task**. It shows that even a smaller, more efficient LLM, when properly adapted with in-domain data, can master the complexities of annotating pedagogical discourse. For practitioners seeking to automate the analysis of such dialogues, we strongly recommend fine-tuning over few-shot prompting.

5 The IXquisite test suite

While our experiments show that teaching acts can be reliably annotated with fine-tuning, the presence of individual acts does not guarantee a high-quality explanation. A good instructional dialogue must orchestrate these acts into a coherent and effective structure. Evaluating this holistic quality is challenging for standard automated metrics, which often fail to capture the nuances of conversational flow and engagement (Deriu et al., 2021).

To address this, we employ and extend IXQUISITE, a test suite of quality metrics for instructional explanations grounded in didactic research (Feldhus et al., 2024). The metrics are di-

IXQUISITE: Function metrics				
Abbr.	Category	Description	Origin	Static metric
PK	Check for prior knowledge	The teacher inquires the student about prior knowledge, background, or what their interests might be	Kulgemeyer and Schecker (2009), Leinhardt and Steele (2005)	T01
MI	Mindfulness of common misconceptions	The teacher addresses common misconceptions	Wittwer et al. (2010), Andrews et al. (2011)	T04
RE	Rule-example structure	The teacher states the abstract form of the concept being taught. Then, the teacher gives some examples to assist in understanding	Tomlinson and Hunt (1971)	T05 → T03
ER	Example-rule structure	For procedural knowledge, the teacher first provides examples and then derives the general rule from them	Champagne et al. (1982)	T03 → T05
EA	Example/Analogy connection	The teacher explains how parts of the analogy/example relate to the concept being explored	Ogborn et al. (1996), Valle and Callanan (2006)	T06
UN	Check for understanding	The teacher tests the understanding of the student	Webb et al. (1995)	T08

Table 4: Explanation and teaching acts-related measures in IXQUISITE for instructional explanation quality based on occurrences of classes from our annotation schema. The right arrow between two teaching acts in static metrics refers to passages where two different acts directly follow one another in this exact sequence.

IXQUISITE: Form metrics				
Abbr.	Category	Description	Origin	Static metric
ME	Minimal explanations	Low cognitive load, e.g. avoid redundancies (verbosity) such as introducing named entities	Black et al. (1986)	Frequency of named entities
LC	Lexical complexity	The level of difficulty associated with any given word form by a particular individual or group	Kim et al. (2016)	Frequency of difficult words
SD	Synonym density	Children are proven better aligned with consistent terminology; experts allow more synonyms	Wittwer and Ihme (2014)	Frequency of synonyms for the n terms most connected to the topic
TM	Correlation to teaching model	Correlation of teaching act order to prescribed teaching models	Oser and Baeriswyl (2002), Krabbe et al. (2015)	Edit distance between T01-T08 (asc.) and actual occurrences
AD	Adaptation	The teacher incorporates prior knowledge, misconceptions and interests and uses analogies	Wittwer et al. (2010)	Inverse frequency of synonyms in the text
RL	Readability level	Indicator of how difficult a passage is to understand	Crossley et al. (2017)	Flesch-Kincaid Grade level
CO	Coherence	How sentences relate to each other to create a logical and meaningful flow for the reader or listener	Lehman and Schraw (2002), Duffy et al. (1986)	Frequency of conjunctions and linking language

Table 5: Categories for instructional explanation quality and associated numerical measures in IXQUISITE.

vided into two categories:

- **Function metrics** assess the pedagogical structure of the dialogue. They are calculated based on the presence, frequency, or sequence of the teaching acts annotated in our dataset (e.g., measuring if a *Rule* is followed by an *Example*). These are detailed in Table 4.
- **Form metrics** evaluate linguistic and stylistic features of the explanation that impact cognitive load and readability, such as lexical complexity or coherence. These are detailed in Table 5.

We investigate this suite through three lenses: human validation, traditional static evaluation, and a novel prompt-based LLM evaluation.

5.1 Human Validation of Metrics

Before applying the metrics, we first sought to validate their relevance with our domain experts. As a follow-up task to the span annotation, we asked our four annotators to assess each of the 13 metrics for every dialogue, with reference to the descriptions provided in Table 4 and Table 5. Using a 3-point Likert scale, they rated the **presence** of each metric and its **contribution** to the explanation’s quality for the given knowledge level. This step anchors our framework in the expertise and judgment of education professionals.

The results of the annotators’ assessment of metric presence are shown in Figure 5a, based on the normalized average of the ratings. The analysis reveals a strong alignment between the perceived presence of most function metrics and the explainee’s knowledge level. For instance, *Check*

for prior knowledge (*PK*), Rule-example (*RE*), and Example-rule (*ER*) structures are rated as more present in dialogues with less expert explainees. In contrast, form-based metrics like *Adaptation (AD)*, *Readability (RL)*, and *Coherence (CO)* are consistently rated as important across all levels, indicating that they serve as foundational elements of any strong explanation.

5.2 Static vs. Prompt-based Evaluation

We then evaluated the dialogues automatically using two different methods to see how well they could replicate the nuanced judgments of our human experts.

Static Evaluation. Our first approach uses "static" or rule-based calculations. For function metrics, this involves counting the tokens in the corresponding gold-standard teaching act spans (e.g., T01 for PK). For form metrics, we use standard linguistic feature calculations like the Flesch-Kincaid grade level for readability (RL). The results, shown in Figure 5b, reveal a key limitation of this approach. While some form-based metrics (e.g., LC, SD, RL) show a clear trend across knowledge levels, the function-based metrics appear noisy and fail to show a consistent correlation. The static method seems too superficial to capture the functional quality of the instructional discourse.

Prompt-based Evaluation. To overcome these limitations, we developed a "prompt-based" evaluation framework inspired by Rooein et al. (2024). Instead of relying on simple counts, we leverage an LLM's reasoning capabilities. We prompted GPT-4o with the full dialogue and asked it to rate each metric on a scale from 0 to 10 (e.g., "On a scale from 0 to 10, how well does the explainer check for understanding?").

The results, shown in Figure 5c for the function metrics, are strikingly different from the static evaluation. The prompt-based scores align remarkably well with the human judgments from our validation step. There is a clear, graded relationship between the metric scores and the explainees' knowledge levels, especially for *PK*, *EA*, *RE*, and *ER*. This demonstrates that an LLM-based "judge" is far more capable of capturing the nuanced, functional aspects of instructional quality than simple static heuristics. For form-related metrics (Appendix F), the prompt-based scores were high and stable across levels, confirming the human assessment that these are universally important. This suggests a hybrid approach for future work: static metrics

may suffice for form, but evaluating the functional discourse structure of explanations requires the inferential power of prompt-based LLM evaluation.

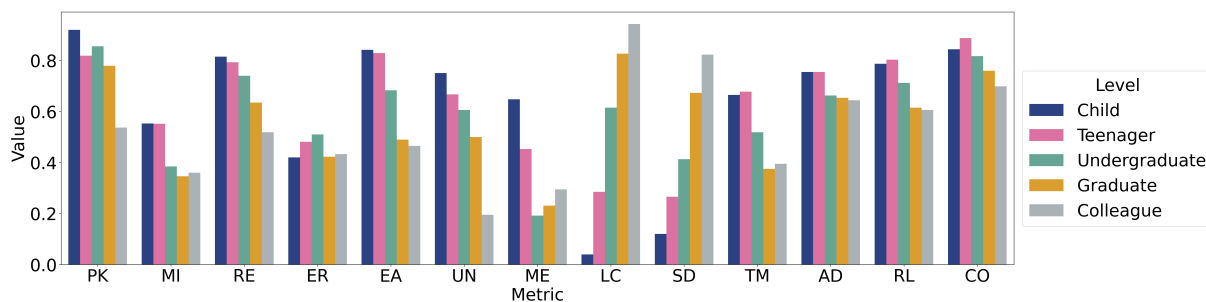
6 Discussion

Our findings offer several key implications for the fields of computational discourse analysis, educational technology, as well as NLP practices such as fine-tuning and automated evaluation.

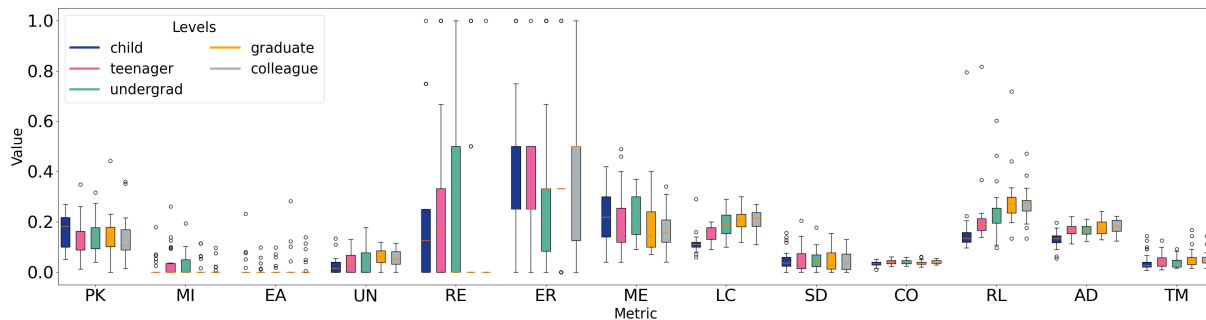
Implications for Discourse Analysis. Our work treats teaching as a complex, goal-oriented discourse phenomenon. By creating a fine-grained, span-level annotation scheme for pedagogical strategies, we provide a new lens for analyzing dialogue structure. The *teaching acts* in ReWIRED can be viewed as a domain-specific set of discourse relations that govern how instructional conversations are built. Our dataset, with its unique five-level structure of explainee expertise, offers a controlled environment to study **audience adaptation** at a granular level. Future work can analyze the typical sequences and flows of these acts to uncover the "discourse grammar" of effective explanation.

Implications for Educational Technology and XAI. Our contributions provide a direct pathway toward more effective and pedagogically-aware AI systems, e.g.:

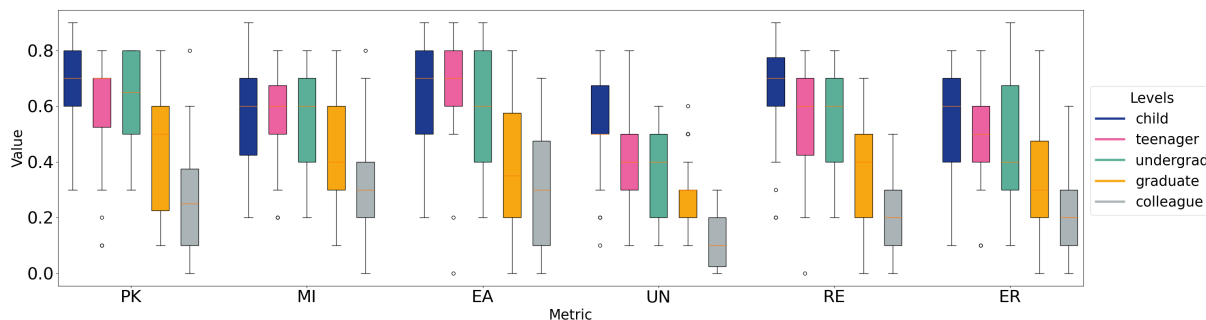
- **AI Tutors:** An automated tutor could use our models to self-assess its own dialogue strategies in real-time (Wang et al., 2024). If it produces too many 'Knowledge Statement's without a corresponding 'Check for Understanding', it could adapt its strategy to be more interactive. The IXQUISITE metrics could serve as a reward function for RL-based dialogue managers.
- **Tools for Human Educators:** Our framework could power tools that provide feedback to trainee teachers. By analyzing a transcript of a practice lesson, such a tool could highlight strengths (e.g., "Great use of analogy here!") or suggest improvements (e.g., "Consider first checking for prior knowledge.").
- **Advancing Explainable AI (XAI):** True XAI should go beyond presenting information to actively fostering human understanding. Our work offers a blueprint for pedagogically sound explanatory dialogue, shifting the focus



(a) Annotators assessment on presence of each metric in IXQUISITE for in each level.



(b) IXQUISITE metrics: Static evaluation of our dataset.



(c) IXQUISITE function-related metrics: prompt-based evaluation of the five levels in the dataset.

Figure 5: IXQUISITE results.

from producing static explanations to enabling interactive and adaptive exchanges (Feldhus et al., 2023).

Methodological Takeaways for NLP. Finally, our experiments offer two clear methodological lessons. First, for complex, domain-specific structured prediction tasks like identifying teaching acts, **in-domain fine-tuning is critical**. It vastly outperforms even the most capable few-shot LLMs, demonstrating that task-specific adaptation remains essential for high-fidelity discourse analysis. The exceptional performance can be explained with the fact that the ground truth is a consolidation from multiple annotators. The model is exposed to many examples of the already consolidated teaching acts, which is in contrast to how human annotators are typically introduced to labeling efforts,

namely with explicit instructions and few-shot examples. This is reinforced by our observation that models exposed only to few-shot examples without fine-tuning performed substantially worse.

Second, our work combines the strengths of two approaches: from the **LLM-as-a-judge** paradigm and static metrics. Our analyses suggest that for evaluating nuanced pragmatic qualities of discourse, leveraging the contextual reasoning of LLMs is a more promising path forward than relying on surface-level heuristics. However, it should be taken into consideration that, depending on the task, judge models’ agreement with human annotators can vary across datasets and domains (Bavaresco et al., 2025). In future work, applying the same principles across multiple LLMs may yield different outcomes.

7 Conclusion

In this paper, we introduced ReWIRED, a dataset of instructional dialogues significantly extending prior work with expert, span-level annotations of teaching acts. We demonstrated that while automatically labeling these acts is challenging for few-shot LLMs, fine-tuning achieves excellent performance with both smaller and larger models, establishing a reliable methodology for analyzing pedagogical discourse at scale. Furthermore, we proposed a framework for evaluating the quality of these explanations, showing that while static metrics are limited for certain dimensions, a prompt-based approach using LLMs as evaluators more effectively captures how instructional strategies are adapted to explainees’ knowledge levels.

Our contributions provide a crucial bridge between pedagogical theory and computational discourse analysis. The dataset and validated evaluation suite offer a concrete methodology for building and assessing systems that engage in instructional dialogue. This paves the way for a new generation of applications, from more adaptive and effective automated tutors to AI-powered tools that provide feedback to human educators. Ultimately, by modeling the structure of effective teaching, our work helps advance the broader goal of creating AI systems that can not only explain, but explain well.

Limitations

We acknowledge that, despite our annotators’ high expertise in the field of education, some teaching acts seem not as easily distinguishable as the other act dimensions, resulting in a relatively low inter-annotator agreement. However, the single aggregation-based Fleiss’ κ score might be too superficial to capture the complexity behind. Ultimately, the annotation variations also convey the subjectivity of teaching-related explanations, following the idea that human label variation should be encouraged (Plank, 2022).

Further limitations include that a portion of the test suite relies on human annotation, which may introduce inconsistencies. Replicating or extending the test suite might be difficult without a reliable teaching act prediction model. Also, the dataset we present is extracted from videos—audio and visual elements not present in the transcription. The efficacy of our approach may vary depending on the complexity and diversity of the multimodal inputs, if present. Last but not least, the generalizability of

our findings may be constrained by the narrow domain of dialogues examined, limiting extrapolation to broader conversational contexts.

Ethical statement

We do not see immediate ethical concerns regarding research and development. The data included in the corpus are readily available from WIRED Web resources. Following the ACM Code of Ethics (1.2, 1.6), all participants consented to be recorded as far as perceivable from the WIRED web resources, which are free to use for research purposes. The four annotators in our study were recruited over online platforms (LinkedIn, university forum). The annotation of each dialogue took an annotator an average of 10 minutes; depending on their workload, the annotation duration was between 12 and 20 hours. In our view, the provided prediction models target dimensions of dialogue turns that are not prone to misuse for ethically doubtful applications.

Acknowledgements

We thank Christoph Kulgemeyer and David Buschhüter for their input and advice on didactics, the annotators and data curators, including Kevin Kornetzki-Lucza, Iris Vassiliadi, Marie-Theres Ebner, Sören Wulff, Lavanya Govindaraju, Elif Kara, Christopher Ebert, and Emiliano Valdes Menchaca. Finally, we thank Qianli Wang for providing preprocessing code at an early stage of the project and David Schlangen and Tatiana Anikina for their feedback on earlier drafts of the paper. This work was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant numbers 03RU2U151C (news-polygraph), 01IW23002 (No-IDLE), and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg, and partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 - 438445824.

References

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. [Modeling the quality of dialogical explanations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11523–11536, Torino, Italia. ELRA and ICCL.

- Tessa M Andrews, Michael J Leonard, Clinton A Colgrove, and Steven T Kalinowski. 2011. [Active learning not associated with student learning in a random sample of college biology courses](#). *CBE—Life Sciences Education*, 10(4):394–405.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- John B. Black, John M. Carroll, and Stuart M. McGuigan. 1986. [What kind of minimal instruction manual is the most effective](#). *SIGCHI Bull.*, 18(4):159–162.
- Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. [The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts](#). In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.
- Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. [Assessing student explanations with large language models using fine-tuning and few-shot learning](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.
- Audrey B Champagne, Leopold E Klopfer, and Richard F Gunstone. 1982. [Cognitive research and the design of science instruction](#). *Educational Psychologist*, 17(1):31–53.
- Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. [Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas](#). *Discourse Processes*, 54(5-6):340–359.
- Dorottya Demszky and Heather Hill. 2023. [The NCTE transcripts: A dataset of elementary math classroom transcripts](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.
- Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. [Survey on evaluation methods for dialogue systems](#). *Artif. Intell. Rev.*, 54(1):755–810.
- 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Gerald G. Duffy, Laura R. Roehler, Michael S. Meloth, and Linda G. Vavrus. 1986. [Conceptualizing instructional explanation](#). *Teaching and Teacher Education*, 2(3):197–214.
- Nils Feldhus, Aliko Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. [Towards modeling and evaluating instructional explanations in teacher-student dialogues](#). In *Proceedings of the 2024 International Conference on Information Technology for Social Good, GoodIT '24*, page 225–230, New York, NY, USA. Association for Computing Machinery.
- Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. [InterroLang: Exploring NLP models and datasets through dialogue-based explanations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.
- Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine L. Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, and 55 others. 2024. [Towards responsible development of generative AI for education: An evaluation-driven approach](#). *CoRR*, abs/2407.12687.
- Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. [SimpleScience: Lexical simplification of scientific terminology](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071, Austin, Texas. Association for Computational Linguistics.
- Heiko Krabbe, Simon Zander, and Hans Ernst Fischer. 2015. [Lernprozessorientierte Gestaltung von Physikunterricht - Materialien zur Lehrerfortbildung](#). Waxmann.
- Christoph Kulgemeyer. 2018. [Towards a framework for effective instructional explanations in science teaching](#). *Studies in Science Education*, 54(2):109–139.
- Christoph Kulgemeyer and Horst Schecker. 2009. [Kommunikationskompetenz in der physik: Zur entwicklung eines domänenspezifischen kompetenzbegriffs](#). *Zeitschrift für Didaktik der Naturwissenschaften*, 15:131–153.

- Stephen Lehman and Gregory Schraw. 2002. [Effects of coherence and relevance on shallow and deep text processing](#). *Journal of Educational Psychology*, 94(4):738–750.
- Gaea Leinhardt and Michael D. Steele. 2005. [Seeing the complexity of standing to the side: Instructional dialogues](#). *Cognition and Instruction*, 23(1):87–163.
- Jeanne McClure, Machi Shimmei, Noboru Matsuda, and Shiyan Jiang. 2024. [Leveraging prompts in llms to overcome imbalances in complex educational text data](#). *arXiv*, abs/2407.01551.
- Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Matrix*. Cambridge University Press.
- Shikib Mehri and Maxine Eskénazi. 2020. [Unsupervised evaluation of interactive dialog with dialogpt](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235. Association for Computational Linguistics.
- Tim Miller. 2019. [Explanation in artificial intelligence: Insights from the social sciences](#). *Artificial Intelligence*, 267:1–38.
- Jon Ogborn, Gunther Kress, Isabel Martins, and Kieran McGillicuddy. 1996. *Explaining science in the classroom*. McGraw-Hill Education (UK).
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Fritz Oser and Franz Baeriswyl. 2002. *AERA’s Handbook of Research on Teaching, 4th Edition*, pages 1031–1065. Washington: American Educational Research Association (AERA).
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. [Structured prediction as translation between augmented natural languages](#). In *International Conference on Learning Representations*.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *CoRR*, abs/2403.05530.
- Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. [Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. [GoLLIE: Annotation guidelines improve zero-shot information-extraction](#). In *The Twelfth International Conference on Learning Representations*.
- Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. [CIMA: A large open access dialogue dataset for tutoring](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2024. [Label Studio: Data labeling software](#). Open source software available from <https://github.com/HumanSignal/label-studio>.
- Peter D Tomlinson and David E Hunt. 1971. [Differential effects of rule-example order as a function of learner conceptual level](#). *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 3(3):237.
- Araceli Valle and Maureen A Callanan. 2006. [Similarity comparisons and relational analogies in parent-child conversations about science topics](#). *Merrill-Palmer Quarterly (1982-)*, pages 96–124.
- Henning Wachsmuth and Milad Alshomary. 2022. [“Mama always had a way of explaining things so I could understand”: A dialogue corpus for learning to construct explanations](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Rose E. Wang, Ana T. Ribeiro, Carly Robinson, Susanna Loeb, and Dora Demszky. 2024. [Tutor CoPilot: A human-AI approach for scaling real-time expertise](#). *arXiv*, abs/2410.03017.
- Noreen M Webb, Jonathan D Troper, and Randy Fall. 1995. [Constructive activity and learning in collaborative small groups](#). *Journal of educational psychology*, 87(3):406.
- Jörg Wittwer, Matthias Nückles, Nina Landmann, and Alexander Renkl. 2010. [Can tutors be supported in giving effective explanations?](#) *Journal of Educational Psychology*, 102(1):74.
- Jörg Wittwer and Natalie Ihme. 2014. [Reading skill moderates the impact of semantic similarity and](#)

causal specificity on the coherence of explanations. *Discourse Processes*, 51(1-2):143–166.

Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. 2024. [Learning to extract structured entities using language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6817–6834, Miami, Florida, USA. Association for Computational Linguistics.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. [The promises and pitfalls of using language models to measure instruction quality in education](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4375–4389. Association for Computational Linguistics.

Zining Zhu and Frank Rudzicz. 2023. [Measuring information in text explanations](#). *CoRR*, abs/2310.04557.

Appendix

A Examples for acts

Figure 6 shows examples from ReWIRED for each of the acts as provided to the annotators.

B Label distributions

Figure 9 shows the number of distinct acts per dialogue turn as per annotated.

C Models

Table 6 lists how the models in §4 were employed. We used the following GPUs: A100, RTXA6000, RTX3080. For the BERT fine-tuning, we reinitialized the BERT model for token classification at the start of every fold ($k = 5$) and used a batch size of 4, an AdamW optimizer with a learning rate of $5 * 10^{-6}$, epsilon of $1 * 10^{-8}$, and warmup.

D Prompt design

Figure 10 and Figure 11 depict the prompts used with LLMs such as GPT-4o to produce the predictions whose evaluation is shown in Table 3. For few-shot demonstrations, we first presented the three preceding turns of the same dialogue (or from the end of last dialogue if the turn in question is at the start of a dialogue) and their corresponding gold spans (in the format required by the respective prompting paradigm) just as we elicit it from the model in the zero-shot setup. Figure 12 and Figure 13 show the results from GoLLIE and TANL prompts for Gemini 1.5 Pro and GPT-4o, respectively.

E Annotation instructions

To annotators, we provided examples from Appendix A as well as further delineations of the acts with examples and descriptions of how to differentiate between them. We also provided a screencast with instructions on how to use LABEL STUDIO and walk-through examples for each act. The introductory text shown to all annotators before watching the recording and accessing LABEL STUDIO is the following (unformatted version):

Your objective is annotating linguistic information about the multi-layered objectives each person performs when communicating. The dataset is comprised of transcribed conversations in which an expert in a field explains some concept to multiple people at varying levels of education: child, teenager, undergraduate, graduate and expert.

Your task as an annotator will be, given a transcript of one of these conversations, to use a highlighting tool to mark which “acts” are present in different parts of the text. These acts highlight some unspoken objectives present in the text. For example, the text “Do you understand that?” could be said to have both an objective of asking a yes/no question and checking for understanding.

Some of these will be straightforward to label and say “that is clearly the intention behind that sentence”, while some will be a bit more complicated. We often have many intentions behind what we say, and we account for that by letting you tag any segment of text with as many labels as you see fit, even none at all.

Your annotation task is about labeling the aforementioned objectives from the perspective of Teaching Acts, which focus on conversation mechanics in terms of lesson planning and didactics.

Model name	#Params	URL	Training times	Inference times
BERT	110M	https://huggingface.co/bert-base-uncased	13 hours	<1 hour
GPT-4o-mini (fine-tuned)	?	https://platform.openai.com/docs/guides/fine-tuning	6 hours	6 hours
GPT-4o	?	https://platform.openai.com/docs/api-reference/chat	n.a.	9 hours
Gemini 1.5	?	https://ai.google.dev/gemini-api/docs	n.a.	11 hours

Table 6: Language models with parameter counts, training times, inference times, and API costs.

fractals are really nice for computer graphics is because the algorithms that we use to draw images also have this kind of recursive flavor. [What's recursion?](#)
•T01 - Assess...

Undergrad: [Recursion is a function that uses itself or calls itself in its definition. And basically with that, you can figure out minute details such as searching for a value in](#)

(a) T01: Assess Prior Knowledge

Explainer: [So here's some toys. We're gonna build some dimensions, right? So what](#)
•T03 - Active...

[would you say about this?](#)

Child: [That's one dimensional.](#)
•T03 - Active...

(c) T03: Active Experience

Explainer: [When we were much smaller societies, you and I could trade in our community pretty easily. As the distance in our trade grew, we ended up inventing institutions, right? If you Uber or you use Airbnb or you use Amazon even, these are](#)

(e) T05: Knowledge Statement

Explainer

That's right. And we could live there. The world we see around us, the three dimensions of space around us could reflect the fact that we are somehow stuck on a three dimensional brane trying to escape.

(g) T07: Generalization

Explainer: [We're gonna talk about some science. Do you like science?](#)
•T02 - Lesson... •T09 - Engage...

Child: [Yes, a lot.](#)
•T02 - Lesson...

(b) T02: Lesson Proposal

Explainer: [Exactly. It's not really one dimensional, right?](#)
•T03 - Active...

Child: [So everything has to be one or two dimensional before it's three dimensional.](#)
•T04 - Reflec...

(d) T04: Reflection

Undergrad: [How long does this process take?](#)
•T06 - Compar...

Explainer: [Well, because people who really need to use these subdivision services for everything, people who worked hard over the years to make this super, super fast, in](#)

(f) T06: Comparison

Explainer

It's even better. It's the theory of everything. What would you tell a friend of yours if they asked you what dimensions are, what extra dimensions are, what a brane is?

(h) T08: Test Understanding (vermillion) and T05: Knowledge Statement (blue)

Explainer: [That was awesome, Daniel, thank you.](#)
•T09 - Engage...

(i) T09: Engagement Management

Figure 6: Examples for teaching acts T01-T09.

F IXQusite: additional information

F.1 Annotator’s assessment of contribution of metrics in each level

Besides validating the presence of each IXQUSITE metric in every dialogue, annotators were additionally asked to assess their importance/contribution, especially in regards to the level of knowledge of the explainee. Figure 7 shows the annotator’s assessment of the importance/contribution of each metric at each level.

F.2 Form metrics: prompt-based evaluation

Figure 8 presents the results of the prompt-based evaluation of the form metrics in the dataset. The results do not exhibit a clear correlation with the five levels, predominantly falling within the range of 0.8 to 0.9. This may be attributed to the formulation of the prompts. This might be related to the way the prompts were formulated.

F.3 Prompt-based metric questions

Table 7 shows the metrics formulated as questions for prompt-based evaluation of the explanatory dialogues in the ReWIRED dataset according to the IXQUSITE test suite.

Abbr.	On a scale from 0 to 10...
PK	... how well does the explainer inquire about prior knowledge?
MI	... how well does the explainer deal with common misconceptions?
RE	... how well does the explainer state the abstract form of a statement and then some example to assist understanding?
ER	... how well does the explainer provide examples prior to deriving a rule?
EA	... how well does the explainer explain ... how parts of the analogy/example relate to the concept being explored?
UN	... how well does the explainer check the understanding of the student?
ME	... how appropriate is the cognitive load for the explainee’s level?
LC	... how appropriate is the lexical complexity for the explainee’s level?
SD	... how appropriate is the amount of synonyms and technical language used for the explainee’s level?
AD	... how well-adapted is the content of the dialogue to the explainee?
RG	... how appropriate is the readability level for the explainee’s level?
CO	... how appropriate is the number of conjunction and subordination for the explainee’s level?
TM	... how coherent is the text for the explainee’s level?"

Table 7: IXQUSITE metrics formulated as questions for prompt-based dialogue evaluation.

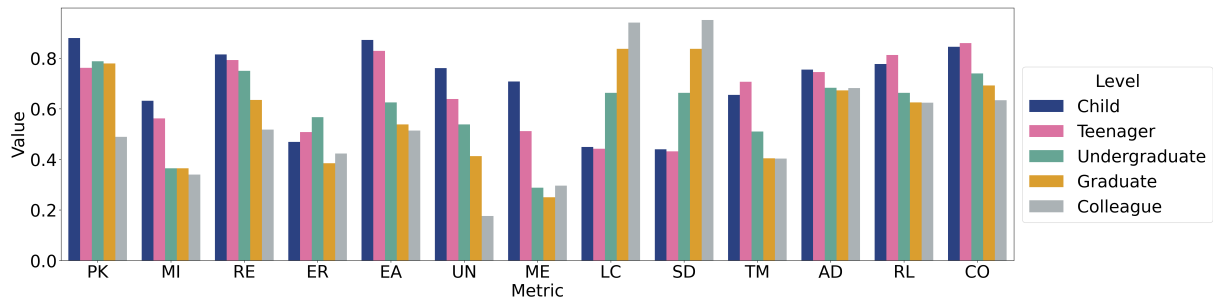


Figure 7: Annotators assessment on contribution of each metric present in IXQUISITE for each level.

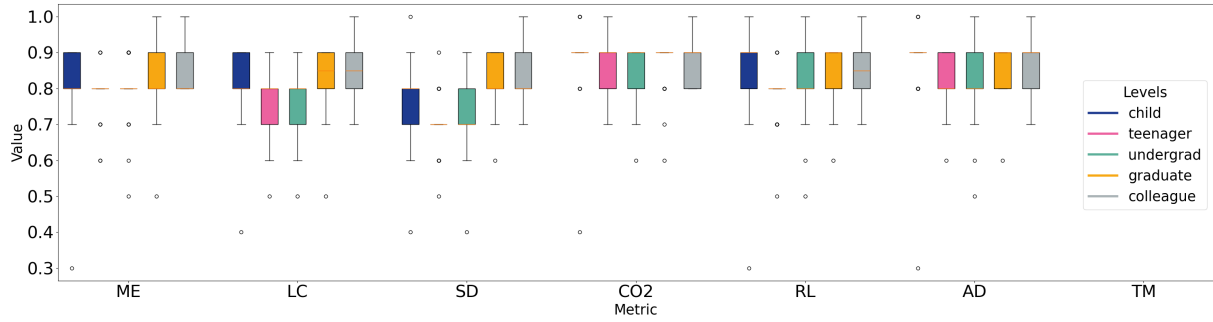


Figure 8: IXQUISITE form metrics: prompt-based evaluation of the five levels in the dataset.

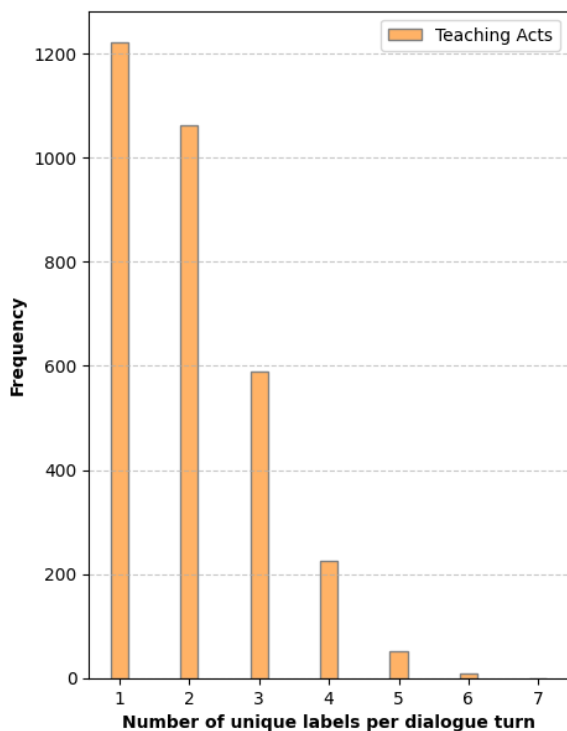


Figure 9: Number of unique teaching acts per turn in ReWIRED. The bar chart reveals that more than half of all dialogue turns in ReWIRED contain more than one distinct teaching act.

```

1 # Example label mapping (dialogue acts)
2 ReWIRED_ta_str_2_int = {
3     'T01 - Assess Prior Knowledge': 1,
4     'T02 - Lesson Proposal': 2,
5     'T03 - Active Experience': 3,
6     'T04 - Reflection': 4,
7     'T05 - Knowledge Statement': 5,
8     'T06 - Comparison': 6,
9     'T07 - Generalization': 7,
10    'T08 - Test Understanding': 8,
11    'T09 - Engagement Management': 9,
12    'T10 - Other Act': 0
13 }
14 label_schema = ("The label schema consists of the following 10 classes:\n* " + "\n*
↳ ".join(list(ReWIRED_ta_str_2_int.keys())) + "\n")

```

Figure 10: Label schema.

```

1 system_prompt = (f"You are an expert annotator. ")
2 read_instruction = (f"Here is one turn from a dialogue between an explainer and a {student_role}
↳ on the topic of {topic}:\n{turn_text}\n")
3
4 task_instruction_JSON = ("Please extract the spans from the turn and assign a label to each of
↳ the spans. It is possible that the whole turn is just one span, because the act applies to
↳ its entirety. Please present your predictions in a JSON format like this:
↳ {\n\t{\n\t\t'Span': '...', \n\t\t'Predicted label': '...' \n\t},\n}\n")
5 task_instruction_TANL = ("Please annotate the spans in the turn by marking them inline using the
↳ format [ span | label ]. It is possible that the whole turn is just one span if the act
↳ applies to its entirety.")
6 task_instruction_GoLLIE = ("Task: Annotate the following text with {TASK_NAME[task]}
↳ labels.\n\n'docstring += 'Guidelines:\n'docstring += '- Identify spans in the text that
↳ correspond to the following acts.\n'docstring += '- The act classes are defined below.")
7
8 entire_input = system_prompt + read_instruction + label_schema + task_instruction

```

Figure 11: Simplified version of the Python code showing the span-labeling task prompt for ReWIRED.

```

1 Text = "Explainer: \"So machine learning is a way that we teach computers to learn things about
↳ the world by looking at patterns and looking at examples of things. So can I show you an
↳ example of how a machine might learn something?\""
2
3 labels = [
4     {'span': "So machine learning is a way that we teach computers to learn things about the
↳ world by looking at patterns and looking at examples of things.", 'label':
↳ 'T05___Knowledge_Statement'},
5     {'span': "So can I show you an example of how a machine might learn something?", 'label':
↳ 'T02___Lesson_Proposal'},
6 ]

```

Figure 12: Example for a result from a GoLLIE prompt with Gemini 1.5 Pro.

```

1 "Explainer: "\"It's a lot of practice and analysis. [Really, an advanced chess player was not
↳ born an advanced chess player. They have probably hundreds, if not thousands of more games
↳ in their mind, in their past, in their history that they've analyzed, that they've studied.
↳ It's like any athlete, you know? | T07 - Generalization] [I put my weight on this foot, and
↳ so I wasn't able to hit the shot back that well. So the next time that that happens, I'm
↳ gonna be more prepared. | T06 - Comparison]\""

```

Figure 13: Example for a result from a TANL prompt with GPT-4o.

Where Frameworks (Dis)agree: A Study of Discourse Segmentation

Maciej Ogrodniczuk¹ Anna Latusek¹ Karolina Saputa^{1,†} Alina Wróblewska¹
Daniel Ziembicki¹ Bartosz Żuk¹ Martyna Lewandowska¹ Adam Okraśiński²
Paulina Rosalska³ Anna Śliwicka¹ Aleksandra Tomaszewska¹ Sebastian Żurowski³

¹Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

²Faculty of Psychology, University of Warsaw, Warsaw, Poland.

³Faculty of Humanities, Nicolaus Copernicus University, Toruń, Poland.

{m.ogrodniczuk, a.latusek, alina, d.ziembicki, b.zuk}@ipipan.waw.pl

Abstract

This study addresses the fundamental task of discourse unit detection – the critical initial step in discourse parsing. We analyze how various discourse frameworks conceptualize and structure discourse units, with a focus on their underlying taxonomies and theoretical assumptions. While approaches to discourse segmentation vary considerably, the extent to which these conceptual divergences influence practical implementations remains insufficiently studied. To address this gap, we investigate similarities and differences in segmentation across several English datasets, segmented and annotated according to distinct discourse frameworks, using a simple, rule-based heuristics. We evaluate the effectiveness of rules with respect to gold-standard segmentation, while also checking variability and cross-framework generalizability. Additionally, we conduct a manual comparison of a sample of rule-based segmentation outputs against benchmark segmentation, identifying points of convergence and divergence.

Our findings indicate that discourse frameworks align strongly at the level of segmentation: particular clauses consistently serve as the primary boundaries of discourse units. Discrepancies arise mainly in the treatment of other structures, such as adpositional phrases, appositions, interjections, and parenthesised text segments, which are inconsistently marked as separate discourse units across formalisms.

1 Introduction

Several linguistic discourse theories have been developed to model the structure and coherence of texts, such as Rhetorical Structure Theory (RST, Mann and Thompson, 1988; Taboada and Mann, 2006) and Penn Discourse Treebank (PDTB, Prasad et al., 2008). Each of them proposes a distinct set of discourse relations and discourse unit

types, as well as specific assumptions about the hierarchical or relational structures that reflect discourse organization. These theoretical frameworks form the basis for a range of practical implementations, e.g., the creation of annotated datasets and the development of discourse parsers.

Computational approaches to discourse adopt various taxonomies, but they typically divide discourse parsing into two main subtasks: the identification of discourse units and the classification of the relations between them (Braud et al., 2023). Discourse units (DUs) correspond to spans of text that convey discourse-relevant content, such as events, states, facts, and propositions. Discourse relations (DRs), in turn, connect DUs and assign labels to them based on a predefined taxonomy, with categories such as contrast, elaboration, or purpose.

In this study, we focus on the initial and foundational stage of discourse parsing – DU detection. This step is particularly critical, as it significantly impacts the overall accuracy of the subsequent discourse parsing task – discourse relation classification. We review how various discourse frameworks conceptualize and define DUs, with particular attention to the taxonomies they propose (Section 2). We then verify how these theoretical definitions are operationalized in practice by evaluating their implementations in existing datasets. To support this analysis, we propose a simple, rule-based heuristics for discourse segmentation (Section 3). We apply it to several English discourse datasets to validate its effectiveness and generalizability (Section 4). Furthermore, we conduct a detailed manual analysis of segmentation differences between datasets representing different discourse formalisms, aiming to identify where these frameworks align and where they diverge (Section 5).

Our current objective is to compare various approaches to discourse segmentation in real English datasets and to identify their commonalities and distinctions. Ultimately, we aim to develop a dis-

[†]To the memory of Karolina, tragically deceased on August 14, 2025.

course segmentation strategy that is both theoretically grounded and practically robust, i.e., consistent across various discourse frameworks and applicable to multiple languages.

Contributions

- A systematisation of discourse frameworks and their associated resources.
- An interpretable, syntax-based discourse segmentation approach.
- An analysis of segmentation inconsistencies in multiple English discourse datasets.

2 Discourse Segmentation in Various Formalisms

Based on [Marcu \(2000\)](#), [Wolf and Gibson \(2005\)](#) point out that there is no agreement on the notion of discourse unit: “discourse segments should be prosodic units ([Hirschberg and Nakatani, 1996](#)), others argue for intentional units ([Grosz and Sidner, 1986](#)), phrasal units ([Longacre, 1983](#); [Lascarides and Asher, 1993](#); [Webber et al., 1999](#)), or sentences ([Hobbs, 1985](#)).”

In general, existing discourse representation formalisms adopt one of two approaches to text segmentation: DUs are either independent of DRs, or they function only as relation arguments. The former implies that a text can contain DUs that are not part of any DR (but may be part of other coherence relations). In contrast, the latter approach limits the definition of DUs to those that participate in DRs, omitting segments that do not conform to a particular formalism.

Further questions to consider include whether DUs can overlap or form hierarchies within each formalism, whether they cover the text completely, and whether discourse markers are part of DUs. In this section, we will address each of these questions by reviewing the most prominent existing discourse representation formalisms and summarizing the findings of this review in [Table 2](#) in [Appendix A](#).

2.1 Non-implemented Discourse Formalisms

Hobbs’ Theory of Discourse Coherence (HTDC) [Hobbs \(1985\)](#) proposes an early text coherence theory that treats the connection of discourse units through discourse relations as an indicator of text coherence. This theory is closely linked to coreference, meaning subsequent sentences in a coherent

text should refer to the same entity ([Hobbs, 1979](#)). Clauses consist of predicates about the entities referred to in the text.

According to [Hobbs’s](#) theory, a DU is a sentential unit, which is also known as *a segment of discourse*. A segment of discourse is a set of clauses and other sentential units. Every clause is a sentential unit. Two segments form a segment of discourse if they are connected by a relation and assertions of predicates in both segments can be connected into one set of assertions. [Hobbs’s](#) theory is an attempt at a computable and implementable theory of discourse. However, the idea of assertions – the propositions in clauses that are asserted in constructing larger sentential units – or even the possibility to parse a clause into a set of assertions is difficult to implement ([Hobbs, 1985, 2013](#)).

Cognitive Approach to Coherence Relations (CCR) CCR, as elaborated by [Sanders et al. \(1992, 1993\)](#), represents a functional grounded framework for understanding discourse coherence. CCR emphasizes the cognitive processes and constraints that underlie how language users identify, categorize, and interpret coherence relations between discourse units.

Key aspects of the situation in CCR include the description of reality: objective relations in CCR typically pertain to DUs that describe situations or events that occur in the real world or in the world described by the text. Clues for the segmentation procedure are that DUs must be small enough to be a single information unit and interpretable on their own. In objective relations, the speaker, or author, merely reports facts and is not actively involved in constructing or evaluating the relation itself. This is evident, for example, in causal relations based on real-world causality, where one situation or event is presented as the cause of another. The concept of situation in CCR thus helps distinguish objective relations from subjective relations, which typically express the speaker’s opinion, argument, or evaluative stance. Historically, one may have referred to these as “semantic” and “pragmatic” relations, respectively.

ISO 24617:2: Dialogue Acts The ISO standard for Dialogue Acts ([Bunt et al., 2012](#)) provides a framework for semantic annotation of dialogues. It differs significantly from other discourse segmentation approaches by adopting a fundamentally functional perspective on discourse units.

The primary segmentation units are dialogue

acts, interpreted in terms of their communicative functions and semantic content, and distributed across multiple dialogue dimensions (e.g., Task Management, Feedback, Turn Management).

2.2 Discourse Frameworks Implemented as Datasets

GraphBank Wolf and Gibson (2005) present an annotation method for discourse coherence and evidence that trees are not an appropriate form for representing discourse structure. The authors argue that coherence structure can be represented in the form of a graph, in which nodes represent *discourse segments* and edges represent the discursive relations connecting these segments.

Discourse segments (aka DUs) correspond to clause units, non-restrictive relative clauses, and modifying prepositional phrases. The authors exclude complex nouns or verb phrases and restrictive relative clauses as DUs. Discourse units are typically marked by coordinating and subordinating conjunctions and punctuation marks. However, the conjunction ‘and’ does not mark the boundaries of segments if it connects nominal expressions and verb groups. As separate DUs, the authors also distinguish attributions that enable the distinction between different sources that comment on the same event. Attributions may be separated only if the attributed material is a complementizer phrase.

Prague Discourse Treebank 4.0 (PDiT) Intra- or inter-sentential discourse relations in PDiT 4.0 (Synková et al., 2024), labeled with semantic-pragmatic types, are determined by *explicitly* expressed discourse connectives that link exactly two *discourse arguments* (aka DUs). The primary and secondary discourse connectives are distinguished.

Since PDiT is built upon the Prague Dependency Treebank (Hajič et al., 2020), each DU is anchored in a single node of a tectogrammatical (deep-syntactic) tree representing a sentence, typically the root of the corresponding subtree. As a result, DUs correspond to text spans centered around a finite verb (i.e., the root node), with their boundaries determined by the extent of the subtree.

ISO 24617:8 ISO 24617-8 (International Organization for Standardization, 2016) is a standard created in 2016 for annotation of local discourse relations in any language or genre.

The basic discourse unit is *situation*, which covers any eventuality, fact, proposition, condition, belief, or dialogue act that can be realized by

a clause, nominalization, full sentence, utterance, or extended DU. The standard is deliberately neutral on span adjacency: an argument may be minimal or extended, continuous or discontinuous, provided it denotes the intended situation. Relations may be symmetric, assigning identical roles to both arguments, or asymmetric, assigning distinct roles. Crucially, relations are defined independently of the presence or absence of discourse markers.

2.3 Discourse Frameworks Implemented as Tools

Rhetorical Structure Theory (RST) Discourse units are essentially sentences, coherent fragments of text characterized by functional integrity (Mann and Thompson, 1988, 248–249). The minimum unit refers to spans: nuclei and their satellites. The nuclei are crucial for maintaining the coherence of the text. Determining the relationship between nucleus units and satellite units allows the analyst to present schemata, which can then be used to represent RST structures in trees. To define the smallest DU, later works referring to RST began to use the term *elementary discourse unit* (EDU, Carlson et al., 2001).

RST has proven to be very useful for both linguistics and natural language processing, as numerous RST parsers have been developed for various languages, e.g. Carlson et al. (2001), Hernault et al. (2010), Cardoso et al. (2011), Stede and Neumann (2014), Irukieta and Zafirain (2015). It should be noted that the assumptions of RST presented by Mann and Thompson (1988), including those concerning the segmentation of text, have been modified, sometimes significantly.

Penn Discourse Treebank (PDTB) PDTB (Prasad et al., 2008) does not explicitly formulate a definition of a DU. Instead, the framework is grounded in a lexically driven, minimal-pair approach, in which the primary focus is the discourse relation. This relation annotated between two spans of text, Arg1 and Arg2, is linked by a discourse connective (e.g. *because*, *and*, *since*), either explicitly present or implicitly inferred from the context. In the annotation process, the connective is identified first, and then the relation built around it is labeled.

The spans are usually equivalent to clauses; however, the overarching principle in PDTB is that they must be interpretable within the context of a discourse relation. As a result, they may extend beyond a single sentence or consist of only part of

a clause. Since PDTB focuses on annotating discourse connectives rather than predefined discourse units, it employs a function-based, non-overlapping segmentation with no hierarchy or nesting.

Segmented Discourse Representation Theory (SDRT) SDRT (Lascarides and Asher, 1993; Asher and Lascarides, 2005) is a framework for modeling discourse semantics that extends Discourse Representation Theory.

The most basic building blocks are called *Elementary Discourse Units* (EDUs). An EDU is defined as the smallest unit of text or dialogue that is semantically independent enough to participate in discourse relations. Typically, this corresponds to a simple clause (or subclause), a complete utterance in dialogue, or sometimes a larger phrase when it functions as a standalone informational unit. Sometimes EDUs are combined in complex discourse units (CDUs).

3 Syntax-based Discourse Segmentation Across Frameworks

3.1 Preliminaries and Rationale

A wide range of discourse frameworks exists (see Section 2), each differing significantly in its theoretical foundations and descriptive conventions. However, it remains an open question to what extent their practical implementations – specifically, annotated datasets and discourse parsers – also diverge. Do these resources reflect fundamental conceptual differences, or do they share underlying similarities? In this study, we aim to shed light on this issue by focusing on the task of discourse segmentation, which represents the initial step in discourse parsing. Using a simple, rule-based heuristics (see Section 3.2), we perform segmentation on a selection of English datasets annotated according to various discourse frameworks.

Predicative-argument structures – comprising sentence-level predicates and their associated arguments – form the backbone of meaning representation in natural language. These structures not only organize the semantics of individual sentences and clauses, but also serve as building blocks for larger discourse-level constructions. Due to their universality across languages, predicate-argument structures provide a promising foundation for identifying DUs.

We assume that boundaries of DUs align with the surface realization of predicate-argument structures. Without a doubt, these surface realizations

vary significantly across languages, influenced by factors such as word order, morphology, and accepted argument or predicate ellipses. Despite this variation, Universal Dependencies (UD, de Marneffe et al., 2021) approximate predicate-argument relations using a cross-linguistically consistent schema. This makes UD trees a practical and theoretically grounded resource for implementing a discourse segmentation method that generalizes across typologically diverse languages.

Universal Dependencies are structured around two fundamental linguistic concepts: the nominal, typically used to represent entities, and the clause, generally used to denote events and states. Clauses consist of a main predicate along with its arguments and modifiers. They may function as either independent sentences or embedded clauses (i.e. realizations of arguments or modifiers). Most of the formalisms discussed in Section 2 define discourse units roughly as clauses. Identifying clauses within UD trees makes it possible to extract corresponding DUs and thus to segment discourse.

3.2 UD-based Discourse Segmentation Rules

Building on the above-mentioned observations, and in line with prior research (Braud et al., 2017; Desai et al., 2020), we assume that discourse relations primarily hold between DUs realized as clauses. Consequently, we define a set of simple UD-based rules to identify clause structures. This clause identification serves as the foundation for detecting DU boundaries in a consistent and linguistically motivated way.

R1. Clauses UD distinguishes several dependency types that are realized as clausal structures and can be directly used for discourse segmentation. In particular, the heads of clauses are typically marked with the following types:

- *root* – the root of a sentence,
- *ccomp* – a clausal complement of a verb or adjective; a finite clause with an internal subject,
- *advcl* – an adverbial clause modifying a predicate or modifier word,
- *acl* – an adnominal clause, i.e., a finite or non-finite clause modifying a nominal (nominal postmodifier).

The full subtree headed by each of these dependency types (see Ex. (1) and (2)) corresponds to a distinct DU.

- (1) ⟨You’re so stupid_{root}⟩ ⟨thinking_{advcl}⟩ ⟨I spent_{ccomp} the night.⟩
- (2) ⟨This is a trend_{root}⟩ ⟨that bears_{acl:relacl} more scrutiny⟩ ⟨than it has received_{acl}.⟩

R2. Parataxis The *parataxis* relation captures constructions in which clauses or constituents are placed side by side without an explicit coordination or subordination structure. Each subtree governed by a *parataxis* head is treated as a distinct separate segment (Ex. (3)).

- (3) ⟨As each task becomes_{advcl} more specialized,⟩
⟨Smith noted_{parataxis},⟩ ⟨it engages_{root} less of the person.⟩

R3. Relative clauses The dependency type *acl:relcl* is used to annotate relative clauses (Ex. (4), (5)), as well as subordinate clauses introduced by a relative pronoun that simultaneously serves as an argument of the main predicate (Ex. (6)). All such constructions are treated as individual DUs.

- (4) ⟨Such a scenario may be found_{root} in different situations,⟩ ⟨including when one studies_{acl:relcl} a language in a classroom...⟩
- (5) ⟨I’m supposed_{root} to trust you every time⟩
⟨you tell_{acl:relcl} the truth.⟩
- (6) ⟨But how am I supposed_{root} to know⟩ ⟨when you’re telling_{acl:relcl} the truth?⟩

R4. Coordination Clauses that are coordinated with other clauses, such as those mentioned above, along with *parataxis* or *xcomp* (a clausal complement of a verb or adjective with an obligatorily controlled subject), are always treated as separate DUs. Similarly, clauses with elided predicates are also considered independent segments (Ex. (7)).

- (7) ⟨Finally, in some cases a gain in performance has been observed_{root}: after 1.5 years of limited exposure in one study [...],⟩ ⟨and [observed]_{ellipsis} in another study after 2 years,⟩ ⟨though [are observed]_{ellipsis} only for some abilities [...].⟩

R5. Parenthesis Parenthesized content, e.g., bibliographic references, is treated as a separate DU. In contrast, bracketed elements that function as appositions, predicate arguments or modifiers are not segmented (Ex. (8)). This distinction reflects the assumption that bracket usage in such cases carries a higher-level, likely pragmatic, interpretation rather than signaling a distinct DU.

- (8) ⟨Higher levels of proficiency (or exposure) may be associated_{root} with less attrition⟩
⟨[17_{dep}], [18], [21], [23]⟩ ⟨or even with no observed losses_{discontinuation}⟩ ⟨[21_{dep}].⟩

While other punctuation marks (i.e., periods, semicolons, and commas) can align with DU boundaries, they are not dependable cues for segmentation. They may be inconsistent or redundant, or reflect higher-level phenomena, e.g., at the pragmatic or stylistic level, rather than indicating clear DUs boundaries.

Discourse connectives, such as complementizers and subordinating conjunctions (annotated as *mark*), often introduce embedded DUs. Similarly, discourse markers (*discourse*) may start new DUs. However, at this preliminary stage, these elements are not segmented separately. Our goal is to examine how they are handled across different discourse frameworks to identify a consistent and unified approach to their treatment.

3.3 Discourse Segmentation Algorithm

A key component of the discourse segmentation algorithm involves identifying which tokens within a sentence correspond to DU heads, i.e., tokens that anchor individual DUs. The selection of head tokens is guided by:

- their dependency relation types,
- their part-of-speech tags,
- and the dependency types and part-of-speech tags of their particular dependent tokens.

Once the head tokens are identified (e.g., *likes* and *lost* in Figure 1), the next step involves determining the token span of each DU using the structure of the UD subtree headed by the identified tokens. A segment *X* consists of its head token *x* and possibly all tokens contained within the subtree rooted at *x*, denoted as T_x . However, if another head token *y*, representing a separate DU, is nested within T_x , the span of segment *X* is restricted to

tokens in T_x that are not part of the subtree rooted at y (i.e., T_y). In cases where segment X continues after the interruption caused by segment Y , a *discontinuity* discourse relation can be directly added, based on the hierarchical structure of T_x .

We adopt a rule-based approach to discourse segmentation rather than training a dedicated model or relying on large language models (LLMs). This decision is motivated by several factors. First, our objective is to develop a unified segmentation method applicable across multiple discourse datasets, annotated according to different discourse frameworks. Training a model on a single dataset would likely result in overfitting to the specific annotation conventions and discourse structure assumptions of that dataset. On the other hand, training a single model on a compilation of all available datasets introduces the risk of learning from inconsistent or conflicting instances, which may lead to poor or unpredictable performance. Furthermore, discourse segmentation models show limited generalizability, even across datasets in the same language and framework (Muller et al., 2019). Moreover, while LLMs can be fine-tuned or prompted to perform discourse segmentation, their output is often difficult to interpret without a complete manual revision or further postprocessing. In particular, it is difficult to systematically verify which syntactic or semantic cues underlie the identified DUs. Additionally, LLMs are prone to unintended alterations in the input, making them unreliable for the current study, where preserving the original text is crucial. Given these constraints, a rule-based approach offers a controlled and interpretable framework that facilitates cross-formalism comparison.

4 Evaluation

As a basic proof of concept, we evaluate our approach on English datasets from DISRPT 2023 (Braud et al., 2023) Task 1 (Treebank Segmentation). The datasets span three discourse frameworks (RST, SDRT, and DEP) across diverse domains:

- eng.dep.scidtb (Yang and Li, 2018) contains scientific abstracts,
- eng.dep.covdtb (Nishida and Matsumoto, 2022) contains COVID-19 research abstracts,
- eng.rst.rstdt (Lynn Carlson, 2002) contains Wall Street Journal articles from the Penn Treebank,

- eng.sdrt.stac (Asher et al., 2016) contains chat dialogues from the Settlers of Catan game,
- eng.rst.gum (Zeldes, 2017) contains mixed genres including essays, interviews, and on-line forum discussions.

It is important to articulate that the DEP datasets follow the RST annotation guidelines for DU segmentation, and their discourse relation set is based on PDBT.

Using the provided tokenization and dependency trees from the DISRPT repository¹, we apply our rule-based discourse segmenter to obtain discourse segmentation and report precision, recall, and F1 scores in Table 1.

We compare against two DISRPT 2023 participants: DisCut (Metheniti et al., 2023) and HITS (Liu et al., 2023). Both systems used pre-trained language models (XLM-RoBERTa for DisCut, RoBERTa-large for HITS) fine-tuned separately on each dataset. For the out-of-distribution eng.dep.covdtb dataset, which only provided dev and test splits, both teams used models trained on eng.dep.scidtb.

While our rule-based approach trails DisCut and HITS by 5.61 and 5.17 F1 points respectively, it demonstrates strong generalization ability. It achieves consistent performance across three different discourse frameworks and vastly different domains. The performance gap narrows significantly on out-of-distribution data: on eng.dep.covdtb, our system comes within just 2.01 points of HITS and 3.84 of DisCut. This convergence highlights a key advantage – our single rule-based system comes close to state-of-the-art performance without requiring data labeling and training separate model for each domain and discourse framework.

These results may suggest that the three discourse frameworks do not differ much concerning segmentation, indicating that they share common assumptions about how discourse should be divided into segments. Furthermore, the observed alignment between our rule-based segments and those in the datasets supports our intuition that DUs generally correspond to clauses. However, despite this apparent similarity, some differences remain between our rule-based segmentation heuristics and the datasets, and they should be examined more

¹<https://github.com/disrpt/sharedtask2023>

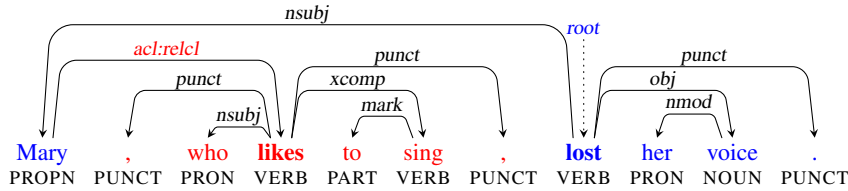


Figure 1: A dependency tree with two head tokens and the corresponding DUs highlighted in blue and red.

Dataset	Our			DisCut			HITS		
	P	R	F1	P	R	F1	P	R	F1
eng.rst.rstdt	87.56	87.00	87.28	97.21	98.04	97.62	96.46	97.66	97.06
eng.rst.gum	90.76	93.14	91.93	94.59	96.42	95.50	95.08	95.29	95.19
eng.sdrst.stac	87.92	90.37	89.13	95.75	94.70	95.22	96.71	95.09	95.89
eng.dep.scidtb	90.84	90.88	90.86	94.96	95.18	95.07	94.77	95.09	94.93
eng.dep.covdtb*	89.60	87.02	88.29	94.04	90.31	92.13	90.22	90.38	90.30
Mean	89.34	90.20	89.50	95.51	94.93	95.11	94.65	94.70	94.67

Table 1: Segmentation precision, recall and F1 score on English DISRPT 2023 datasets (Braud et al., 2023) comparing our approach to DisCut (Metheniti et al., 2023) and HITS (Liu et al., 2023) (Treebanked track). *indicates out-of-distribution datasets without training data.

closely. It is important to ask whether certain segments result from formalism-specific segmentation principles and are unique to a particular theoretical framework. Addressing these questions is essential for understanding the concept of DUs and the underlying theoretical assumptions in each discourse framework.

5 Analysis of Segmentation Discrepancies

After applying our heuristics to five different datasets, we conduct a comparative manual analysis of their samples (i.e. about 10 sentence pairs with segmentation discrepancies from each dataset). Its main goal is to identify the areas where rule-based segmentation and gold-standard segmentation diverge, assess the extent of the differences between them, and determine whether these discrepancies significantly exceed the boundaries of the DU definition we adopted.

We argued that the segmentation method – using clause boundaries as the foundation for defining DUs – is grounded in solid linguistic principles. Our analyses largely confirm this assumption, as the majority of approaches converge with our clause-level segmentation. Differences occur rather in individual cases than globally and are caused by various factors, often due to incorrect morphosyntactic annotation.

Concerning discrepancies, they are grouped into

true discrepancies, reflecting systematic differences in segmentation principles (see Section 5.1), and other discrepancies, caused by preprocessing errors and gold data inconsistencies (see Section 5.2).

5.1 True Discrepancies

R1. Adverbial clauses (*advcl*) The highest level of agreement concerns adverbial clauses modifying a predicate or modifier word, which in most datasets – similarly to our approach – are segmented. The only exception is `eng.sdrst.stac`, where adverbial clauses are very consistently not segmented (see B.1).

R1. & R3. Adnominal clauses (*acl*) and relative clauses (*acl:relcl*) A similar degree of overlap (high agreement) between our segmentation and the compared datasets can be observed in the case of all adnominal clauses, including relative clauses. However, `eng.sdrst.stac` again stands out, where such segmentation is regularly not performed (see B.2).

R1. Clausal complements (*ccomp*) The segmentation of complement clauses brings similar conclusions to the previous ones. In situations involving a verb complement, two UDs are distinguished in most datasets. The only exception shows the `eng.sdrst.stac` dataset (see B.3).

R4. Verb coordination In the case of verb coordination, there is no longer such agreement. In a larger number of datasets, when there is a single subject and a series of coordinated verbs related to it, no division into separate parts may be made. This is particularly noticeable in `eng.dep.scidtb`). Our approach always segments each of the predicates into separate DUs, no matter whether they are coordinated (see B.4).

R5. Apposition in brackets This is the area of the least agreement between our approach and the others. In RST-type datasets, such as `eng.rst.rstdt` and `eng.rst.gum`, all information in brackets, including appositions, is frequently annotated as separate DUs (see B.5, B.6). In DEP- and SDRT-type approaches, similarly to our heuristics, appositions in the brackets are not marked as separate DU.

Punctuation As we mentioned in Section 3, we do not rely on punctuation in our segmentation. However, in some datasets, colons or semicolons were used to mark DU boundaries. This is particularly evident in RST-type datasets, where a colon after an introductory phrase always signals a new unit (see B.7).

Interjections Interjections, such as 'right', 'well', 'no', 'sorry', 'hi', are marked as separate DUs in `eng.sdrt.stac`. In contrast, both `eng.erst.gum` and our approach treat them as part of an adjacent DU (see B.8).

Adpositional phrases In `eng.dep.covdtb`, gerunds (e.g. 'including', 'regarding') and gerunds followed by an adposition (e.g. 'according to') are annotated as adpositional phrases that constitute separate DUs (see B.9).

5.2 Other Discrepancies

Discourse datasets are annotated with morphosyntactic information, including sentence and token segmentation, part-of-speech tags, morphological features, and dependency trees. These annotations were either derived from existing resources or predicted automatically. Since automatic preprocessing is prone to errors, it can negatively impact the quality of discourse segmentation.

Preprocessing errors Some of the observed mismatches can be traced to inaccuracies in POS tagging or dependency parsing errors in the gold-standard data. For instance, multiple errors occur

in `eng.dep.covdtb`, including misidentification of sentence predicates (see B.10), incorrect analyses of coordination structures, erroneous assignment of part-of-speech tags, etc.

Inconsistencies in gold standard In RST-type approaches, appositions are treated as separate DUs only when enclosed in parentheses. When they are set off by commas, they remain part of the main clause. This inconsistent treatment of appositions leads to differences in segmentation between this dataset and our approach (described in the previous section).

Over-segmentation Our rules sometimes produce unnecessary splits, which becomes particularly evident when it comes to identifying and segmenting spans containing proper names, titles, or compound nouns (see B.11).

6 Conclusions

All of our analyses – initial, focusing on the most popular approaches to discourse segmentation; subsequent, examining how these theories are implemented in annotated datasets; and concluding, comparing our proposal of the UD-based discourse segmentation with already existing ones – have allowed us to preliminarily confirm the assumption that, in practice, all these approaches share a significant number of common characteristics.

Moreover, our UD-based approach to discourse segmentation, grounded in simple and clear rules, has yielded very promising results. We are aware that our rule-based approach does not reach the performance of state-of-the-art methods. However, surpassing these methods was not our primary objective. It is nevertheless worth emphasizing that segmentation based on five simple rules, applied uniformly across all tested datasets annotated according to different formalisms, approaches the performance of models trained separately for each dataset.

Our main goal was to investigate whether syntax can contribute to discourse segmentation, and the results suggest that it constitutes a key factor in identifying discourse units. This finding stands in partial contrast to the results reported by Braud et al. (2017). However, a direct comparison is difficult, given the substantially different assumptions and experimental setups. This issue, therefore, calls for further dedicated research.

Our findings encourage us to research the topic

further and set a goal of developing a universal, cross-approach method for detecting DUs. Naturally, certain areas remain that warrant particular attention. They will need to be addressed in greater depth in the next stages of our study, e.g., manual comparative analyses, in which we focused on identifying as many points of divergence in segmentation as possible, revealed discrepancies primarily at the level of segmenting certain types of subordinate clauses, parataxis, appositions or units' discontinuity. These are areas that require special attention in future work.

It should also be noted that our comparative analysis of the proposed UD-based discourse segmentation was conducted across a selection of datasets that, in our view, appeared to be the most representative. The choice of these particular datasets was motivated by our understanding of discourse as a phenomenon in which the text operates as a whole. In this perspective, discourse refers to a specific way of organizing the text holistically, where every element plays a defined role and, therefore, cannot be omitted in segmentation.

Nonetheless, we acknowledge that developing a universal approach to discourse segmentation will require further evaluations and analyses of cross-linguistic and cross-dataset similarities and differences in future work.

Key Insights

- High level of agreement on DUs segmentation across the evaluated datasets.
- Promising results of the UD-based method for DUs segmentation, demonstrating strong generalization ability.
- Potential for unifying rules for cross-linguistic discourse segmentation.

Acknowledgments

This research was funded in whole by the National Science Centre, Poland, grant 2023/50/A/HS2/00559 ("Universal Discourse: a multilingual model of discourse relations").

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. *Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus*. In *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC'16), pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. *Does syntax help discourse segmentation? Not so much*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442, Copenhagen, Denmark. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. *The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification*. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

H Bunt, J Alexandersson, J-W CHOE, Alex Chengyu FANG, K Hasida, V Petukhova, A Popescu-Belis, and D Traum. 2012. *Iso 24617-2: 2012: Language resource management—semantic annotation framework (semaf)—part 2: Dialogue acts*.

Paula C. F Cardoso, Erick G Maziero, Maria Lucía R. Castro Jorge, Eloize M. R Seno, Ariani Di Felippo, Lucia H. M Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. *Cstnews: a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese*. In *Brazilian Symposium in Information and Human Language Technology - STIL*. SBC.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. *Building a discourse-tagged corpus in the framework of rhetorical structure theory*. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16, SIGDIAL '01*, page 1–10, USA. Association for Computational Linguistics.

Ta-Chung Chi and Alexander Rudnicky. 2022. *Structured dialogue discourse parsing*. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335, Edinburgh, UK. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.

Takshak Desai, Parag Pravin Dakle, and Dan Moldovan. 2020. *Joint learning of syntactic features helps discourse segmentation*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1073–1080, Marseille, France. European Language Resources Association.

- Barbara J. Grosz and Candace L. Sidner. 1986. *Attention, Intentions, and the Structure of Discourse*. *Computational Linguistics*, 12(3):175–204.
- Grigorii Guz, Patrick Huber, and Giuseppe Carenini. 2020. *Unleashing the power of neural discourse parsers - a context and structure aware approach using large scale pretraining*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805, Barcelona, Spain. International Committee on Computational Linguistics.
- Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. *Prague dependency treebank - consolidated 1.0*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.
- Hugo Hernault, Helmut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. *HILDA: A Discourse Parser Using Support Vector Machine Classification*. *Dialogue & Discourse*, 1(3).
- Julia Hirschberg and Christine H. Nakatani. 1996. *A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues*. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Santa Cruz, California, USA. Association for Computational Linguistics.
- Jerry R. Hobbs. 1979. *Coherence and Coreference**. *Cognitive Science*, 3(1):67–90.
- Jerry R. Hobbs. 1985. *On the coherence and structure of discourse*. Technical Report CSLI-85-37, Center for the Study of Language and Information (CSLI), Stanford, CA.
- Jerry R. Hobbs. 2013. *ACL lifetime achievement award: Influences and inferences*. *Computational Linguistics*, 39(4):781–798.
- International Organization for Standardization. 2016. *ISO 24617-8:2016 Language resource management — Semantic annotation framework (SemAF) — Part 8: Semantic relations in discourse (DR-Core)*. Standard ISO 24617-8:2016, International Organization for Standardization, Geneva, Switzerland.
- Mikel Iruskieta and Beñat Zapirain. 2015. *Euseduseg: A dependency-based edu segmentation for basque*. *Proces. del Leng. Natural*, 55:41–48.
- Yangfeng Ji and Jacob Eisenstein. 2014. *Representation learning for text-level discourse parsing*. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.
- Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. *Top-down discourse parsing via sequence labelling*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.
- Alex Lascarides and Nicholas Asher. 1993. *Temporal interpretation, discourse relations and commonsense entailment*. *Linguistics and Philosophy*, 16(5):437–493.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. *A pdtb-styled end-to-end discourse parser*. Preprint, arXiv:1011.0835.
- Wei Liu, Yi Fan, and Michael Strube. 2023. *HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification*. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.
- Zhengyuan Liu and Nancy Chen. 2021. *Improving multi-party dialogue discourse parsing via domain integration*. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Robert E. Longacre. 1983. *The Grammar of Discourse*. Plenum, Mew York.
- Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. *RST Discourse Treebank LDC2002T07*.
- William C. Mann and Sandra A. Thompson. 1988. *Rhetorical Structure Theory: Toward a functional theory of text organization*. *Text & Talk*, 8:243–281.
- Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.
- Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. *DisCut and DiscReT: MELODI at DISRPT 2023*. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.
- Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. *ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents*. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.
- Noriki Nishida and Yuji Matsumoto. 2022. *Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation*. *Transactions of the Association for Computational Linguistics*, 10:127–144.
- Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembecki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024. *Polish Discourse Corpus (PDC): Corpus design, ISO-compliant annotation, data highlights, and parser development*.

- In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12829–12835, Torino, Italy. ELRA and ICCL.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. [The Penn Discourse Tree-Bank 2.0](#). In *International Conference on Language Resources and Evaluation*.
- Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1992. [Toward a taxonomy of coherence relations](#). *Discourse Processes*, 15(1):1–35.
- Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 1993. [Coherence relations in a cognitive theory of discourse representation](#). *Cognitive Linguistics*, 4(2):93–134.
- Purificação Silvano, João Cordeiro, António Leal, and Sebastião Pais. 2023. [DRIPPS: a corpus with discourse relations in perfect participial sentences](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 470–481, Vienna, Austria. NOVA CLUNL, Portugal.
- Manfred Stede and Arne Neumann. 2014. [Potsdam commentary corpus 2.0: Annotation for discourse research](#). In *International Conference on Language Resources and Evaluation*.
- Pavlna Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. [Announcing the Prague Discourse Treebank 3.0](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279, Torino, Italia. ELRA and ICCL.
- Pavlna Synková, Jiří Mírovský, Marie Pačlíková, Lucie Poláková, Magdaléna Rysová, Veronika Scheller, Jana Zdeňková, Šárka Zikánová, and Eva Hajičová. 2024. [Prague Discourse Treebank 4.0 \(PDiT 4.0\)](#).
- Maite Taboada and William C. Mann. 2006. [Rhetorical structure theory: looking back and moving ahead](#). *Discourse Studies*, 8(3):423–459.
- Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. [Llamipa: An incremental discourse parser](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.
- Jianxiang Wang and Man Lan. 2015. [A Refined End-to-End Discourse Parser](#). In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning – Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.
- Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. [Discourse Relations: A Structural and Presuppositional Account Using Lexicalised TAG](#). In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, College Park, Maryland, USA. Association for Computational Linguistics.
- Florian Wolf and Edward Gibson. 2005. [Representing discourse coherence: A corpus-based study](#). *Computational Linguistics*, 31(2):249–287.
- Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2005. [Discourse Graphbank](#). LDC Catalog No.: LDC2005T08.
- An Yang and Sujian Li. 2018. [SciDTB: Discourse dependency TreeBank for scientific abstracts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.
- Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. [RST discourse parsing with second-stage EDU-level pre-training](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.
- Amir Zeldes. 2017. [The GUM Corpus: Creating multilayer resources in the classroom](#). *Language Resources and Evaluation*, 51(3):581–612.

A Properties of Discourse Units Across Major Discourse Frameworks

Discourse Approach	DU Name	DU Form	FTC	OV	SC	I	Dataset / Parser
Hobbs’ Theory of Discourse Coherence (HTDC)	segment / sentential unit	clause / set of clauses	N	N	N	N	N
Cognitive Approach to Coherence Relations (CCR)	discourse segment	any text span, cognitively motivated	N	Y	Y	N	N
ISO 24617-2 (Dialogue acts)	functional segment	Turns, utterances, sub-utterances	Y	Y	N	Y	N
GraphBank	discourse segment	clause	Y	N	Y	Y	Y
							<ul style="list-style-type: none"> • Discourse Graphbank (Wolf et al., 2005)
Prague Discourse Treebank 4.0 (PDiT)	discourse argument	Root nodes of tectogram-matical subtrees (i.e. heads of finite clauses; text spans centred around finite verbs)	N	N	Y	Y	Y
							<ul style="list-style-type: none"> • PDiT 4.0 (Synková et al., 2024)
ISO 24617-8	situation	eventuality, fact, proposition, condition, belief or dialogue act	N	Y	Y	Y	Y
							<ul style="list-style-type: none"> • DRIPPS (Silvano et al., 2023) • PDC (Ogrodniczuk et al., 2024)

Discourse Approach	DU Name	DU Form	FTC	OV	SC	I	Parser / Dataset
Rhetorical Structure Theory (RST)	elementary discourse units (EDU)	nucleus or satellite span (essentially clauses)	Y	N	N	Y	Y <ul style="list-style-type: none"> • DPLP parser (Ji and Eisenstein, 2014), • Hilda (Hernault et al., 2010), • RST parser (Guz et al., 2020), • Top-Down parser (Koto et al., 2021), • RST parser (Yu et al., 2022)
Penn Discourse Treebank (PDTB)	argument	a minimal span of text that conveys a single discourse function (most often clauses)	N	N	Y	Y	Y <ul style="list-style-type: none"> • PDTB parser (Lin et al., 2010), • CoNLL-2015 shared task parsers (Wang and Lan, 2015)
Segmented Discourse Representation Theory (SDRT)	discourse unit (EDU/CDU)	propositions / clauses	N	Y	Y	Y	Y <ul style="list-style-type: none"> • DDP parser (Liu and Chen, 2021), • SDD parser (Chi and Rudnicky, 2022), • Llamipa (Thompson et al., 2024)

Table 2: Properties of discourse units across major discourse representation frameworks. Abbreviations: Y – Yes, N – No, FTC – Full Text Coverage; OV – DU Overlap; SC – Separate Connective (i.e. connectives treated as separate units); I – Implementability (i.e., the feasibility of implementing the framework in a dataset or a discourse parser.

B Types of Segmentation Discrepancies with Corresponding Examples

ID	Category	Dataset	Gold segmentation	Our segmentation
B.1	Adverbial clauses	eng.sdrt.stac	⟨Please only start the game when all four participants are there⟩	⟨Please only start the game⟩ ⟨when all four participants are there⟩
B.2	Adnominal clauses, relative clauses	eng.sdrt.stac	⟨I respect popular music from the time in which it was actually musical.⟩	⟨I respect popular music from the time⟩ ⟨in which it was actually musical.⟩
B.3	Relative clauses	eng.sdrt.stac	⟨I didn't know you could have them.⟩	⟨I didn't know⟩ ⟨you could have them.⟩
B.4	Verb coordination	eng.dep.scidtb	⟨We observe, identify, and detect naturally occurring signals of interestingness in click transitions on the Web between source and target documents,⟩ ⟨which we collect from commercial Web browser logs.⟩	⟨We observe,⟩ ⟨identify,⟩ ⟨and detect naturally occurring signals of interestingness in click transitions on the Web between source and target documents,⟩ ⟨which we collect from commercial Web browser logs.⟩
B.5	Parenthesis	eng.rst.gum	⟨Also beginning trading today on the Big Board are El Paso Refinery Limited Partnership , El Paso , Texas ,⟩ ⟨(ELP)⟩ ⟨and Franklin Multi-Income Trust , San Mateo , Calif. ,⟩ ⟨(FMI) .⟩	⟨Also beginning trading today on the Big Board are El Paso Refinery Limited Partnership , El Paso , Texas ,(ELP)⟩ ⟨and Franklin Multi-Income Trust , San Mateo , Calif. ,(FMI) .⟩
B.6	Appositions in brackets	eng.dep.scidtb	(...) ⟨health practices⟩ ⟨(exercise, tobacco and alcohol consumption, sleep efficiency)⟩ ⟨and genetics contribute to CLI risk.⟩	(...) ⟨health practices (exercise, tobacco and alcohol consumption, sleep efficiency) and genetics contribute to CLI risk.⟩
B.7	Punctuation	eng.rst.gum	⟨Respondents were asked to indicate their race from among the following categories:⟩ ⟨White; Black or African American; Hispanic; American Indian or Native American; and Asian or Pacific Islander.⟩	⟨Respondents were asked to indicate their race from among the following categories: White; Black or African American; Hispanic; American Indian or Native American; and Asian or Pacific Islander.⟩
B.8	Interjections	eng.sdrt.stac	⟨we 're waiting for 2 other players⟩ ⟨right⟩ ⟨no⟩ ⟨sorry⟩	⟨we 're waiting for 2 other players right⟩ ⟨no sorry⟩

B.9	Adpositional phrases	eng.dep.covid	<p>⟨Severe infections can lead to a variety of diseases,⟩</p> <p>⟨including poliomyelitis, aseptic meningitis, myocarditis and neonatal sepsis.⟩</p>	<p>⟨Severe infections can lead to a variety of diseases, including poliomyelitis, aseptic meningitis, myocarditis and neonatal sepsis.⟩</p>
B.10	Preprocessing errors	eng.dep.covid	<p>Phylogenetic analysis based on S3 gene showed that the Brazilian TReoV isolates_{s_{comp}} clustered in a single group with 98-100% similarity to TReoV strains circulating in the United States.</p>	
B.11	Compounds	eng.rst.gum	<p>⟨Among his smaller works, the seventh Humoresque and the song "Songs My Mother Taught Me" are also widely performed and recorded.⟩</p>	<p>⟨Among his smaller works, the seventh Humoresque and the song "Songs) ⟨My Mother Taught Me"⟩ ⟨are also widely performed) ⟨and recorded.⟩</p>

Table 3: Segmentation discrepancy types categorized by type of linguistic phenomenon.

Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser

Elena Chistova
FRC CSC RAS
chistova@isa.ru

Abstract

We introduce UniRST, the first unified RST-style discourse parser capable of handling 18 treebanks in 11 languages without modifying their relation inventories. To overcome inventory incompatibilities, we propose and evaluate two training strategies: *Multi-Head*, which assigns separate relation classification layer per inventory, and *Masked-Union*, which enables shared parameter training through selective label masking. We first benchmark mono-treebank parsing with a simple yet effective augmentation technique for low-resource settings. We then train a unified model and show that (1) the parameter efficient Masked-Union approach is also the strongest, and (2) UniRST outperforms 16 of 18 mono-treebank baselines, demonstrating the advantages of a single-model, multilingual end-to-end discourse parsing across diverse resources.¹

1 Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) represents discourse as a hierarchical tree of elementary discourse units (EDUs) connected by rhetorical relations. Over the years, RST has inspired the creation of multiple discourse treebanks across different languages. However, large-scale annotated corpora are scarce and predominantly available in English. For other languages, the high cost of annotation and inconsistent guidelines have resulted in smaller, heterogeneous resources with incompatible relation inventories.

The English RST Discourse Treebank (RST-DT) (Carlson et al., 2001), the primary benchmark for RST parsing, defines 56 fine-grained rhetorical relations, usually mapped to 18 coarse-grained classes for training and evaluation. Many discourse treebanks in other languages define considerably fewer

relations. Aligning them with the RST-DT inventory often requires collapsing relations, such as merging CAUSE with EFFECT, CONTRAST with CONCESSION, or ELABORATION with ENTITY-ELABORATION. This process erases distinctions that can be crucial for downstream applications such as coreference resolution, narrative analysis, and opinion mining. Moreover, when no direct equivalents exist, alignment is frequently based on surface-level label similarity, which compromises annotation reliability across languages.

End-to-end RST parsing involves three interconnected subtasks: EDU segmentation, tree structure prediction, and nuclearity and relation labeling. The definitions of these tasks are shaped by the relation inventory and constraints of each treebank. For instance, segmentation decisions can be influenced by fine-grained intra-sentential relations. Mono- or multinuclearity of certain overlapping relations (LABEL_NS, LABEL_SN, LABEL_NN) varies across treebanks. When datasets with different inventories are merged and collapsed into a coarser label set, inconsistencies in relation definitions and nuclearity distributions can introduce substantial noise into both training and evaluation.

Despite these challenges, training on multiple treebanks offers clear benefits. RST-style parsers are known to generalize poorly across domains (Liu and Zeldes, 2023), and training a unified parsing model on all available treebanks may yield broader applicability. The skewed label distributions within individual corpora complicate model training, particularly in low-resource settings; pooling datasets with overlapping labels can mitigate this issue. Although larger treebanks provide sufficient data for accurate EDU segmentation and local relation labeling, they remain too small to support robust learning of global document structures. Leveraging all annotated structures across corpora can thus strengthen structural prediction. Altogether, these considerations motivate the development of univer-

¹Our models and code: <https://github.com/tchewik/UniRST>.

sal discourse parsers that effectively integrate all available resources, regardless of language, genre, or domain.

In this work, we propose methods for building a unified RST parser from heterogeneous treebanks. Our contributions are:

1. The first large-scale RST parsing study covering 18 treebanks in 11 languages.
2. Data augmentation technique allowing for strong end-to-end mono-treebank RST parsing baselines even in low-resource settings.
3. Two strategies for jointly modeling divergent relation inventories: Multi-Head and Masked-Union.
4. Evaluations showing that: (i) dataset-specific segmentation heads are essential for handling varying EDU definitions; (ii) the Masked-Union approach enables sufficient model training by leveraging label overlap while respecting treebank-specific relation inventories, and (iii) our unified model outperforms 16 out of 18 mono-treebank baselines.

2 Related Work

Cross-Lingual RST Parsing Cross-lingual rhetorical structure parsing has gained increasing attention in recent years. Braud et al. (2017) introduced a unified set of coarse-grained (harmonized) rhetorical relations and presented the first data-driven cross-lingual RST parser, transferring across English, Brazilian Portuguese, Spanish, German, Basque, and Dutch. Their study demonstrated that rhetorical structure parsing from pre-segmented texts successfully transfers beyond English and across typologically diverse languages. Building on this foundation, Liu et al. (2020) leveraged multilingual embeddings and proposed EDU-level machine translation to enrich training data. Subsequently, Liu et al. (2021) introduced DMRST, a unified framework performing joint EDU segmentation and discourse tree parsing, enabling end-to-end RST parsing evaluation across multiple languages under harmonized inventories. Extending this line of work, Chistova (2024) applied DMRST to parallel English–Russian data, highlighting the importance of aligned corpora for assessing cross-lingual transfer in the context of RST treebank incompatibilities.

Training on Incompatible Treebanks Research on integrating incompatible treebanks has largely focused on syntax parsing. Early work by Johansson (2013) introduced two adaptation techniques for training syntax parsers on treebanks with differing annotation schemes. Their methods involved concatenating the feature spaces of two treebanks and using a parser trained on one treebank to guide the other. These approaches were applied to treebanks pairs within the same language (German, Swedish, Italian, and English). Stymne et al. (2018) explored three strategies: treebank concatenation with and without fine-tuning, and the inclusion of treebank-specific embeddings. Their results showed consistent improvements in dependency parsing for most of the nine languages evaluated when using treebank-specific embeddings. A similar approach was applied by Barry et al. (2019) to train a cross-lingual parser for low-resource Faroese syntax parsing. Johansson and Adesam (2020) trained a Swedish constituency parser on six incompatible treebanks by sharing word representations across corpora while maintaining separate neural parsing modules for each treebank, thus accommodating both constituency and dependency annotations. Kankanampati et al. (2020) leveraged two Arabic dependency treebanks to build a parser with a unified attachment scorer. Sayyed and Dakota (2021) conducted multilingual experiments with treebank-specific biaffine parsing layers for UD and SUD syntactic annotations, ultimately finding that combining distinct annotation schemes could degrade parsing performance.

Notably, in syntactic parsing, terminal nodes correspond to words, so efforts to resolve annotation inconsistencies are confined to structure building and label assignment. In contrast, rhetorical structure parsing additionally requires segmentation, which is affected by treebank-specific constraints on elementary discourse units. In our work, we aim to develop the first end-to-end RST parser benefiting from each annotation scheme in a wide range of diverse discourse treebanks.

3 UniRST

We address joint training over heterogeneous RST corpora while preserving each treebank’s native relation inventory, EDU segmentation, and relational definitions. Building on the DMRST architecture (Liu et al., 2021), we explore extensions that enable training across incompatible tree-

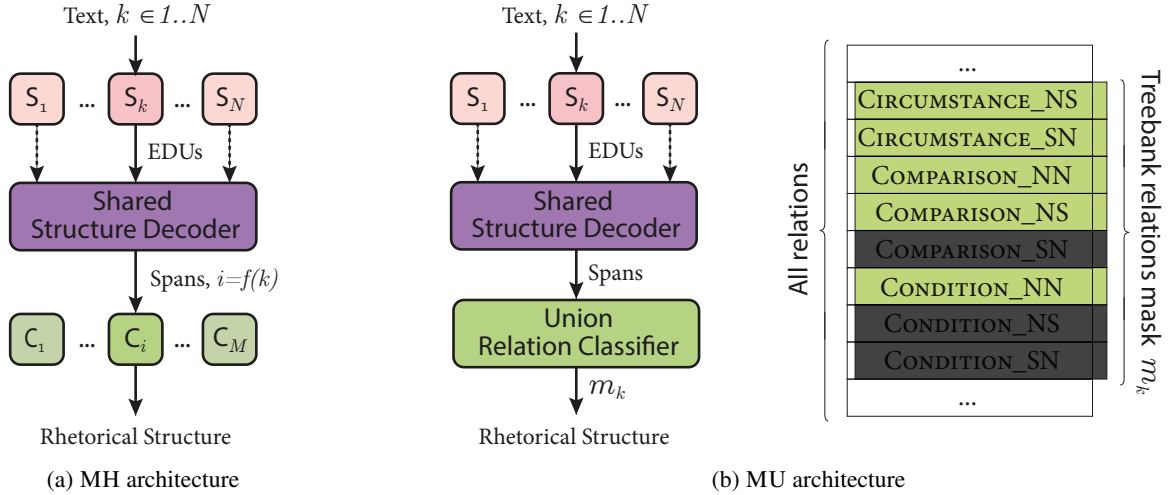


Figure 1: Model variants in the UniRST framework. (a) Multi-Head: independent classifiers per relation inventory. (b) Masked-Union: shared classifier with treebank-specific label masking.

banks. Specifically, we propose two strategies: *Multi-Head* (MH), which maintains separate classification heads per inventory, and *Masked-Union* (MU), which uses a single classifier constrained by treebank-specific masks. For reference, we additionally implement *Unmasked-Union* (UU), which lacks label masking and serves as a lower bound. Unless otherwise noted, models use treebank-specific segmentation heads, though shared segmentation is also tested. Figure 1 illustrates the architectures.

3.1 DMRST

DMRST (Liu et al., 2021) is an end-to-end RST parsing model that integrates EDU segmentation, discourse tree construction, and relation/nuclearity labeling. Its pipeline has four stages: (1) a pre-trained language model encodes input tokens, (2) an LSTM-CRF module detects EDU boundaries, (3) a recurrent pointer network decoder constructs the discourse tree, and (4) a biaffine classifier assigns nuclearity and relation labels. The model is trained jointly, with dynamically weighted loss balancing segmentation, structure prediction, and labeling. This unified design enables consistent end-to-end parsing.

UniRST extends this backbone to multi-treebank training. The pretrained encoder and recurrent decoder are shared across corpora, while segmentation and relation classification are treebank-tailored. This design aims to achieve robust structural prediction while respecting each corpus’s definitions and constraints.

3.2 Multi-Head (MH)

Our first method for multi-inventory RST parsing assigns a separate classification head to each distinct relation inventory. Given the set of inventories $\mathcal{G} = \{G_1, \dots, G_M\}$, treebanks sharing the same inventory (e.g., eng.gum, rus.rrg, zho.gcdt) share a relation/nuclearity classifier $\mathbf{W}^{(m)} \in \mathbb{R}^{d \times |G_m|}$. In this configuration, cross-treebank information about relation and nuclearity is exchanged only implicitly, through fine-tuning of the language model and shared structural decoder.

3.3 Masked-Union (MU)

Let $\mathcal{U} = \bigcup_k \mathcal{L}_{T_k}$ be the unified set of all relation types across treebanks. MU employs a single shared classifier $\mathbf{W} \in \mathbb{R}^{d \times |\mathcal{U}|}$ that predicts over this unified label space. To enforce inventory constraints, for each treebank T_k , we apply a binary mask $\mathbf{m}^{(k)} \in \{-1 \times 10^9, 1\}^{|\mathcal{U}|}$ to the classifier logits. This parameter-efficient design promotes explicit parameter sharing and enables direct transfer for overlapping relations (e.g., ELABORATION_NS) across all components of the model.

3.4 Unmasked-Union (UU)

UU mirrors the MU architecture but omits the treebank-specific masking, thereby allowing predictions over the entire concatenated label set without restriction. Consequently, it can produce labels that do not exist in the target corpus, limiting its practical utility. We include UU scores as a lower-bound baseline.

Treebank	Language	# Docs	# Tokens	# EDUs	# Labels	# Classes	# Rels
ces.crdt (2024)	Czech	54	14,623	1,345	23	34	1,288
deu.pcc (2014)	German	176	32,836	2,842	25	37	2,665
eng.gentle (2023) (test only)	English	26	17,799	2,328	15	27	2,552
eng.gum v11.1 (2025)	English	255	250,290	34,428	15	27	32,173
eng.oll (2008)	English	327	46,177	3,026	21	35	2,716
eng.rstdt (2002)	English	385	205,829	21,789	18	42	21,404
eng.sts (2008)	English	150	70,422	3,208	21	35	3,058
eng.umuc (2024)	English	87	61,292	5,421	28	46	5,334
eus.ert (2013)	Basque	88	45,780	2,509	24	31	2,421
fas.prstc (2021)	Persian	150	66,694	5,789	18	26	5,638
fra.annodis (2012)	French	86	32,699	3,307	18	20	3,221
nld.nldt (2012)	Dutch	80	24,898	2,326	27	45	2,246
por.cstn (2011)	Portuguese	140	58,793	5,527	22	38	5,387
rus.rrg (2024)	Russian	213	172,405	25,222	15	27	25,010
rus.rtt (2017)	Russian	233	262,495	28,247	17	25	25,892
spa.rststb (2011)	Spanish	267	58,717	3,351	29	43	3,084
spa.sctb (2018)	Spanish	50	16,515	744	20	26	694
zho.gcdt (2022)	Chinese	50	62,905	9,403	15	28	9,345
zho.sctb (2018)	Chinese	50	15,496	744	20	26	684

Table 1: Treebank statistics.

4 Data

This study leverages training data from 18 RST treebanks covering 11 languages, aiming to create the most universal end-to-end RST parser to date. The treebanks span Czech, German, English, Basque, Persian, French, Dutch, Brazilian Portuguese, Russian, Spanish, and Chinese. Treebank statistics are summarized in Table 1.

For the English RST-DT benchmark, we adopt the coarse-grained relation labels used in prior work. Corpora annotated using the GUM RST schema (eng.gum, zho.gcdt, rus.rrg) retain their predefined coarse-grained labels. For other corpora, if applicable, we merged infrequent classes (less than 10 instances) with related ones based on nuclearity, following the mapping suggested by Braud et al. (2017). This ensures both label diversity and sufficient representation for training. Detailed class distributions are illustrated in Appendix B.

To ensure consistency and reproducibility, we use the standardized training, validation, and test splits² provided by the DISRPT 2025 shared task for segmentation, connective identification, and relation classification across discourse annotation frameworks.

4.1 Data Augmentation

While several large RST treebanks dominate end-to-end discourse parsing research, smaller corpora

remain underutilized due to limited training data. To address this gap and establish strong mono-treebank baselines, we propose a simple yet effective data augmentation technique to improve performance in low-resource settings. Crucially, our method enriches training data without modifying the original texts or local annotations.

DMRST model employs a recurrent structure prediction module that relies heavily on contextual signals. As each annotated tree yields a single training instance, the number of examples is limited, particularly in smaller treebanks. To address this, we introduce an augmentation approach based on extracting structurally coherent subtrees from annotated documents. While these subtrees omit full-document context, their internal discourse structure remains valid and informative.

Our procedure involves: (1) identifying sentence boundaries to avoid extracting subtrees spanning sentence fragments; (2) extracting all connected subtrees not spanning sentence fragments and including at least three rhetorical relations; and (3) sampling a proportion p_{aug} of these subtrees for augmentation. Sampling is critical to prevent overfitting, particularly for the segmentation subtask.

This augmentation allows the model to train on a wider range of partial structures, potentially improving end-to-end RST parsing training in low-resource settings. We set p_{aug} to 50% to enrich the training data manifold.³

²We employ the open version of eus.ert treebank from <https://ixa2.si.ehu.es/diskurtsoa/en/>, containing 88 annotations.

³For RST-DT, $p_{\text{aug}} = 50\%$ produces 5.4 times more training samples. Over all treebanks, it multiplies number of training samples by 7.7.

Treebank	Baseline										+ Augmentation									
	Gold seg			End-to-end							Gold seg			End-to-end						
	S	N	R	Full	Seg	S	N	R	Full	S	N	R	Full	Seg	S	N	R	Full		
ces.crdt	60.2	31.1	18.2	16.9	90.1	47.3	24.0	13.7	12.3	58.9	31.1	18.0	17.1	90.6	46.2	23.4	11.5	11.2		
deu.pcc	68.7	38.8	24.7	23.6	95.2	58.9	33.9	21.0	20.0	67.2	42.7	26.4	25.6	96.0	59.7	36.9	21.7	21.2		
eng.gum	73.3	60.5	52.6	51.5	95.5	66.9	55.4	48.3	47.4	72.8	59.9	52.6	51.4	95.2	66.1	54.3	47.9	46.9		
eng.o11	65.9	48.2	29.5	29.3	89.7	51.4	36.7	21.9	21.5	61.8	43.2	29.0	28.4	91.2	54.1	36.4	24.5	24.0		
eng.rstdt	77.5	66.6	56.1	54.6	97.6	73.8	63.4	53.3	51.8	78.3	67.5	57.0	55.2	97.7	74.9	64.5	54.6	52.9		
eng.sts	46.5	35.1	21.7	21.1	89.7	38.2	28.8	18.2	18.0	44.1	32.9	20.5	19.6	88.4	33.5	24.5	16.1	15.7		
eng.umuc	68.8	50.8	33.1	32.4	89.6	51.2	36.9	24.3	23.7	67.1	48.4	31.8	31.2	89.0	49.1	35.1	24.3	24.0		
eus.ert	71.0	47.3	29.9	29.2	89.7	54.8	38.0	23.1	22.7	66.5	44.5	25.7	25.7	89.0	52.3	35.3	19.9	19.8		
fas.prstc	65.0	51.3	40.2	40.1	93.8	55.3	44.6	34.4	34.4	65.3	50.9	40.7	40.4	93.8	55.3	42.9	34.3	34.0		
fra.annodis	62.5	51.6	33.0	33.0	92.1	53.2	44.7	28.6	28.6	62.5	51.4	32.9	32.9	91.5	52.4	43.3	27.6	27.6		
nld.nldt	63.8	47.4	30.7	29.1	96.3	58.2	42.9	28.5	26.9	61.7	46.4	30.6	28.8	96.4	57.2	42.8	28.9	27.3		
por.cstn	76.0	62.2	50.9	50.8	93.9	68.2	53.3	43.9	43.8	76.1	61.6	49.9	49.9	94.0	66.3	53.3	43.0	43.0		
rus.rg	71.2	57.2	49.4	48.2	97.2	67.6	54.3	47.1	46.0	70.3	55.9	47.9	46.8	96.8	65.6	52.2	44.9	43.9		
rus.rt	79.7	61.4	51.5	51.3	91.1	62.8	49.0	41.4	41.3	80.0	61.9	52.2	51.9	91.0	63.0	49.9	42.5	42.3		
spa.rststb	68.1	52.2	35.0	35.0	91.5	54.9	40.9	28.1	28.1	71.1	54.4	38.9	38.9	91.9	57.7	44.0	32.8	32.8		
spa.sctb	66.7	41.5	35.2	35.2	74.1	34.3	25.4	23.1	23.1	66.7	43.6	31.7	31.7	84.1	47.5	35.3	27.3	27.3		
zho.gcdt	76.3	58.1	52.3	50.7	91.2	61.0	46.1	40.9	39.6	75.3	58.2	51.9	50.4	91.9	64.0	49.5	43.5	42.3		
zho.sctb	66.7	37.7	32.1	32.1	91.1	56.3	34.3	28.5	28.5	60.2	40.9	32.3	32.3	92.5	52.3	38.1	29.6	29.6		

Table 2: Performance of the treebank-specific models, with and without train data augmentation.

5 Experimental Setup

We employ `xlm-roberta-large` as the multilingual encoder across all experiments. The batch size is set to 2, with a hidden size of 200 for the segmenter and 512 for the parsing module. The DMRST model is trained with a learning rate of $1e-5$, while the encoder is fine-tuned using a learning rate of $2e-5$. Early stopping is set to a patience of 5 in mono-treebank settings and reduced to 3 in UniRST due to the larger concatenated dataset.

Evaluation follows the original Parseval metrics for rhetorical structure parsing, with micro F1 scores reported for segmentation (Seg), span (S), relation (R), nuclearity (N), and full structure (Full). Each model is trained using three different random seeds, and all reported results are averaged across these runs.

6 Experimental Results

6.1 Mono-Treebank Evaluations

Table 2 reports the performance of treebank-specific models trained with and without data augmentation. Augmentation yielded substantial gains on smaller corpora such as `eng.o11`, `spa.sctb`, `zho.sctb`, and `zho.gcdt`, but improvements were not uniform across all treebanks. Interestingly, on the `eng.rstdt` benchmark with diverse document lengths, augmentation led to an average 1.1% F1 improvement in unlabeled structure prediction (S), highlighting its potential even for larger datasets. On the other hand, the data augmentation

Model	In-treebank			All avg.		
	Seg	S	N	Seg	S	N
ces.crdt	90.5	49.5	24.5	76.4	32.6	16.0
deu.pcc	96.0	59.7	36.9	76.6	32.9	19.3
eng.gum	95.5	66.9	55.4	78.9	41.3	30.2
eng.o11	91.2	54.1	35.4	77.3	31.9	18.7
eng.rstdt	97.7	74.9	64.5	78.4	40.2	28.8
eng.sts	89.7	38.2	32.6	75.8	29.1	16.8
eng.umuc	89.6	51.2	36.9	77.6	35.1	22.4
eus.ert	89.7	54.8	38.0	77.5	33.8	18.4
fas.prstc	93.8	58.3	44.6	78.3	36.2	20.3
fra.annodis	92.1	53.2	44.7	77.5	32.6	16.5
nld.nldt	96.4	57.0	43.9	80.2	35.9	22.0
por.cstn	94.1	68.8	57.2	78.2	37.1	25.2
rus.rg	97.2	67.6	54.3	80.0	40.9	28.7
rus.rt	91.0	63.0	49.9	81.2	40.9	27.3
spa.rststb	91.9	57.7	44.0	78.5	35.4	22.6
spa.sctb	84.1	47.5	35.3	74.0	29.9	16.2
zho.gcdt	91.9	64.0	49.5	76.9	36.7	23.3
zho.sctb	92.5	52.3	38.1	57.4	17.1	10.2
UniRST	—	—	—	92.9	60.7	47.3

Table 3: Evaluation across all treebanks. We only assess segmentation (Seg), unlabeled structure construction (S), and nuclearity assignment (N), as relation inventories are incompatible.

resulted in performance degradation on two large-scale GUM-based corpora (`eng.gum`, `rus.rg`), likely due to segmenter overfitting on long documents. Overall, augmentation yielded the best mono-treebank parsing performance on 10 of the 18 treebanks. For comparison, Appendix A summarizes previous end-to-end RST parsing results on eight treebanks. DMRST+ denotes the architecture used as a baseline in this work.

Method	Segmentation	Gold seg				End-to-end				
		S	N	R	Full	Seg	S	N	R	Full
MH	single	73.5	58.4	47.6	46.6	93.4	63.4	50.7	41.7	40.8
	multiple	73.6	59.0	48.5	47.6	93.7	63.7	51.3	42.4	41.6
UU	single	73.8	58.7	47.0	46.8	93.4	64.3	51.9	42.8	41.8
MU	single	74.1	59.3	48.8	47.8	93.7	64.5	52.1	43.2	42.3
	multiple	74.4	59.6	49.3	48.3	93.9	64.8	52.1	43.4	42.5

Table 4: Performance of the UniRST model in different setups.

Treebank	Seg	S	N	R	Full
ces.crdt	94.2 (+4.1)	57.9 (+10.6)	38.6 (+14.6)	27.3 (+3.3)	26.8 (+14.5)
deu.pcc	96.5 (+0.5)	66.3 (+6.6)	45.5 (+8.6)	32.8 (+11.1)	31.1 (+9.9)
eng.gum	95.2 (-0.3)	66.7 (-0.2)	54.7 (-0.7)	48.0 (-0.3)	46.9 (-0.5)
eng.oll	93.8 (+2.6)	56.7 (+2.6)	40.6 (+4.2)	27.6 (+3.1)	27.1 (+3.1)
eng.rstdt	97.8 (+0.1)	75.6 (+0.7)	65.1 (+0.6)	55.2 (+0.6)	53.5 (+0.6)
eng.sts	91.0 (+1.3)	40.4 (+2.2)	30.7 (+1.9)	19.4 (+1.2)	18.8 (+0.8)
eng.umuc	88.8 (-0.8)	52.0 (+0.8)	40.1 (+3.2)	26.1 (+1.8)	25.6 (+1.9)
eus.ert	92.0 (+2.3)	62.8 (+8.0)	47.4 (+9.4)	35.4 (+12.3)	35.3 (+12.6)
fas.prstc	94.6 (+0.8)	61.7 (+6.4)	50.2 (+5.6)	40.7 (+6.3)	40.5 (+6.1)
fra.anodis	90.9 (-1.2)	58.1 (+4.9)	47.3 (+2.6)	31.1 (+2.5)	30.7 (+2.1)
nld.nldt	97.6 (+1.2)	59.3 (+2.1)	45.3 (+2.5)	33.5 (+4.6)	31.7 (+4.4)
por.cstn	94.3 (+0.4)	67.7 (-0.5)	54.9 (+1.6)	45.7 (+1.8)	45.4 (+1.6)
rus.rrg	96.5 (-0.7)	66.8 (-0.8)	53.5 (-0.8)	45.5 (-1.6)	44.1 (-1.9)
rus.rrt	90.6 (-0.4)	63.0 (0.0)	49.8 (-0.1)	42.6 (+0.1)	42.4 (+0.1)
spa.rststb	92.5 (+0.6)	63.5 (+5.8)	50.1 (+6.1)	35.3 (+2.5)	35.2 (+2.4)
spa.sctb	86.0 (+1.9)	55.8 (+8.3)	48.0 (+12.7)	40.8 (+13.5)	40.8 (+13.5)
zho.gcdt	92.1 (+0.2)	62.9 (-1.1)	48.7 (-0.8)	44.0 (+0.5)	42.7 (+0.4)
zho.sctb	94.3 (+1.8)	64.3 (+12.0)	50.5 (+12.4)	40.7 (+11.1)	40.7 (+11.1)

Table 5: UniRST performance per treebank. Improvements over the strongest mono-treebank baseline, as listed in Table 2, are shown in parentheses.

To assess generalization, each best-performing treebank-specific model was evaluated on all 18 corpora. Table 3 reveals a consistent transferability gap: models tend to overfit to treebank-specific language, domains, relation usage, and document styles. Segmentation scores also decline in transfer settings, though less severely than Span or Nuclearity scores. In certain cases, however (e.g., eng.oll, eng.gum), segmentation drops sharply, reflecting variation in EDU definitions across corpora. Despite strong in-treebank Span F1 (e.g., 74.9% for eng.rstdt, 68.8% for por.cstn), transfer performance degrades substantially (dropping to 40.2% and 37.1%, respectively). This disparity demonstrates that in-domain success is a poor indicator of cross-corpus robustness and highlights the need for more generalizable RST parsers, such as UniRST.

6.2 UniRST

Performance of the Multi-Head and Masked-Union strategies is reported in Table 4. UniRST performs

best when segmentation is handled by treebank-specific heads, which capture differences in EDU annotation schemes, whereas a universal segmentation head primarily learns broader segmentation patterns. The Masked-Union (MU) strategy consistently outperforms Multi-Head (MH), offering both greater efficiency and higher parsing accuracy. Its masking mechanism ensures that each treebank’s inventory is respected, while still enabling transfer for overlapping relations, which in turn improves parsing performance over the unmasked baseline. The strongest configuration is MU with treebank-specific segmentation heads. We refer to this variant as “UniRST” throughout the remainder of the paper.

As shown in Table 3, UniRST achieves higher average performance across combined test set compared to any mono-treebank parser. This demonstrates the robustness of UniRST model as a cross-lingual parser capable of learning shared representations that generalize effectively across diverse

RST corpora.

Detailed results by treebank are provided in Table 5. The unified model outperforms the strongest mono-treebank baselines on 16 out of 18 treebanks. Notable improvements in end-to-end Full F1 are observed across most datasets, particularly for smaller-scale treebanks such as *ces.crdt*, *deu.pcc*, *eus.ert*, *spa.sctb*, *zho.gcdt*, and *zho.sctb*. Similar to data augmentation in mono-treebank training, joint training does not benefit the large-scale *eng.gum* and *rus.rrg* corpora, whose annotations appear sufficient on their own. Importantly, the performance drop on *eng.gum* under joint training remains marginal. The only corpus where UniRST fails to exceed 50% Span F1 and 25% Full F1 is *eng.sts*. Given the limited documentation of this dataset, the cause is unclear, but the low scores may stem from poor inter-annotator agreement or inconsistently applied segmentation and structural constraints. Joint training nonetheless improved performance, suggesting that it provides some stabilization even under noisy conditions. Across nine corpora in English, Persian, Portuguese, Russian, Spanish, and Chinese, UniRST achieves more than 40% Full end-to-end F1 while preserving original relation inventories.

To further assess out-of-domain generalization, we evaluate GUM-compatible models on the GENTLE benchmark, which follows GUM annotation guidelines.⁴ As shown in Table 6, UniRST achieves the highest Full end-to-end parsing score. The *eng.gum* model performs best in segmentation (93.0% F1) and structure prediction (58.0%) due to its alignment with GENTLE’s language and annotation conventions. However, UniRST outperforms it on Relation and Full F1, highlighting the benefits of shared relation classification training across multiple treebanks. Notably, UniRST supports 11 languages, while *eng.gum* is English-only. Training on multiple multi-domain treebanks, including five English treebanks, did not lead to a substantial improvement in out-of-domain performance over the GUM-specific model. These findings highlight the importance of treebank-specific annotation schemes and show that the universal model remains most effective within the domains and genres present in its training data.

⁴GENTLE includes annotations for eight unconventional genres: dictionary entries, esports commentaries, legal documents, medical notes, poetry, mathematical proofs, syllabuses, and threat letters. None of these genres are represented in the training corpora used in this work.

Model	Seg	S	N	R	Full
<i>eng.gum</i>	93.0	58.0	47.2	39.1	38.6
<i>rus.rrg</i>	85.2	44.7	34.9	28.8	28.3
<i>zho.gcdt</i>	76.4	34.1	23.5	18.4	18.0
UniRST	92.7	57.4	46.0	39.9	39.4

Table 6: Performance of the GUM-compatible models on GENTLE out-of-domain benchmark.

7 Conclusion

While previous approaches to multilingual parsing have often advocated for reducing relation inventories to a small standardized set of RST relations, such strategies fail to fully account for the broader divergences among RST treebanks. These include differences in discourse segmentation, the treatment of mono- versus multinuclearity, and the granularity, specificity, and definitions of rhetorical relations. In this work, we introduced UniRST, the first unified RST-style discourse parser capable of effectively processing 18 treebanks across 11 languages without altering their original relation inventories. To address the challenge of inventory incompatibility, we proposed two approaches: Multi-Head and Masked-Union. Our results show that the latter yields superior performance, particularly when paired with treebank-specific segmentation heads. UniRST outperforms 16 out of 18 mono-treebank baselines, demonstrating that end-to-end multilingual discourse parsing is achievable despite considerable annotation diversity. The results indicate that embracing annotation heterogeneity can benefit multilingual discourse parsing.

Limitations

The main limitation of a multilingual RST parser that preserves multiple relation inventories lies in the need to account for inventory differences in downstream applications. This issue is not unique to our approach, as it also arises when deploying separate treebank-specific models per language or domain. Even under label harmonization to a reduced set, variation in the number and distribution of relations across languages can persist. While UniRST demonstrates strong generalization across most treebanks, it shows a marginal performance drop on two large, multi-domain corpora (*eng.gum*, *rus.rrg*), likely because their annotations are sufficient to support strong mono-treebank models. Furthermore, *eng.sts* remains the only dataset where Span F1 remains below 50%, with both mono- and

multi-treebank models performing poorly. These observations suggest that data quality and annotation consistency substantially affect performance, and that future work may benefit from treebank filtering or weighting.

Acknowledgments

The research was carried out using the infrastructure of the shared research facilities «High Performance Computing and Big Data» of FRC CSC RAS (CKP «Informatics»).

References

- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. [An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).
- Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. [GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation](#). In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.
- James Barry, Joachim Wagner, and Jennifer Foster. 2019. [Cross-lingual parsing with polyglot training and multi-treebank learning: A Faroese case study](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 163–174, Hong Kong, China. Association for Computational Linguistics.
- Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. [Cross-lingual RST discourse parsing](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.
- Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. [The RST Spanish-Chinese treebank](#). In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, M. Eloize, R. Kibar Aji Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. [CST-News - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese](#). In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, Brazil.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. [Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory](#). In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.
- Elena Chistova. 2024. [Bilingual rhetorical structure parsing with large parallel annotations](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9689–9706, Bangkok, Thailand. Association for Computational Linguistics.
- Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. [On the development of the RST Spanish treebank](#). In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.
- Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. [The RST Basque TreeBank: An online search interface to check rhetorical relations](#). In *4th Workshop on RST and Discourse Studies*, pages 40–49, Fortaleza, Brasil.
- Richard Johansson. 2013. [Training parsers on incompatible treebanks](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–137, Atlanta, Georgia. Association for Computational Linguistics.
- Richard Johansson and Yvonne Adesam. 2020. [Training a Swedish constituency parser on six incompatible treebanks](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5219–5224, Marseille, France. European Language Resources Association.
- Yash Kankanampati, Joseph Le Roux, Nadi Tomeh, Dima Taji, and Nizar Habash. 2020. [Multitask easy-first dependency parsing: Exploiting complementarities of different dependency representations](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2497–2508, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yang Janet Liu and Amir Zeldes. 2023. [Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. [Multilingual neural RST discourse parsing](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. [DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. RST Discourse Treebank LDC2002T07.
- William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.
- Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Draoulec, and Laure Vieu. 2012. Manuel d’annotation en relations de discours du projet annodis.
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. [RST parsing from scratch](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.
- Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. [GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.
- Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, A. Nasedkin, S. Nikiforova, I. Pavlova, and A. Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 194–204.
- Lucie Polakova, Jiří Mírovský, Šárka Zikánová, and Eva Hajicova. 2024. [Developing a Rhetorical Structure Theory treebank for Czech](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4802–4810, Torino, Italia. ELRA and ICCL.
- Andrew Potter. 2008. Interactional coherence in asynchronous learning networks: A rhetorical approach. *The Internet and Higher Education*, 11(2):87–97.
- Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. [Multi-layer discourse annotation of a Dutch text corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).
- Zeeshan Ali Sayyed and Daniel Dakota. 2021. [Annotations matter: Leveraging multi-task learning to parse UD and SUD](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3467–3481, Online. Association for Computational Linguistics.
- Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.
- Manfred Stede and Arne Neumann. 2014. [Potsdam commentary corpus 2.0: Annotation for discourse research](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. [Parser training with heterogeneous treebanks](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.
- Karolina Zaczynska and Manfred Stede. 2024. [Rhetorical strategies in the UN security council: Rhetorical Structure Theory and conflicts](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28, Kyoto, Japan. Association for Computational Linguistics.
- Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. [eRST: A signaled graph theory of discourse relations and organization](#). *Computational Linguistics*, 51(1):23–72.
- Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. [A top-down neural architecture towards text-level parsing of discourse rhetorical structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.

A Reference Results from Prior Work

Table 7 summarizes previously reported results for end-to-end RST parsing. It is important to note that prior results may differ in experimental setup,⁵ limiting direct comparability. All results reported in Section 6.1 are obtained through single-treebank

⁵Most notably, in the use of multicorpus training with harmonized label sets, or non-standard train/dev/test splits.

training using the original relation sets and the standardized DISRPT 2025 splits. To the best of our knowledge, the remaining treebanks are evaluated here for the first time in a full end-to-end RST parsing setting.

System	Seg	S	N	R	Full
<i>eng.rstdt</i>					
SegBot (2020)	92.2	62.3	50.1	40.7	39.6
Nguyen et al. (2021)	96.3	68.4	59.1	47.8	46.6
DMRST (2021)	96.3	68.4	59.1	47.8	46.6
DMRST+ (2024)	97.8	74.8	64.5	54.5	53.0
<i>deu.pcc</i>					
DMRST (2021)	96.5	70.4	60.6	n/c	n/c
<i>eus.ert</i>					
DMRST (2021)	88.7	53.3	39.1	n/c	n/c
<i>nld.nldt</i>					
DMRST (2021)	95.5	62.3	46.6	n/c	n/c
<i>por.cstn</i>					
DMRST (2021)	92.8	62.5	51.6	n/c	n/c
<i>rus.rrg</i>					
DMRST+ (2024)	96.9	66.5	53.3	45.8	44.6
<i>rus.rrt</i>					
DMRST+ (2024)	92.2	65.9	51.0	43.9	43.8
<i>spa.rststb</i>					
DMRST (2021)	92.8	62.5	51.6	n/c	n/c

Table 7: Reference end-to-end parsing evaluations across RST treebanks. **n/c** indicates incompatible (harmonized) label sets.

B Relation Classes across Treebanks

Figures 2 and 3 illustrate the distribution of all relation labels across 19 treebanks (including the test-only *eng.gentle*). UniRST handles all 96 unique LABEL_NUCLEARITY relations as they appear in each corpus. Note that while some treebanks (e.g., GUM-style and RST-DT) internally group ANTITHESIS, CONTRAST, and CONCESSION as ADVERSATIVE, and CAUSE with RESULT as CAUSAL, others treat some of these relations separately or organize them under alternative groupings.

During preprocessing, only relations with equivalent definitions and comparable granularity were unified under a single label (e.g., CONDITION and CONTINGENCY; ADVERSATIVE and coarse-grained CONTRAST). CONDITION is a coarse-grained label encompassing, in most treebanks, the underrepresented fine-grained relations OTHERWISE, UNLESS, and UNCONDITIONAL, each of which appears too infrequently to be modeled reliably on its own. Labels without clear counterparts, such as GRADATION_SN in *ces.crdt* (Polakova et al., 2024) or FRAME_NS in *fra.annodis*

(Muller et al., 2012), remain unique to their respective treebanks.

Variations in the representation of overlapping labels across treebanks reflect underlying genre and linguistic differences. For instance, *zho.gcdt* features more instances of ELABORATION_SN than ELABORATION_NS, in stark contrast to other languages, where the satellite in ELABORATION typically follows the nucleus.

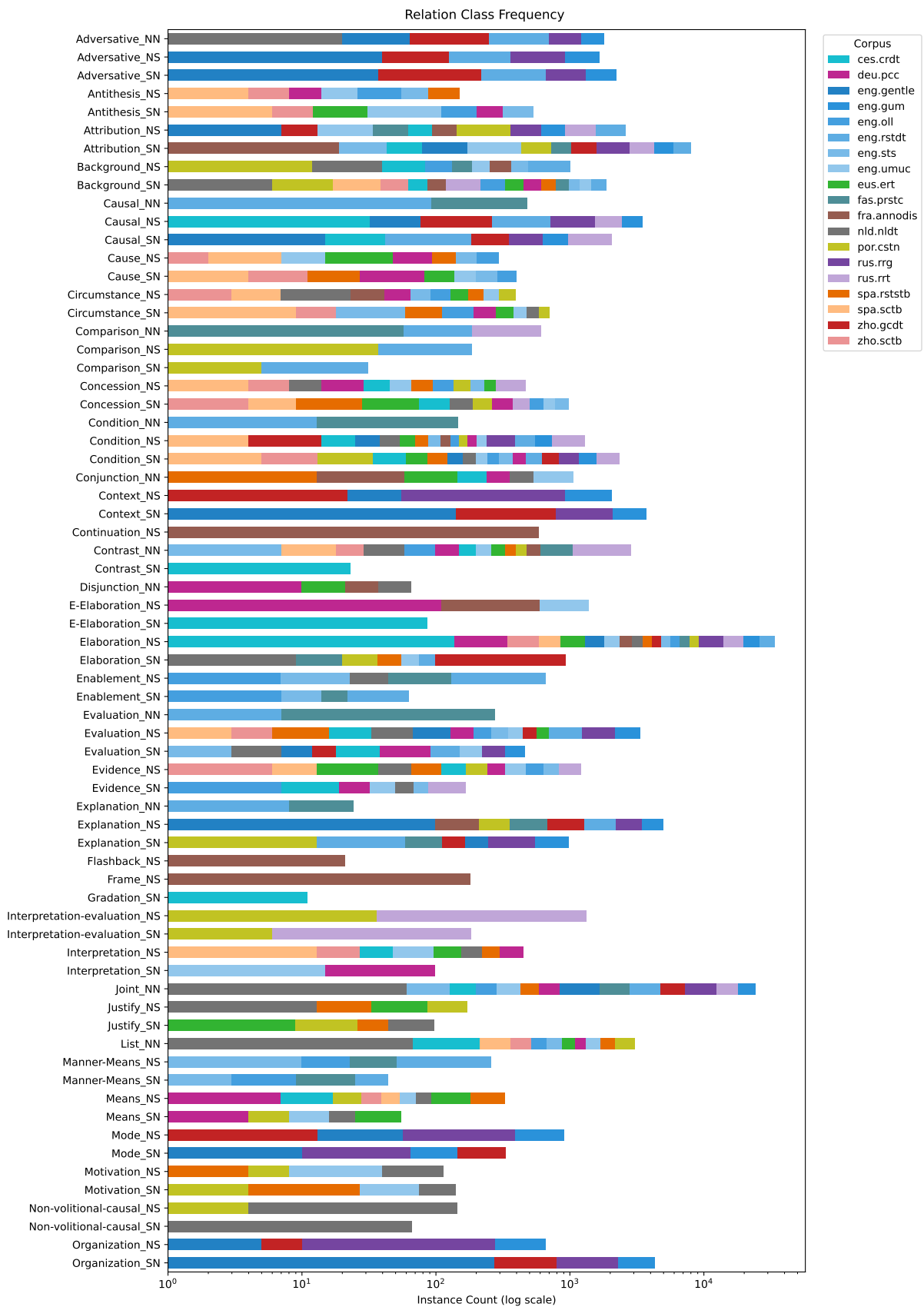


Figure 2: Relation class frequency across treebanks.

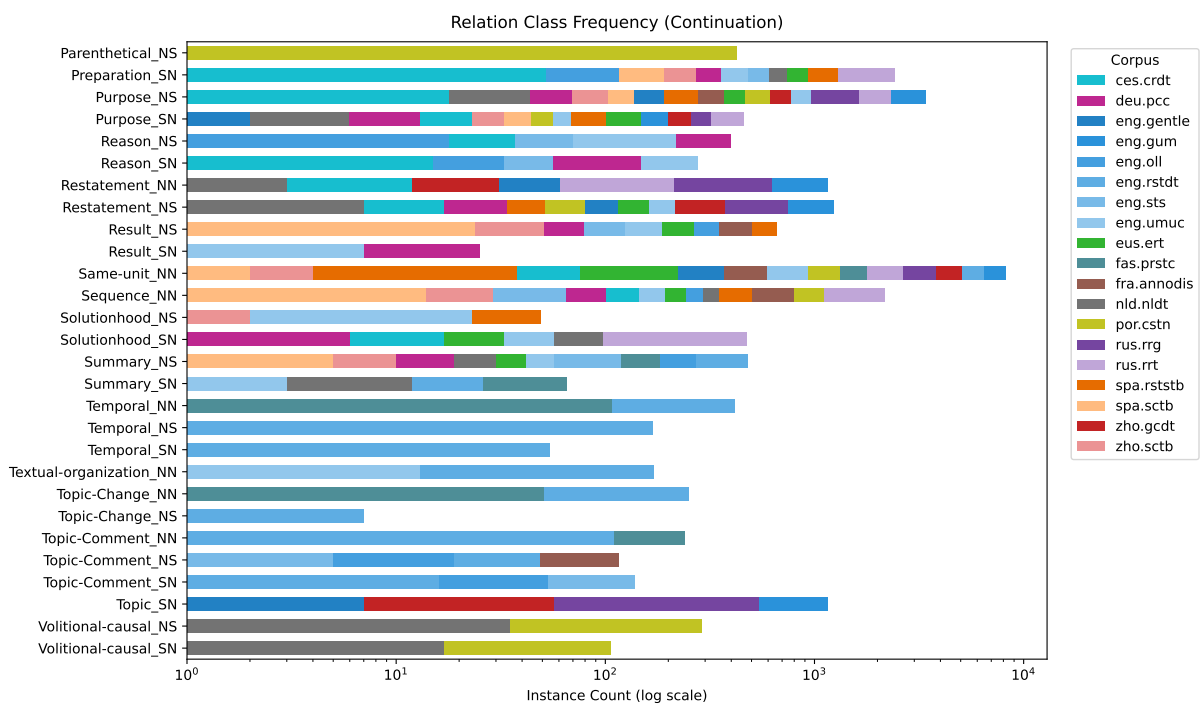


Figure 3: Relation class frequency (continuation).

Corpus-Oriented Stance Target Extraction

Benjamin D. Steel

McGill University

benjamin.steel@mail.mcgill.ca

Derek Ruths

McGill University

derek.ruths@mcgill.ca

Abstract

Understanding public discourse through the frame of stance detection requires effective extraction of issues of discussion, or stance targets. Yet current approaches to stance target extraction are limited, only focusing on a single document to single stance target mapping. We propose a broader view of stance target extraction, which we call corpus-oriented stance target extraction. This approach considers that documents have multiple stance targets, those stance targets are hierarchical in nature, and document stance targets should not be considered in isolation of other documents in a corpus. We develop a formalization and metrics for this task, propose a new method to address this task, and show its improvement over previous methods using supervised and unsupervised metrics, and human evaluation tasks. Finally, we demonstrate its utility in a case study, showcasing its ability to aid in reliably surfacing key issues of discussion in large-scale corpora.

1 Introduction

Disagreement is a critical part of discourse. As such, understanding discourse requires inferring the constituent disagreements. This task becomes increasingly complex as discussions scale to online environments (Gottfried, 2024), where the pressing need to ensure healthy dialogue is compounded by threats from inauthentic influence attempts and harmful platform mechanisms (Saurwein and Spencer-Smith, 2021; Goldstein et al., 2023; Commission, 2024). To address these challenges, we need both easy-to-use analytical methods and clear representations of discussion data. However, developing such tools presents challenges, given that online media documents typically mix many different related issues, topics, and contexts.

Stance detection (i.e. the task of identifying the attitude of the author of a text on a stance target (a claim, entity etc.) (Mohammad et al., 2016)) is a

well-developed method for understanding disagreement. But the current state of stance detection is such that, unless one knows a priori the stance targets one wants to know the documents' stance on, one must undertake the difficult task of defining those stance targets oneself via the arduous task of understanding the entire corpus. While there are initial methods available for finding targets in documents, we propose that they are insufficient at the corpus-level, and that such a method needs four key features in order to faithfully and clearly capture stance in a discussion corpus:

1. The stance targets need not be known a priori to the researcher - avoiding human bias in issue selection, and improving scalability.
2. A single document can articulate a position on multiple, or hierarchical (i.e. more abstract, or more general), targets - which frequently occurs in the real-world - and as such, the method should map the document to these targets.
3. Targets should be determined in the context of the corpus - meaning both that the discussion as a whole aids the inference of the targets of a document, and that documents should be clustered to targets to allow aggregation for downstream application.
4. Documents should be mapped to clear representations of these stance targets, to aid understanding, and allow use in downstream tasks

Existing approaches do not address all of these features. Most stance target extraction methods produce a single stance target for a single given document, without attending to the broader context of a discussion, or allowing for multiple issues to be addressed in a document (Irani et al., 2024; Akash et al., 2024; Li et al., 2023; Zhang et al., 2021). Disagreement discovery methods from outside the stance-detection literature that *do* consider a corpus

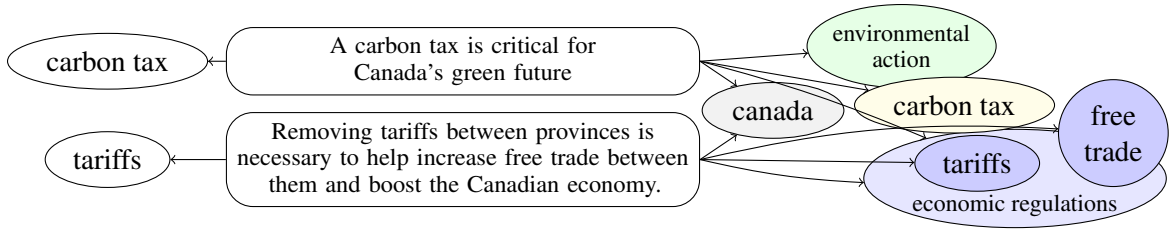


Figure 1: Comparison of assigning a single stance target to each document (left), versus assigning multiple hierarchical stance targets that overlap with other texts as proposed here (right).

as a whole (Paschalides et al., 2021; He et al., 2021) do not produce a clear mapping of documents to stance targets.

We make four contributions. We formalize this task of mapping issues/targets of disagreements in a corpus into a computational task which we call *corpus-oriented stance target extraction* (COS-TE_x). We then provide a metric for evaluating a method’s performance on this task. We present a method which addresses the task, and show it outperforms existing methods on our task. Finally, we conduct a case study using our method, which shows that it can retrieve key issues (stance targets and stances) from a discussion represented by a corpus.

With the evaluation and development of a method that performs well on the task we outline here, we can unlock powerful insights in large-scale media corpora, giving us new tools to understand large-scale natural language behaviour such as polarization and public opinion. We release a library for this method at <https://github.com/benda-vidsteel/stancemining>

2 Background

Subjectivity Detection The fields of stance detection, aspect-based sentiment detection, and argument mining have produced methods to identify targets of subjective perspective, and classifying the subjective judgement of documents towards those targets. Li et al. (2023) and Steel and Ruths (2024) look at stance target extraction, but both methods require a priori knowledge and manual target choice at a point in the method, and therefore do not fulfill feature 1. Akash et al. (2024), Irani et al. (2024) and Zhang et al. (2021), all look at open-target extraction (where there are no predefined targets) in stance detection, argument mining, and sentiment detection respectively. All three focus on inferring targets for documents in isolation and, as a result, none of these methods consider the multiple or hierarchical stance targets possible

from a document (as represented in Fig. 1 and defined as a required feature 2), or the need for large stance target clusters —groups of documents mapped to the same stance target—if we want to aggregate the data for further analysis (feature 3). Nevertheless, we compare our developed method against WIBA (Irani et al., 2024) in this work.

Polarized and Controversial Topics Topic modelling derived methods are a common approach to this problem space, and naturally handle the desired aggregation process from feature 3. But converting topic clusters to stance target clusters is not trivial. Topic and stance targets clusters don’t map neatly one-to-one, as demonstrated in Fig. 2a. Work on topic cluster representation, such as Pham et al. (2023) and Grootendorst (2022), uses large language models (LLMs) to improve the interpretability of cluster names, working towards feature 4. But mapping a topic cluster to a stance target is difficult, as it requires domain knowledge and reasoning to convert topic descriptions into a stance target (Fig. 2b).

Fukuma et al. (2022) use a network method to find polarized topics, but this method is designed for X/Twitter specific features. Garimella et al. (2018) use hashtags to define conversational graphs, and find partitions in those graphs in order to find controversial topics. This method however relies on hashtags, limiting it to corpuses with heavy hashtag usage. Paschalides et al. (2021) and He et al. (2021) produce methods to find polarized topics, and we evaluate these methods in this work.

3 Problem Definition

Motivated by our desired features from Section 1, we define COST_{Ex} as follows: given a corpus of documents, we seek to identify labeled clusters of those documents where all documents in a cluster share the same stance target, which is captured by the label of the cluster. Crucially, clusters can be overlapping, allowing a document to be assigned

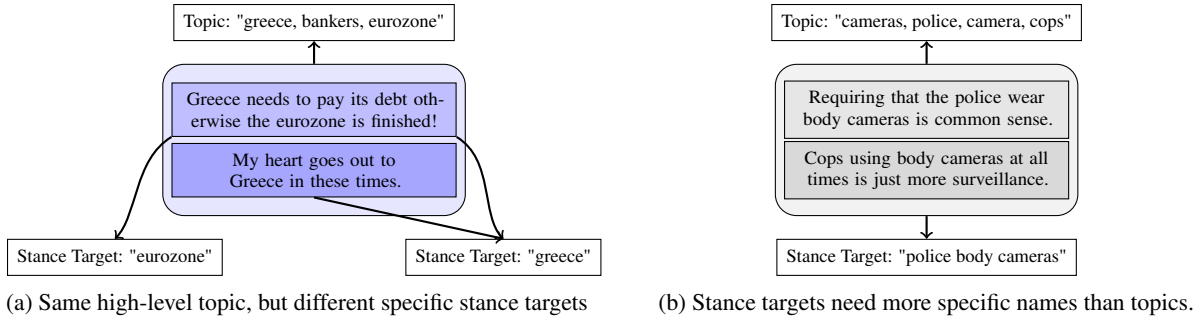


Figure 2: Representation of the differences between stance target clusters and topic clusters, showing hierarchical relationships, one-to-many mappings, and different cluster naming requirements, as discovered in manual analysis.

more than one stance target. Formally, for a corpus of documents $D = \{d_1, \dots, d_N\}$, we want to find a set stance target clusters, $C = \{c_1, c_2, \dots, c_M\}$ where $c_i \subset D$, and their corresponding stance targets $T = \{t_1, t_2, \dots, t_M\}$. As stance detection has not previously considered corpus-aware methods, we propose new criteria that define success in COSTEx, that measure the extent to which a method that implements this task fulfills the desired features. As such, the COSTEx problem seeks C and T such that they reflect the following criteria:

1. **Clusters with Large Stance Variance:**

Given the stance of each document on the stance target $stance(t_i, d_j) \rightarrow \{-1, 0, 1\}$, we want to find stance targets that maximize the stance variance for all related documents:

$$\frac{1}{|C|} \sum_{c_i \in C} Var(\{stance(t_i, d) : d \in c_i\})$$

This is a metric for picking “controversial” stance targets. Intuitively, stance targets that no-one disagrees on are less interesting than stance targets that people disagree on.

2. **Stance Target Range and Relevance:** We want to find many stance targets that are relevant to the documents. We can measure relevance of targets by ensuring that the stance targets adhere to human judgments of stance targets, via comparison to labeled datasets and custom human annotation (as in Akash et al. (2024)), and we can measure ‘many stance targets’ by measuring the mean number of targets per document:

$$\frac{1}{|D|} \sum_{d \in D} |\{c_i \in C : d \in c_i\}|$$

3. **Balanced Stance Target Clusters:** We want to optimize for clusters of a range of sizes,

including large clusters to allow for useful aggregations, that still capture clusters of meaningful grouping. To measure meaningful clusters, we will use human evaluation. And to measure cluster size, we will use the cluster size Shannon entropy multiplied by normalized cluster size range, to ensure there are a balanced number of clusters of a range of sizes:

$$\frac{\max_i c_i - \min_i c_i}{\max_i c_i} \left(- \sum_{i=1}^n \frac{c_i}{C} \log_2 \frac{c_i}{C} \right)$$

where $C = \sum_{k=1}^n c_k$

Naturally, in most situations it will be impossible to perfectly satisfy all of these. Solutions to this task will have to make careful trade-offs between these criteria. In practice, some of these metrics are trivially measurable, and some of them are much harder to measure (i.e. the ones requiring human evaluation). We will seek to do so via quantitative supervised and unsupervised metrics, and metrics from human evaluation tasks.

Finally, we must address the question of what we mean by *stance targets* in the formulation above. In the literature, it is common to define stance targets either as noun-phrases (e.g., “police body cameras”), or claims (“police should wear police body cameras”) (Zhao and Caragea, 2024). A document assigned to this stance target contains content that takes a position on it. Note that, where stance targets are concerned, the problem definition requires only a means of scoring a document’s position on a stance target (i.e., $stance(t_i, d)$). As a result, the problem admits either noun-phrases or claims as stance targets.

4 Methods

Here, we propose our method that fulfills all the features in Section 1. Rather than clustering documents directly (which conflates topics with stance

targets as discussed in Section 2), we first extract multiple stance targets (fulfilling feature 2), then cluster those targets. This lets us find large stance target clusters, where documents can naturally belong to multiple large clusters. We then generate stance targets for each of those clusters to find higher-level stance targets (i.e. targets that are hierarchically more abstract, or more general to the cluster). Collecting all these stance targets together for each document, we then have small, specific stance target clusters, and larger, high level stance target clusters. We call this method *ExtractCluster* (EC), formally define it in Algorithm 3, and show a simplified system diagram of it in Fig. 4. Our method aims to meaningfully achieve each criteria from Section 3.

The base stance targets are produced using an LLM fine-tuned on document - stance target pairs, using diverse beam search (Vijayakumar et al., 2016) to generate multiple targets. We cluster the targets using BERTopic (Grootendorst, 2022), which provides an easy-to-use and configurable topic modelling solution. The default clustering configuration of BERTopic gives us one layer of clusters, meaning there are two hierarchical levels of stance targets. The higher-level stance targets are generated using an LLM with a few-shot prompt (shown in Appendix B.6). To avoid producing stance targets for each document that are paraphrases of each other, we remove stance targets where their sentence embeddings have high cosine similarity based on a configurable threshold (Reimers and Gurevych, 2019) (detailed in Appendix B.4).

4.1 Comparison Methods

We selected three methods to compare to EC on our task COSTEx. Although these methods do not fully address our proposed task, they address it sufficiently to warrant evaluation.

POLAR (Paschalides et al., 2021) uses entity extraction and network methods to find polarized topics. While this method is designed to find polarized topics, we apply it here to the similar but more general COSTEx task. Though the method does not explicitly map documents to stance targets, we extend it to use any entities or noun phrases that are tagged as part of a polarized topic as stance targets for their respective documents.

PaCTE (He et al., 2021) combines topic modeling and a partisanship classification model to find

```

1: function EXTRACTCLUSTER( $D$ )
2:   for each document  $d \in D$  do
3:      $T_d \leftarrow$  ExtractStanceTargets( $d$ )
4:      $T_d \leftarrow$  RemoveSimilarTargets( $T_d$ )
5:   end for
6:    $C \leftarrow$  TopicModelTargets( $T$ )
7:   for each cluster  $c \in C$  do
8:      $T_c \leftarrow$  GenerateHigherLevelTargets( $c$ )
9:      $T_c \leftarrow$  RemoveSimilarTargets( $T_c$ )
10:    for each  $d : \exists t \in T_d : t \in c$  do
11:       $T_d \leftarrow T_d \cup T_c$ 
12:    end for
13:  end for
14:  for each document  $d \in D$  do
15:    for each target  $t \in T_d$  do
16:       $S_{d,t} \leftarrow$  ClassifyStance( $d, t$ )
17:    end for
18:  end for
19:  return  $D, T, S$ 
20: end function

```

Figure 3: Algorithm used by EC. Topic modelling is done on the flat list of stance targets using BERTopic. Removal of similar targets is based on high cosine similarity between stance target sentence embeddings.

topics of partisan disagreement. We adapt it here to finding targets of stance disagreement.

WIBA (Irani et al., 2024) uses three fine-tuned LLMs to determine whether a document features an argument, extracts the claim topic of the argument, then determines the stance of the document on that argument. In this application we remove the argument detection step, instead relying on the neutral label in stance classification. While this method is defined for argument detection, it maps neatly to stance detection. Although a more stance detection-centric method is now available (Akash et al., 2024), we use Irani et al. (2024) because it was available with an implementation at the time of this work’s inception. However, the two methods are functionally similar enough as to be interchangeable in this context.

Comparison To summarize, these three methods from the literature fulfill different features of the COSTEx task as defined in Section 1. We summarize the ways in which the representative methods—which we will evaluate here—fulfill those requirements in Table 1. As shown, none of the methods achieve all of the necessary attributes, but they

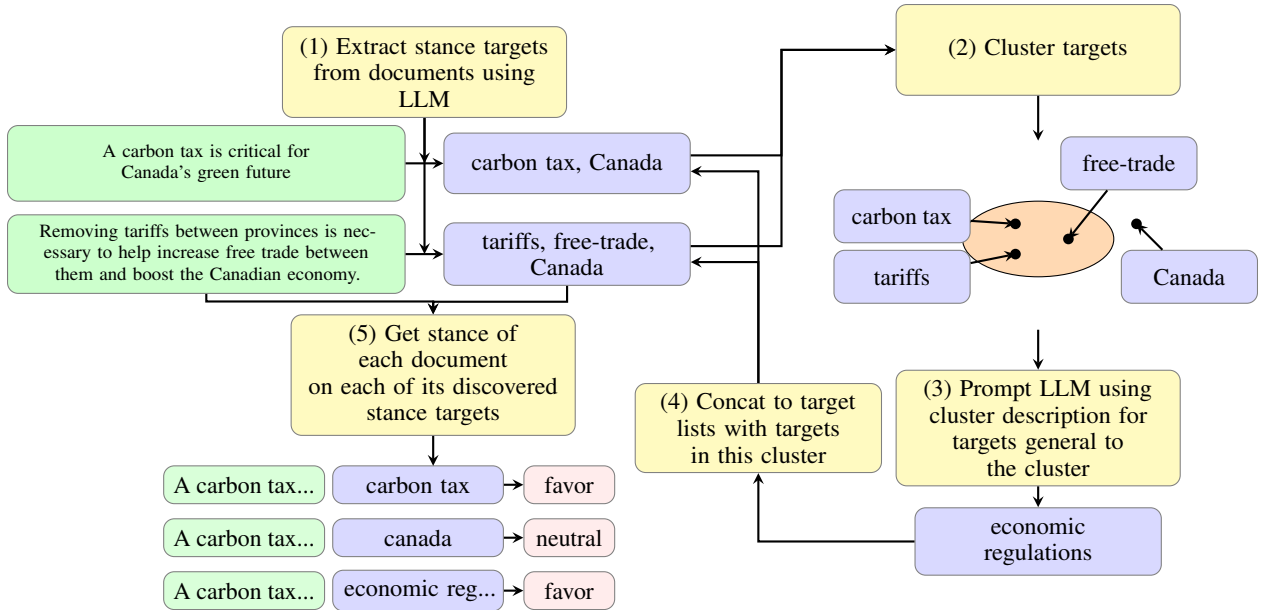


Figure 4: Simplified system diagram of the *ExtractCluster* (EC) method. Green boxes represent documents, yellow boxes represent system component steps, blue boxes represent stance targets, the orange circle represents a cluster, and pink boxes represent stance classifications. Numbers in each component step indicate the sequence of operations. We have excluded the stance target de-duplication step for brevity.

each achieve most aspects of the desired method.

5 Experiments

With our method in hand, we now want to see to what extent it fulfills COSTEx by testing it using metrics and human evaluation methods derived from our formulation, and comparing it to our comparison methods.

Datasets We use two large stance detection datasets to evaluate the methods, VAST (Allaway and McKeown, 2020) and EZ-STANCE (Zhao and Caragea, 2024). These datasets come from two domains, New York Times comments and Twitter respectively, enabling testing across diverse text types. Importantly, both datasets derive their stance targets from each document —as opposed to a dataset designed around a specifically chosen set of stance targets —allowing us to grade the produced stance targets against the annotated stance targets from the datasets. We report statistics from the datasets (Tab. 2).

Configuration We fine-tune a stance target extraction model and a stance detection model, on both VAST and EZ-STANCE for both tasks, using *Llama-3.2-1B-Instruct* as a base model (an open-weight 1B parameter model) (Meta, 2024). We use these fine-tuned models for both EC and WIBA. For diverse beam generations in EC, we sample 3 generations, as this is the ceiling integer above the

highest mean number of targets in each dataset (Tab. 2). We use a cosine similarity value of 0.8 for EC, as through manual validation, this de-duplicates stance targets that are functionally identical. We use *phi-3.5-mini-instruct* (Abdin et al., 2024) to generate higher-level stance targets for EC, a 4B model suitable for few-shot prompting. We list all other experimental implementation details of each comparison method in Appendix B.

5.1 Automated Evaluation

Metrics As previously highlighted (Sec. 3), some of the outcomes that we want to optimize in our method are trivially measurable, and some are much more difficult to measure. We therefore propose a set of metrics that assess the extent to which the method outputs optimize for the objectives defined above. While our method does produce hierarchical stance targets, in evaluation we will treat them as a flat list, while maintaining the valuable property of higher-level stance targets aggregating across more documents.

- **Target F1:** The BERTScore F1 (Zhang et al., 2019) of the discovered targets, compared to the annotated dataset, as in (Akash et al., 2024). As we have a set of annotated stance targets for each document in our labelled dataset, we compute the precision by comparing each predicted stance target to all gold

Feature	PaCTE	POLAR	WIBA	EC
Stance target discovery through aggregation	✓	✓	✗	✓
Multiple stance targets per document	✓	✓	✗	✓
Map documents to stance targets	✗	✗	✓	✓

Table 1: Comparison of different methods against our method, *EC*, for each of features 3, 2, and 4, as defined in Section 1. All of the methods fulfill feature 1.

Dataset	Num. Ex.	Mean. Num. Targets	Stance Split (F/N/A)	Lang.
VAST	784	2.45	0.47/0.02/0.51	en
EZ-STANCE	1561	1.71	0.36/0.35/0.29	en

Table 2: Statistics from the datasets used for testing.

stance targets, and the recall by comparing each gold stance target to all predicted stance targets, and compute the F1 from the resulting precision and recall, as defined in Appendix E.1. This metric measures adherence to Criterion 2.

- **Stance Retrieval F1:** The F1 of the discovered stance of the documents, compared to a labeled dataset. Seeing as we have a potentially different set of predicted stance targets as the gold stance targets, we create a mapping of predicted stance targets to gold stance targets where the sentence embedding cosine similarity is greater than 0.9, then compute the precision by comparing each predicted stance to all gold stances, the recall by comparing each gold stance to all predicted stances, and the F1 score from the precision and recall, as defined in Appendix E.2.
- **Stance Variance:** See Criterion 1.
- **Mean Num. of Targets:** See Criterion 2
- **Balanced Cluster Sizes:** See Criterion 3
- **Walltime:** Method run-time duration (s).

The supervised metrics, the target F1 and stance retrieval F1, are measuring the adherence of the method to a typical stance detection dataset. However, we also want to optimize for multi-target, hierarchical, and clustered stance targets. Optimizing for metrics that measure these aspects will reduce our target F1 score, as the stance targets will be further from the stance targets given in the base datasets. We need to assess our results holistically, and consider that, as part of our task formalization, any solution to this problem is making a trade-off between objectives. We will therefore determine

the overall ranking of the methods via a summed rank order: we find the rank of each method on every metric, sum all the ranks for each method, and the lowest summed rank order is the best method.

Results We report the supervised and unsupervised metrics from the mean of 5 runs for each method on each dataset (Tab. 3). *EC* generally outperforms other methods, except on stance target F1 and precision, and wall-time. Stance retrieval rankings are robust to varying cosine similarity, see Appendix E.2.1. We conducted ablations of cosine similarity threshold and number of beam generations and confirm that our chosen values yield the best results, see Appendix C. We also tested a version of the method that removes lower variance stance targets and find that it achieves higher mean stance variance but worse on all other metrics (See Appendix C.1), showing our method could not be trivially improved in this manner.

Human Evaluation We created two human evaluation tasks to evaluate the method outputs. The first task presents a triad of documents (a base document, another from the same cluster, and one from a different cluster) and has the annotator select which two documents go in the same stance target cluster. We measure how often the annotators agree with the document clustering chosen by each method. A second task presents a base document, and two stance target sets provided by two different methods, and a prompt asks the labeler to choose between the two stance target sets, or neither if neither are suitable. Each annotator received an evaluation guide prior to their evaluation task to explain the concepts in use in the task. We obtained 483 and 492 annotations for each task respectively, from 6 annotators who were students in the authors’ lab. We show the prompts and evaluation guide given to annotators, and example generation process in Appendix D.

To ensure there was agreement between annotators, we had two annotators evaluate the same set of 20 examples from each task. The Fleiss’ Kappa (Fleiss et al., 1981) of the stance target cluster task was 0.53, and for the stance target task it

Method	Target			Stance			Mean Num. Targets ↑	Stance Variance ↑	B. Cluster Sizes ↑	Wall time ↓
	F1 ↑	P. ↑	R. ↑	F1 ↑	P. ↑	R. ↑				
VAST										
PaCTE	0.775	0.779	0.771	0.000	0.000	0.000	1.212	0.226	2.314	155.2
POLAR	0.512	0.524	0.501	-	-	-	<u>2.140</u>	-	3.028	327.7
WIBA	0.910	0.930	0.891	0.116	0.190	0.089	1.000	0.108	7.753	246.9
EC	<u>0.897</u>	<u>0.889</u>	0.907	0.143	0.210	0.119	3.190	<u>0.136</u>	8.031	1569.1
EZSTANCE										
PaCTE	0.766	0.768	0.763	0.000	0.000	0.000	<u>1.038</u>	0.208	4.311	213.7
POLAR	0.038	0.038	0.037	-	-	-	0.218	-	0.168	<u>582.7</u>
WIBA	0.884	0.899	0.871	0.145	0.200	0.120	1.000	0.019	9.495	766.4
EC	<u>0.859</u>	<u>0.851</u>	<u>0.867</u>	0.158	0.202	0.141	3.380	<u>0.039</u>	9.520	3349.8

Table 3: Metrics comparison across datasets and methods averaged across 5 runs for each method and dataset. Best metrics are indicated with arrows. P. and R. stand for precision and recall respectively. We do not include stance results for POLAR as it does not assign stance to individual documents. Bold numbers indicate the best performance, underline indicates second best.

Method	LSR	Agree Pct.	Example Output
PaCTE	-2.23	0.19	school,health,covid...
POLAR	-2.79	0.00	anyone
WIBA	<u>1.51</u>	0.62	medical law
EC	2.23	<u>0.34</u>	jerusalem

Table 4: Luce Spectral Ranking (LSR) pairwise comparison score, calculated by comparing different methods’ stance target sets for each document, alongside an example stance target output from each method for reference (PaCTE example shown truncated). And percentage of examples where annotators agreed with the clustering of a document triad, for each method.

was 0.83, indicating inter-annotator agreement. For the stance target set comparison task, we use the Luce spectral ranking (LSR) (Maystre and Grossglauser, 2015) (via Choix¹) to determine the output stance targets most preferred by human annotators. EC and WIBA are rated the highest, with POLAR and PaCTE rated poorly (Tab. 4). For stance target cluster agreement scores, we simply record the number of times the human evaluator agreed with the method. WIBA, EC, and PaCTE obtain the best results for cluster evaluation, and POLAR obtains no agreement from evaluators (Tab. 4).

Summary We show the summed rank order of each method, for each metric, in Fig. 5. This demonstrates the overall rank of the methods on the COSTEx task we introduce in this work.

6 Discussion

POLAR needs to find many named entities to find polarized topics (being designed for news arti-

¹github.com/lucasmaystre/choix

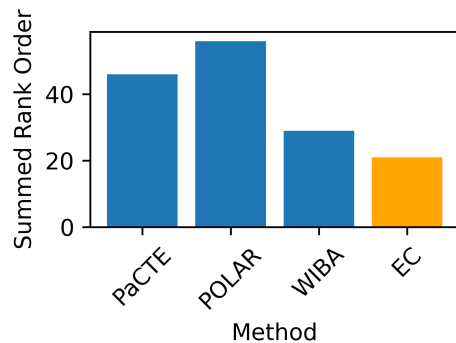


Figure 5: Summed rank order across all metrics for each method. EC outperforms the other methods we trial from the literature across our metrics.

cles), and as such performs poorly on the short text datasets used here, especially the EZ-STANCE dataset (Tab. 3). We observe poor evaluations of naming and clustering performance (Tab. 4).

PaCTE’s use of LDA topic modeling and a small classifier model means that it can quickly find large stance target clusters with high stance variance (Tab. 3). However, the naming of clusters with topic keywords results in a low evaluation score (Tab. 4), and the stance target clusters are only moderately agreed with (Tab. 4).

WIBA’s stance target extraction produces good stance targets (Tab. 3 and 4), and performs highest on our cluster agreement evaluation (Tab. 4). But the small stance target clusters it produces—due to only producing one stance target per document—result in lower stance retrieval F1, and low stance variance and cluster size (Tab. 3).

EC outperforms WIBA in stance target cluster size, stance variance, stance retrieval, and stance

target set preference (Tab. 4). However, we also see that it under-performs WIBA on our cluster agreement evaluation (Tab. 4), and stance target precision. We infer that as EC maps higher-level stance targets to each document—that have no parallel in the annotated datasets we use—which results in large clusters defined by abstract stance targets that are too general for annotators to spot in our cluster agreement exercise. Nonetheless, from the summed rank order (Fig. 5), EC is the most effective method tested here.

7 Case Study

Having empirically shown our method outperforms other methods from the literature, we chose to assess its effectiveness at identifying key characteristics of a discourse under real-world conditions. Topic modelling is frequently used for exploratory analysis of discourse corpuses (Hobson et al., 2024; Falkenberg et al., 2022). Although it does not dis-aggregate expressed valence—a key part of separating discourse (Ghafouri et al., 2024)—the previously missing step of stance detection—stance target discovery—makes it labour-intensive to run as a go-to exploratory step. Crucially, both EC and BERTopic (Grootendorst, 2022) require no notable parameter tuning and, so, are of equal complexity for a domain researcher to use.

We assumed the role of a researcher studying the political views present in a social media dataset. We chose a 2024 Twitter dataset consisting of 1.4m tweets (81% English, 9% French, 10% other languages) from 1.9k prominent Canadian media accounts (Pehlivan et al., 2025). See Appendix F for implementation details. Table 5 shows the largest stance target vs. topic clusters.

Meaningful clusters. Both stance target and topic modeling methods can produce nonsensical clusters. How do we quickly remove the noise? In topic modeling, this is messy: as seen in Table 5, some of the largest topic clusters are meaningless (e.g., “shes, shell, shed, quelle”). In contrast, with EC, an easy way of filtering weak stance targets is by simply dropping small stance clusters, with the intuition being that modal stance targets are more frequently good stance targets. In this case study, we found removing stance targets with less than 50 data-points to be a good level. At this border, there are some good stance targets (‘Organic Food Movement’, and ‘US Col. Lawrence Wilkerson’) but also many non-specific or nonsensical stance targets (‘which will’, ‘candidate nomination’).

Cluster informativeness. Table 5 highlights the informativeness of EC clusters in several ways. First, stance target clusters capture more of the documents than the largest topics, due to EC allowing documents to belong to multiple stance target clusters. If we kept just the first extracted stance target (as previously (Akash et al., 2024)), the ‘trudeau’ stance target would only be assigned to 22k documents, with our method allowing us to know the stance of more documents on ‘trudeau’ where he may be referred to implicitly. Second, the stance targets capture the large ongoing issues of Canadian public discourse (Canada, Justin Trudeau, the Liberal Party), and topical issues (Donald Trump’s presidency, the B.C. election, the Israel-Palestine conflict), where these large issues are missed by the topics - instead emphasizing smaller topics like the Olympics. Even for topic clusters that are not “noise”, the stance target names are consistently more specific, and therefore more usable for further analysis. However, EC needs improved stance target de-duplication, as shown by the presence of ‘j. trudeau’ and ‘trudeau’.

Understanding stance on the target clusters.

We show a map of the 30 largest stance target clusters in Fig. 6. Having stance classifications on so many targets surfaces key aspects of the discourse: allowing us to compare mean stance on party leaders (-0.57 for Trudeau vs. -0.44 for Poilievre), parties (-0.45 for the NDP vs. -0.62 for the Liberal Party), and foreign policy issues (-0.46 for Israel vs. -0.79 for Hamas) with one method application (where we have substituted ‘favor’ for 1, ‘neutral’ for 0, and ‘against’ for -1).

This case study highlights how EC gave the researcher a larger and more detailed map of the discussion in our dataset, alongside more specific and understandable cluster names.

8 Conclusion

We have motivated and conceptualized the task of *COSTEx*, and shown that our new method for this task, *EC*, outperforms previous methods for similar tasks. We then used a large-scale real-world dataset to demonstrate that our method reliably captures and represents clusters of stance target discussion. We hope that this method can aid practitioners in quickly understanding discourse in large and wide-ranging real-world datasets, and help improve understanding of complex behaviors such as polarization and public opinion in our quickly changing information environments.

Stance Targets		Topics	
Name	Count	Name	Count
canada	76k	gaza, israel, israeli, hamas	22k
j. trudeau	54k	olympics, game, olympic, athletes	15k
trudeau	39k	hes, guy, coyne, mrstache9	10k
trump presidency	29k	url, juliemarienolke, thejagmeetsingh, saudet80	9k
liberal party	22k	healthcare, nurses, doctors, doctor	9k
israeli	17k	shes, shell, shed, quelle	8k
trump	17k	housing, rent, rental, homes	7k
b.c. ndp	16k	trudeau, justin, trudeaus, resign	6k

Table 5: Comparing largest stance target clusters to largest topic clusters.

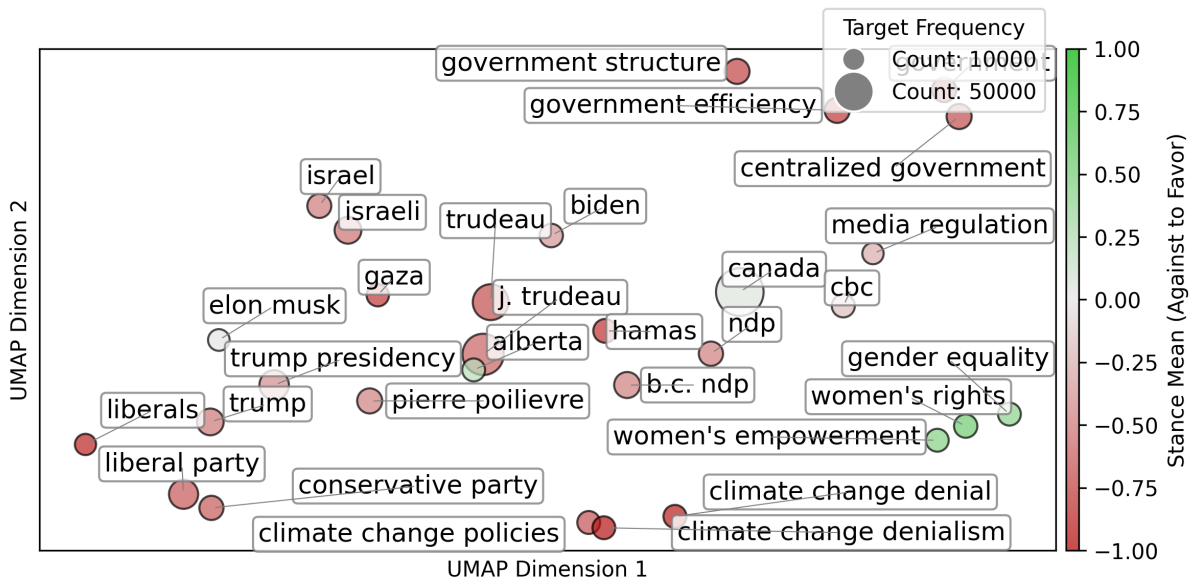


Figure 6: Map of top stance targets, sized by frequency, coloured by average stance. In general, the major issues that Canadian social media users tend to have attitudes on are represented. However, we can also see that improved stance target de-duplication is necessary, along stance target clarity (‘climate change denialism’).

9 Limitations

Datasets We found that the lack of hierarchical stance targets in the stance-detection datasets used in this work made it difficult to evaluate the ability of the methods to find a full breadth of hierarchical, clustered stance targets for each document. We can only use these datasets to assess the extent to which the method found the base stance targets for each document. Future work should develop new datasets to evaluate higher-level stance targets.

Methods Stance target de-duplication became an issue when we applied our method to a larger corpus. We experimented with using DBSCAN to some success, but de-duplicating different ways of spelling names (‘j. trudeau’, ‘trudeau’) while avoiding false positives requires a carefully set distance threshold between embeddings. Additionally, our method of using diverse generation to generate multiple stance targets for each document —while

not requiring re-training of our stance target generation model —could be made faster and more flexible by generating targets as a list.

Task Formulation Optimizing for stance variance deprioritizes stance targets that are generally agreed upon, but when disagreed upon, are interesting, such as conspiracy theories, so optimizing for this metric is a trade-off.

Another key limitation was a lack of a principled framework for defining a hierarchy of targets. In practice, LLM prompting produced sufficiently useful results here, but a more well-defined definition could produce stronger results.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat

- Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. 2024. Can large language models address open-target stance detection? *arXiv preprint arXiv:2409.00222*.
- Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.
- European Commission. 2024. [Commission opens formal proceedings against tiktok on election risks under the digital services act](#).
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociochi, et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121.
- Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.
- Tomoki Fukuma, Koki Noda, Hiroki Kumagai, Hiroki Yamamoto, Yoshiharu Ichikawa, Kyosuke Kambe, Yu Maubuchi, and Fujio Toriumi. 2022. How many tweets does one need?: Efficient mining of short-term polarized topics on twitter: A case study from japan. *arXiv preprint arXiv:2211.16305*.
- Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.
- Vahid Ghafouri, Jose Such, Guillermo Suarez-Tangil, et al. 2024. I love pineapple on pizza! = i hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Empirical Methods in Natural Language Processing*.
- Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.
- Jeffrey Gottfried. 2024. Americans’ social media use. *Pew Research Center*, 31.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Zihao He, Negar Mokherian, António Câmara, Andrés Abeliuk, and Kristina Lerman. 2021. Detecting polarized topics using partisanship-aware contextualized topic embeddings. *arXiv preprint arXiv:2104.07814*.
- David Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032.
- Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23.
- Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024. Wiba: What is being argued? a comprehensive approach to argument mining. *arXiv preprint arXiv:2405.00828*.
- Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. 2023. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*.
- Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

- Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28.
- Meta. 2024. [Llama 3.2: Revolutionizing edge ai and vision with open, customizable models](#).
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.
- Demetris Paschalides, George Pallis, and Marios D Dikaiakos. 2021. Polar: a holistic framework for the modelling of polarization and identification of polarizing topics in news media. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 348–355.
- Zeynep Pehlivan, Saewon Park, Alexei Sisulu Abrahams, Mika Jacques Patel Desblancs, Benjamin David Steel, and Aengus Bridgman. 2025. Can-polnews: A multi-platform dataset of political discourse in canada. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2550–2559.
- Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.
- 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*. Association for Computational Linguistics.
- Florian Saurwein and Charlotte Spencer-Smith. 2021. Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, 9(4):222–233.
- Benjamin Steel and Derek Ruths. 2024. Multi-target user stance discovery on reddit. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 200–214.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. 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 504–510.
- Chenye Zhao and Cornelia Caragea. 2024. Ez-stance: A large dataset for english zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714.

A Methods

In addition to the method we propose in this work, we also trialled a method we call *ClusterExtract*, inspired by PaCTE. It starts by finding hierarchical topics in the corpus using BERTopic (Grootendorst, 2022), then assigns stance targets to each topic. It is described in Algorithm 1. However, we found that it produced inferior results to *EC*, and so do not detail it in the main results of the work.

B Implementations

B.1 POLAR

We used all of the default parameter settings and models for POLAR for the VAST dataset, but for EZ-STANCE, we reduce the noun phrase clustering threshold from 0.8 to 0.6, as the default value was resulting in no found clusters given that the EZ-STANCE dataset is composed of low word count tweets, which have low entity mention counts.

In adapting this method, we need to extend it by mapping the chosen polarized topics back to the documents, to allow our metrics to be applied to the results. We do so by considering a document to be in a stance target cluster when it features a polarized entity, and the discovered noun phrases as the stance targets.

B.2 PaCTE

We train the PaCTE BERT model (Devlin, 2018) using the combined training sets from VAST and EZ-STANCE, removing all neutral examples as the original implementation was only trained on partisan news.

We use online latent dirichlet allocation (LDA) (Hoffman et al., 2010) as a drop in method speed-up, instead of the original single-core method. Other implementation details are all the same as the original implementation.

Algorithm 1 Algorithm used by *ClusterExtract*.

Require: Documents D

```
1: function CLUSTEREXTRACT( $D$ )
2:    $C \leftarrow$  TopicModelDocs( $D$ )
3:    $\triangleright$  Handle outlier documents (Topic = -1)
4:    $D_{out} \leftarrow$  FilterOutliers( $D, C$ )
5:   for each document  $d \in D_{out}$  do
6:      $T_d \leftarrow$  ExtractStanceTargets( $d$ )
7:      $T_d \leftarrow$  RemoveSimilarTargets( $T_d$ )
8:   end for
9:    $\triangleright$  Handle non-outlier documents
10:  for each cluster  $c \in C$  do
11:     $T_c \leftarrow$  ExtractClusterStanceTargets( $c$ )
12:     $T_c \leftarrow$  RemoveSimilarTargets( $T_c$ )
13:  end for
14:   $\triangleright$  Generate hierarchical topic targets
15:   $H \leftarrow$  GetHierarchicalTopics( $T$ )
16:  for each parent cluster  $c \in H$  do
17:     $C_p \leftarrow$  GetChildTopics( $c$ )
18:     $T_p \leftarrow$  AggregateChildTargets( $C_p$ )
19:     $T_p \leftarrow$  RemoveSimilarTargets( $T_p$ )
20:  end for
21:   $\triangleright$  Combine targets and remove duplicates
22:  for each document  $d \in D$  do
23:    if  $d \notin D_{out}$  then
24:       $c \leftarrow$  GetDocumentCluster( $d$ )
25:       $p \leftarrow$  GetParentCluster( $c$ )
26:       $T_d \leftarrow T_c \cup T_p$ 
27:       $T_d \leftarrow$  RemoveSimilarTargets( $T_d$ )
28:    end if
29:    for each target  $t \in T_d$  do
30:       $S_{d,t} \leftarrow$  DetermineStance( $d, t$ )
31:    end for
32:  end for
33:  return  $D, T, S$ 
34: end function
```

B.3 WIBA

We used Llama 3.2 1B (Meta, 2024) as the base LLM for our implementation of WIBA, for its trade-off of performance with small size. Training used the combined VAST and EZ-STANCE train/validation sets. On the combined test sets, it achieved a stance detection F1 of 71.5%, and for stance target extraction it obtained a BERTScore of 90.3%, comparable with the metrics achieved in the original work.

We replaced the system and instruction tuning tokens with a chat template as appropriate for the Llama model. We used a cosine learning rate with warmup that increments every step (Loshchilov and Hutter, 2016), and NEFTune to improve fine-tuned accuracy (Jain et al., 2023). We trained on a 24GB NVIDIA GPU, training took roughly 8 hours.

B.4 EC

For diverse generation, we generate 3 return sequences, by exploring 3 beam groups using 6 beams, with a diversity penalty of 10.0. We use a no repeat n-gram size of 2 to prevent repetition.

We use the *paraphrase-MiniLM-L6-v2* sentence transformer model (Reimers and Gurevych, 2019) to embed candidate stance targets, and remove a target from pairs that have a cosine similarity of higher than 0.8.

We run ablation experiments on the number of beam groups and the cosine similarity threshold used for stance target de-duplication in Section C.

B.5 Datasets

When using VAST as a comparison dataset for the methods, we remove the synthetic neutral examples, as these targets aren't specific for each document. We do however use the synthetic neutral examples to train our stance detection model.

B.6 Prompts

We include the few-shot prompt used for stance target extraction from topic clusters in Prompt 1:

Prompt 1: Prompt used for extracting stance targets from a topic cluster.

⚙️ System:

You are an expert at analyzing discussions across multiple documents.

👤 Human:

Your task is to identify a common stance target that multiple documents are expressing opinions about.

Instructions:

1. Read all provided documents
2. Identify topics that appear across multiple documents
3. Determine if there is a shared target that documents are taking stances on
4. Express the target as a clear noun phrase

Input:

Documents: [list of texts]

Output:

Stance target: [noun phrase or "None"]

Reasoning: [2-3 sentences explaining the choice]

Examples:

Example 1:

Documents:

"The council's new parking fees are excessive. Downtown businesses will suffer as shoppers avoid the area."

"Increased parking rates will encourage public transit use. This is exactly what our city needs."

"Local restaurant owners report 20% fewer customers since the parking fee increase."

Output:

Stance target: downtown parking fees

Reasoning: All three documents discuss the impact of new parking fees, though from different angles. The documents show varying stances on this policy change's effects on

business and transportation behavior."",

Example 2:

Documents:

"Beijing saw clear skies yesterday as wind cleared the air." "Traffic was unusually light on Monday due to the holiday." "New subway line construction continues on schedule."

Output:

Stance target: None

Reasoning: While all documents relate to urban conditions, they discuss different aspects with no common target for stance-taking. The texts are primarily descriptive rather than expressing stances.

Example 3:

Documents:

"AI art tools make creativity accessible to everyone."

"Generated images lack the soul of human-made art."

"Artists demand proper attribution when AI models use their work."

Output:

Stance target: AI-generated art

Reasoning: The documents all address AI's role in art creation, discussing its benefits, limitations, and ethical implications. While covering different aspects, they all take stances on AI's place in artistic creation.

Documents:

{formatted_docs}

🤖 Assistant:

Output:

Stance target:

We include the few-shot prompt used for aggregating stance targets in Prompt 2:

Prompt 2: 3-shot in-context prompt for aggregating stance target clusters.

⚙️ System:

You are an expert at analyzing and categorizing topics.

👤 Human:

Your task is to generate a generalized stance target that best represents a cluster of related specific stance targets.

Instructions:

1. Review the provided stance targets and keywords that characterize the topic cluster
2. Identify the common theme or broader issue these targets relate to
3. Generate a concise noun phrase that:
 - Captures the core concept shared across the targets
 - Is general enough to encompass the specific instances
 - Is specific enough to be meaningful for stance analysis

Input:

Representative stance targets: [list of stance targets]

Top keywords: [list of high tf-idf terms]

Output format:

Generalized target: [noun phrase]

Reasoning: [1-2 sentences explaining why this generalization fits]

Examples:

Input:

Representative stance targets: ["vaccine mandates", "mandatory covid shots", "required immunization for schools"]

Top keywords: ["mandatory", "requirement", "public health", "immunization", "vaccination"]

Output:

Generalized target: vaccination requirements

Reasoning: This captures the common theme of mandatory immunization policies while being broad enough to cover various contexts (workplace, school, public spaces).

Input:

Representative stance targets: ["EVs in cities", "gas car phase-out", "zero emission zones"]

Top keywords: ["emissions", "vehicles", "transportation", "electric", "fossil-fuel"]

Output:

Generalized target: vehicle electrification

Reasoning: This encompasses various aspects of transitioning from gas to electric vehicles, including both the technology and policy dimensions.

Input:

Representative stance targets: ["content moderation", "online censorship", "platform guidelines"]

Top keywords: ["social media", "guidelines", "content", "moderation", "posts"]

Output:

Generalized target: social media content control

Reasoning: This captures the broader issue of managing online content while remaining neutral on the specific approach or implementation.

Representative stance targets: {repr_docs}

Top keywords: {keywords}

🤖 Assistant:

Output:

Generalized target:

C Ablations

We re-ran EC for varying values of the number of beam generations and cosine similarity threshold for stance target de-duplication to explore the impact it had on method outputs. Figs. 7 and 8 shows that 3 generated beam groups produces consistently the best results on our automated metrics (other the number of beam groups and wall-time being linearly directly correlated), out of 2, 3, and 5 as possible values. Figs. 9 and 10 shows that varying the cosine similarity threshold between values of 0.8, 0.9, and 0.95 has minimal effect on the final metrics.

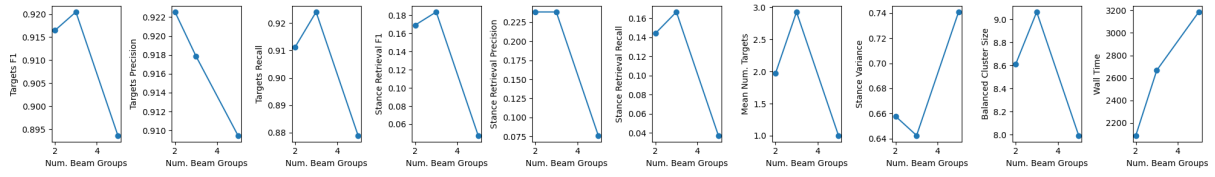


Figure 7: Ablation of the number of beam group generations used for EC for VAST.

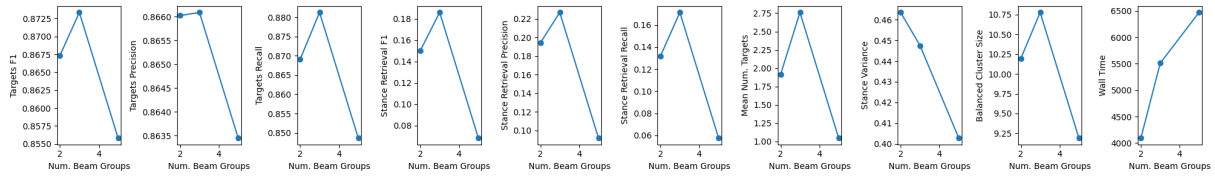


Figure 8: Ablation of the number of beam group generations used for EC for EZ-STANCE.

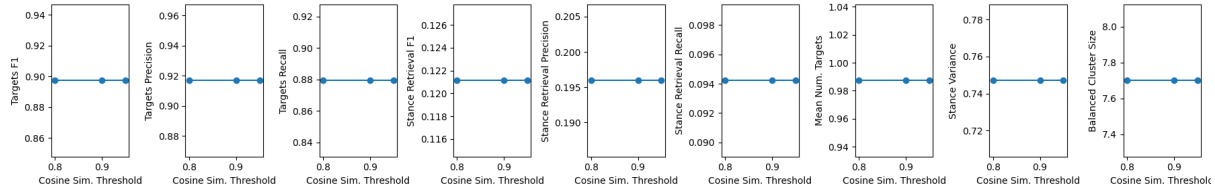


Figure 9: Ablation of the cosine similarity threshold used for stance target de-duplication in EC for VAST.

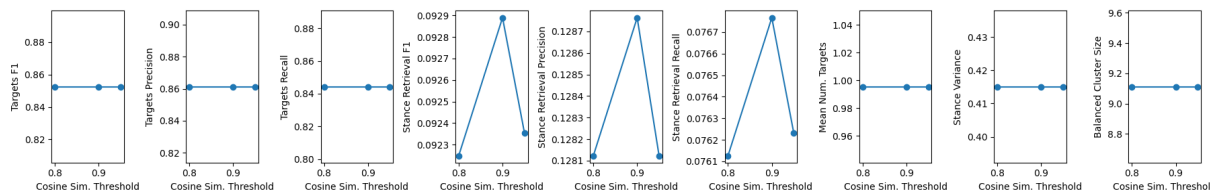


Figure 10: Ablation of the cosine similarity threshold used for stance target de-duplication in EC for EZ-STANCE.

C.1 Stance Variance

Our stance variance metric could be optimized to through de-generate solutions. We wanted to determine the effect that solely optimizing for this metric would have on other metrics. We ran an experiment where we kept only stance targets that had over the 75th and 90th percentiles. Keeping only stance targets over the 75th stance variance percentile did not impact the mean stance variance (0%) change, and reduced stance retrieval F1 by 4.5%. Keeping only stance targets over the 90th stance variance percentile increased mean stance variance by 33%, but decreased stance retrieval F1 (-58%), stance target F1 (-37%), balanced cluster size (-27%), and the mean number of targets per document (-53%).

D Human Evaluation

Human evaluators were fellow students from the authors' lab.

We provided each annotator this explanatory document prior to their evaluation task to help them understand the concepts in Prompt 3.

Prompt 3: Evaluation Guide

What is a stance target? A stance target is a concept that one can have an opinion on. While one can technically have an opinion on almost anything (i.e. one can technically be for or against atoms, but we generally do not consider atoms to be an issue that one is for or against), there are a more constrained set of concepts that we generally put forth opinions, or stances, on.

What is a topic? In computer science, defined as a set of frequently co-occurring words. More generally, synonymous with a theme, or subject that a document can reference or be about. A document can have multiple topics. It is an abstract concept.

Stance Targets So for example, the text: 'I discussed my preference for tariffs over free trade while playing golf today at mar-a-lago' There are 4 prominent concepts: tariffs, free trade, golf, and mar-a-lago. Two topics for this text would be trade policy (tariffs, free trade), and golf (golf, mar-a-lago), as these are frequently co-occurring words/concepts. The two prominent stance targets are tariffs and free-trade, as they are discussed in the context of having a position on them, and are things that one generally has a stance on. Golf could also be considered a stance target in this context, but is discussed with less emphasis on stance.

Stance targets can also exist at a higher conceptual level. For example, here the author is expressing not only their preference for tariffs, but economic regulation, and protectionism. In this way, the most representative set of stance targets for this text would be 'tariffs', 'free trade', 'economic regulations', and 'protectionism'

One can discuss a stance target while staying

neutral. For example: ‘I read about the idea of tariffs recently. Undecided on whether or not they’re effective’. This author is neutral on the stance target of tariffs.

Stance Target Clusters Two documents fit in the same stance target cluster, if they discuss the same stance target, whatever the conceptual level of that stance target. The two documents may both be favoring the stance target, on opposing sides of the stance target, or both neutral on the stance target.

For example, the texts: ‘I think tariffs are a terrible idea’ ‘Taxes should be much higher!’ Are not in the same ‘tariffs’ stance target cluster, but are in a stance target cluster: ‘economic regulations’

The exact text prompt given to human evaluators for the stance target cluster comparison task is shown in Prompt 4:

Prompt 4: Stance target cluster comparison prompt

Which document discusses a stance target that the base document is also discussing? If both documents discuss completely different stance targets from the base document, choose neither.

To generate triads, for each method and document from both datasets, we randomly sample a document that is a stance target cluster that the base document is also in, and randomly sample a document that is not in any of the same stance target clusters. If the method does not place the base document in a stance target cluster with any other document, then two documents that are not in the same stance cluster are sampled. The order of the two comparison documents is randomly swapped to prevent the chosen document being inferred from the order. We then simply check if the annotator agrees with the method.

The exact text prompt given to human evaluators for the stance target comparison task is shown in Prompt 5:

Prompt 5: Stance target comparison prompt

Compare the two sets of stance targets, and choose the set that better covers the stance targets the document discusses. If neither sets fit at all, choose neither.

We sample comparisons from the set of all pairwise stance target set comparisons between methods for all documents from both methods. We randomly swap the order of these sets to ensure the same method does not always appear on the same side.

E Metrics

E.1 Stance Target F1

For the stance target BERTScore, given a set of documents D where each document d has predicted targets P_d and gold targets G_d , we compute the precision, recall and F1 as:

$$P = \frac{1}{|D|} \sum \frac{\sum^{P_d} \max_{g \in G_d} \text{BERTScore}(p, g)}{|P_d|}$$

$$R = \frac{1}{|D|} \sum \frac{\sum^{G_d} \max_{p \in P_d} \text{BERTScore}(g, p)}{|G_d|}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

E.2 Stance Retrieval F1

Given a set of documents D , where each document d has predicted target-stance pairs $P_d = \{(t, s)\}$, and gold target-stance pairs $G_d = \{(t, s)\}$, where stance can be any of $\{favor, against, neutral\}$.

We define a mapping between predicted stance targets and gold stance targets, where stance targets are only mapped to each other if their sentence embedding cosine similarity is higher than $\theta = 0.9$:

$$M = \{(t_p, t_g) : \max_{t' \in G} \text{sim}(t_p, t') \wedge \text{sim}(t_p, t_g) \geq \theta\}$$

For each document d , define the set of correct predictions:

$$C_d = \{(t_p, s) \in P_d : \exists (t_g, s) \in G_d, (t_p, t_g) \in M\}$$

Then:

$$P = \frac{1}{|D|} \sum_{d \in D} \frac{|C_d|}{|P_d|}$$

$$R = \frac{1}{|D|} \sum_{d \in D} \frac{|C_d|}{|G_d|}$$

$$F1 = \frac{2PR}{P + R}$$

E.2.1 Threshold Sensitivity

We looked at the sensitivity of our stance retrieval metrics to the chosen cosine similarity parameter, as seen in Fig. 11. The rankings of the method are robust to varying values of the chosen cosine similarity.

F Case Study Implementation

When deploying EC at scale in the case study, we use smaller models: *SmolLM2-360M-Instruct*² to generate the base targets, and *SmolLM2-135M-Instruct*³ to classify stance. Although this makes applying this method to large datasets more tractable, it occasionally results in poor stance targets. This problem is alleviated by using a strong model for the higher-level stance target generation (huggingface.co/microsoft/Phi-4-mini-instruct).

²huggingface.co/HuggingFaceTB/SmolLM2-360M-Instruct

³huggingface.co/HuggingFaceTB/SmolLM2-135M-Instruct

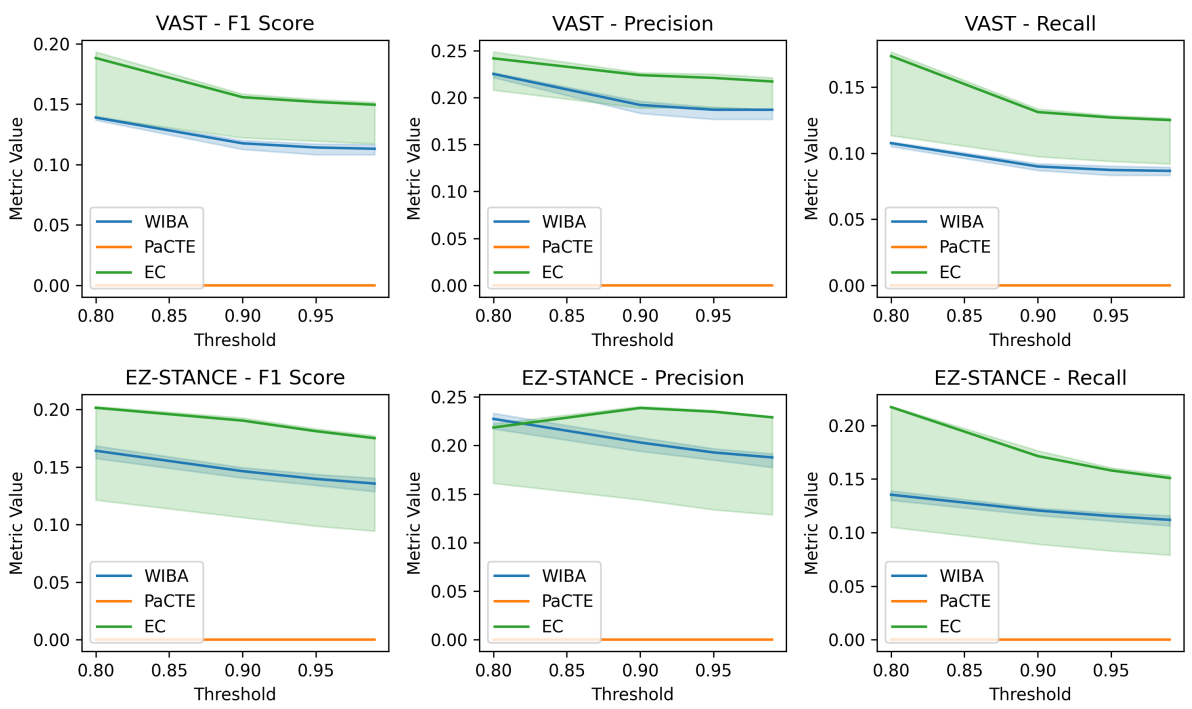


Figure 11: Varying values of the cosine similarity parameter used for calculating stance retrieval against the final value of the metric. Method ranking remains robust to the varying parameter. Shown are the median and quartiles for 5 outputs of each method.

Information-Theoretic and Prompt-Based Evaluation of Discourse Connective Edits in Instructional Text Revisions

Berfin Aktas

Natural Language Understanding Lab
University of Technology Nuremberg
berfin.aktas@utn.de

Michael Roth

Natural Language Understanding Lab
University of Technology Nuremberg
michael.roth@utn.de

Abstract

We present a dataset of text revisions involving the deletion or replacement of discourse connectives. Manual annotation of a replacement subset reveals that only 19% of edits were judged either necessary or should be left unchanged, with the rest appearing optional. Surprisal metrics from GPT-2 token probabilities and prompt-based predictions from GPT-4.1 correlate with these judgments, particularly in such clear cases.

1 Introduction

Discourse relations are essential for maintaining coherence and logical flow in text. This is especially critical in instructional texts such as how-to guides, where a lack of clarity can lead to misinterpretation and misunderstanding (Roth et al., 2022; Aktas and Roth, 2025). Discourse relations are often signaled explicitly through discourse connectives like “because” or “however.” In the case of implicit discourse relations, where coherence is inferred from context rather than without an overt connective, interpretation becomes more ambiguous. Even with explicit connectives, alternative connectives may sometimes express a relation more clearly or appropriately. Identifying when a connective is truly necessary, when it may be redundant or misleading, and which connective best fits the context remains a challenge in discourse processing.

We investigate revisions in which discourse connectives are inserted, replaced or deleted, and explore whether LLMs (GPT-2, GPT-4.1, Mistral-7B), can offer useful cues. Specifically, we explore whether an information-theoretic measure such as Shannon (1948)’s *surprisal* can be used to assess the necessity or appropriateness of a discourse connective in context. Our hypothesis is that connectives inserted during revision (i.e., judged necessary) and those deleted (i.e., judged unnecessary) affect the predictability of the text in

different ways and these effects can be captured through surprisal-based measures such as *average surprisal*, *variance*, and *smoothness* of surprisal change.

An initial analysis of information-theoretic measures over *all* revisions regarding connectives revealed inconsistent patterns: the presence or absence of a discourse connective did not consistently affect surprisal-based metrics (see Appendix B). To further investigate this, we conduct a qualitative study on a subset of connective replacements, which we present in Section 4. Manual annotation of this subset reveals that not all replacements serve the same function: some are optional, others are essential for coherence, and some may even be inappropriate. We apply two complementary approaches to uncover underlying patterns computationally: information-theoretic analysis using GPT-2 (Radford et al., 2019) and prompting-based evaluation with more recent models, namely GPT-4.1 (OpenAI et al., 2024) and Mistral-7B (Jiang et al., 2023). We discuss the patterns uncovered by these methods in Section 5, arguing that comparing these approaches provides valuable insights into the interpretability of model behavior.

Although our exploration of information-theoretic measures with language models yields largely inconclusive results, we believe this line of investigation remains promising. As language models become more transparent in exposing token-level probabilities and better at modeling discourse, information-theoretic metrics may offer an interpretable framework for assessing connective necessity and discourse coherence. We view this as an interesting direction for future research.

We investigate the following research questions:

- RQ1: Can information-theoretic metrics reveal when replacing a connective improves coherence?
- RQ2: To what extent do information-theoretic

and prompt-based methods align in their interpretation of discourse relations?

2 Related Work

Information *surprisal* has been extensively used in computational linguistics to model processing difficulty and expectations in language comprehension (e.g., Levy, 2008; Clark et al., 2023; Oh and Schuler, 2023). From a discourse perspective, Torabi Asr and Demberg (2015) examine the role of discourse connectives through the lens of the Uniform Information Density (UID) hypothesis. They show that connectives can help distribute information more evenly.

Recently, Aktas and Roth (2025) examine discourse connective insertions in revision data and find that multiple relations are often plausible, while language models perform inconsistently in detecting ambiguity. This underscores the challenge of identifying the necessity of connectives and motivates the need for complementary metrics.

3 Overview

In this section, we provide a brief overview of our study design. We introduce the dataset of connective edits extracted from WikiHow revisions (Section 4), including deletions, replacements, and manually annotated subsets of these edits used to assess their function. Building on this, we conduct experiments on the annotated replacement instances, comparing information-theoretic measures with prompt-based judgments from large language models (Section 5). Together, these components allow us to investigate how connective edits affect discourse coherence and how well computational models capture these preferences.

4 Data

As a framework for discourse relations, we adopt the Penn Discourse Treebank (PDTB) (Prasad et al., 2018) and use its inventory of discourse connectives. Building on prior work by Aktas and Roth (2025), we use the wikiHowToImprove dataset (Anthonio et al., 2020), which contains sentence-level revisions. While that prior study primarily focused on connective insertions, we target two complementary operations: the deletion of existing connectives and the replacement of one connective with another. We introduce a subset of instructional text revisions where explicit discourse connectives are ei-

ther deleted from the beginning of a sentence (§4.1) or replaced with another (§4.2).

4.1 Connective Deletions

We identified 13,597 instances where discourse connectives were deleted from the beginning of a sentence, as illustrated in the top row of Table 1. In a quantitative analysis, we find the most commonly deleted connectives to be *then*, *also*, and *and*.¹ To better understand these deletions, we qualitatively analyzed 100 examples (four for each of the 25 most frequently deleted discourse connectives). In 64% of cases, the connective was redundant or its removal improved sentence flow. In 21%, deletion was necessary to resolve syntactic or interpretability issues. In the remaining 15%, deletion should not happen as it is altering the intended meaning.²

We examined the connective deletion statistics in comparison to the connective insertion statistics provided by Aktas and Roth (2025). Across the dataset, there is a clear asymmetry between these two types of revisions: the number of connective deletions (13,597) significantly exceeds the number of insertions (4,274).

As shown in Appendix E, connectives such as *then*, *finally*, *and*, and *so* are deleted far more often than inserted, suggesting they are often perceived as redundant. This supports the findings of Torabi Asr and Demberg (2012), who argue that causal, temporal, and additive relations (when not marking an event shift) are frequently left implicit. In contrast, connectives like *for example* and *in addition* are more often inserted, pointing to their role in improving clarity. Interestingly, although *for example* signals an “instantiation” relation in PDTB, which Torabi Asr and Demberg (2012) describe as typically implicit, our data shows frequent explicit realization. This discrepancy suggests the need for further investigation, possibly by analyzing the full dataset in terms of discourse relations, rather than focusing only on those made explicit or implicit through sentence-initial edits.

4.2 Connective Replacements

We identified 1,841 revisions in which discourse connectives occurring at the beginning of a sentence were replaced by another connective, as

¹We list the 15 most frequently deleted connectives in Table 7 in Appendix E.

²For an example, see Figure 3 in Appendix A where deleting the contrastive connective “Otherwise” causes the instruction to lose its conditional framing.

Deletion	Cut the four shirt pieces out of the sheet material. Then place those pieces on the lining material, face side up, and pin them into place before cutting them out.
Replacement	Pick a music choice that agrees with your style choice. So <u>For example</u> , if your dance is fast, pick a fast song.

Table 1: Examples of connective edits in our data. In the first example, the connective is deleted (*Then*), while in the second example the connective (*So*) is replaced by another connective (*For example*).

Revision necessary	The revised version is the only option that conveys the necessary meaning or fits the syntactic context.	11 cases
Revision better	Both options are plausible, but the revision is preferred in terms of clarity, formality, or fluency.	35 cases
Either way	Both the original and revised versions are similarly acceptable.	36 cases
Original better	Both options are plausible, but the original is preferred in terms of clarity, formality, or fluency.	10 cases
Original should stay	The original version is the only option that conveys the necessary meaning or fits the syntactic context.	8 cases

Table 2: Description of labels for annotation and absolute counts after aggregation by majority vote.

illustrated in the bottom row of Table 1. The most common replacement is *but* \rightarrow *however*.³ While both connectives share the same dominant PDTB sense, COMPARISON.CONCESSION.ARG2-AS-DENIER, *however* is slightly less ambiguous, linked to only 8 different PDTB senses compared to 13 for *but*, and arguably better aligned with the formal tone of instructional texts.

However, reduced ambiguity or stylistic considerations do not fully explain all replacement choices. For instance, *since* is more ambiguous than *because*, yet *because* \rightarrow *since* appears among the top five replacement pairs. Apart from replacements like *if* \leftrightarrow *even if*, which reflect a clear shift in discourse relation, most replacements preserve the original PDTB sense. This suggests many changes are guided more by tone, readability, or syntactic fit rather than discourse semantics. To gain deeper insight into the nature of these replacements, we analyze four randomly selected instances for each of the top 25 replacement pairs.⁴ As this data is used for experiments (§5), we collect three independent annotations for each instance to ensure data quality and aggregate them by majority vote.⁵

The labels and aggregated counts are listed in Table 2. The average agreement with the majority

label is 74%. The high frequency of the *Either_way* label (36%) is expected as the original and revised connectives convey the same discourse relation in 48% of cases.⁶ By contrast, only 11% of the revisions are annotated as clearly necessary, and 8% are seen as inappropriate. The remaining cases appear to reflect stylistic or contextual preferences rather than discourse-level necessity.

5 Pilot Study on Replacements

Using the data described in Section 4.2, we evaluate whether language models can detect subtle discourse preferences between original and revised versions of connectives. We compare two different methodologies: (1) an information-theoretic approach based on token-level log-likelihoods, and (2) prompting-based judgments from more recent large language models. While the former offers an interpretable output grounded in language predictability, the latter captures higher-level discourse reasoning not available through raw probabilities. We describe information-theoretic measures (§5.1) before turning to a comparative evaluation (§5.2).

5.1 Information-Theoretic Measures

To measure a language model’s uncertainty and the predictability of text, we use *surprisal*. For GPT-2, the input is first tokenized, and then, the model

³Frequent pairs are shown in Table 8 in Appendix E.

⁴Since this dataset includes only edits at the beginning of sentences, all annotated cases involve sentence-initial edits.

⁵One annotator is an author of this paper, whereas the other two are PhD students in Computational Linguistics.

⁶As determined by the most common PDTB sense annotations (e.g., *once* and *after* both signal *temporal.asynchronous.succession*).

computes the probability of each token given its preceding context. The *surprisal* of each token is then calculated as the negative base-2 logarithm of its predicted probability.

Average Token Surprisal For each token t_i , its *surprisal* $S(t_i|t_{<i})$ measures how unexpected it is given the preceding context $t_{<i}$ according to the language model:

$$S(t_i|t_{<i}) = -\log_2 P(t_i|t_{<i})$$

The *average token surprisal* for a sequence $T = (t_1, t_2, \dots, t_N)$ is:

$$\text{Avg } S(T) = \frac{1}{N} \sum_{i=1}^N S(t_i|t_{<i})$$

This value reflects the overall predictability of the sequence for the language model, with lower values indicating greater predictability.

Variance of Surprisal The variance of surprisal, $\text{Var}(S)$, quantifies the spread of token-wise surprisal values around their mean:

$$\text{Var}(S) = \frac{1}{N} \sum_{i=1}^N (S(t_i | t_{<i}) - \text{Avg}(S))^2$$

Higher variance indicates larger fluctuations in predictability across tokens, whereas lower variance suggests a more uniform distribution.

Smoothness of Surprisal Surprisal smoothness captures how abruptly the language model’s predictability shifts from one token to the next. It is defined as the mean absolute difference between the surprisals of consecutive tokens:

$$\text{Smoothness}(S) = \frac{1}{N-1} \sum_{i=1}^{N-1} |S(t_{i+1}) - S(t_i)|$$

A lower value indicates more gradual changes (smoother transitions), while a higher value reflects more abrupt predictability shifts.

5.2 Comparative Evaluation

We computed model preferences using GPT2⁷, comparing the “original” and “revision” versions based on surprisal metrics described in Section 5.1. Among these, only the differences in *average token*

⁷Specifically, the GPT2-large model from the transformers library (Wolf et al., 2020).

Gold label	Equal	Orig	Revi
Either_way	8	14	14
Original_better	1	6	3
Original_should_stay	2	6	0
Revision_better	6	9	20
Revision_necessary	2	0	9

Table 3: Prediction distribution from GPT2-large. **Orig** indicates preference for the original connective; **Revi**, for the revised connective; and **Equal** indicates both versions received identical surprisal values.

surprisal were statistically significant across human annotation categories ($p < 0.001$), suggesting a meaningful correlation between average surprisal changes and human preferences. In contrast, differences in *variance* and *smoothness* did not reach statistical significance.

For a comparison with a more recent model, we conducted prompting-based experiments using GPT-4.1-mini. In each prompt, the connective was replaced with “<.>” and the model was explicitly asked to suggest an appropriate discourse connective for that position (see Appendix C for examples). Each item was evaluated across five independent runs. If the original connective was predicted more frequently than the revised one, we interpreted this as a preference for the original; if the revised one was more frequent, it was considered preferred. Equal prediction rates (or no prediction) were treated as indicating no clear preference.

Results Table 3 and Table 4 summarize the predictions of GPT-2. and GPT-4.1, respectively, on the 100 manually annotated instances.⁸ Despite some variation across individual categories, the overall distributional patterns between both models are largely consistent. A *chi-square test* performed across categories revealed no statistically significant differences between two models’ predictions.

We also evaluated GPT-4.1-mini as a classifier by prompting it with the same 5-way classification instructions provided to our human annotators. Each of the 100 instances was labeled through 5 runs, and the majority prediction was compared to the human majority label. The model’s predictions matched the human majority in 43% of the cases, with a moderate association between the two label

⁸Note that zero-shot and few-shot prompting did not yield statistically significant differences in GPT-4.1’s output.

Gold label	Equal	Orig	Revi
Either_way	11	9	16
Original_better	3	5	2
Original_should_stay	2	5	1
Revision_better	10	3	22
Revision_necessary	2	0	9

Table 4: Prediction distribution from GPT-4.1 on the manually annotated dataset.

sets (Cramér’s $V = 0.306$, p -value = 0.002).

We further evaluated open-source LLMs, specifically Mistral-7B and LLaMA-3.1-8B. Both models had difficulty adhering to the prompt instructions unless supplied with a short list of example connectives. In consequence, they heavily favored items from the provided list, resulting in reduced lexical diversity. To address this limitation, we also experiment with a few-shot prompting strategy rather than explicitly listing options. Under this configuration, Mistral-7B exhibited improved performance, producing valid connectives more consistently and with greater lexical variety. However, no statistically significant correlation with human annotations was found ($p > 0.05$; see Table 5 in Appendix D for the predictions of Mistral-7B).

6 Conclusion

We present a dataset of text revisions involving the deletion or replacement of discourse connectives at the beginning of sentences in the WikiHow text revisions.⁹ From this dataset, we manually annotated 100 instances of connective replacements. Only 11% of these edits were judged necessary to convey the correct discourse meaning, whereas in 8% cases, the original connective was favored; the remaining edits appeared to be somewhat optional.

Using GPT-2, we computed information-theoretic metrics (mean surprisal, variance, and smoothness) for these annotations. Of these, only mean surprisal significantly correlated with human judgments (RQ1). A prompt-based evaluation with GPT-4.1-mini showed similar preferences, especially in edge cases of 5-way classification, where a revision was necessary or the original connective should be retained (RQ2). These results suggest that information-theoretic metrics and prompt-based methods capture some patterns in human

⁹The dataset is publicly available at: <https://github.com/berfingit/connective-deletion-replacement>.

decisions on discourse connectives, though their coverage remains limited.

Acknowledgements

The research presented in this paper was funded by the DFG Emmy Noether program (RO 4848/2-1).

Limitations

Our human annotation analysis is limited to a small subset of the connective replacement data; expanding annotations to include more examples and other edit types (e.g., insertions and deletions) would strengthen the generalizability of our findings. Additionally, we do not explicitly ground our analysis in a theoretical framework such as the Uniform Information Density (UID) hypothesis, leaving open questions about the broader cognitive or linguistic implications of our results.

For computing surprisal-based metrics, we rely on GPT-2, a relatively outdated language model. This choice is motivated by the lack of token-level log-probability access in widely used API-based models like GPT-3.5 and GPT-4. While more recent open-source models such as Mistral-7B and LLaMA-3.1 provide access to logits (via the transformers library) and can be used for surprisal computation, integrating them was beyond the scope of this pilot study. We leave surprisal-based experiments with more recent models to future work.

References

- Berfin Aktas and Michael Roth. 2025. [Clarifying under-specified discourse relations in instructional texts](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12237–12256, Vienna, Austria. Association for Computational Linguistics.
- Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. [wikiHowToImprove: A resource and analyses on edits in instructional texts](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.
- Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. [A cross-linguistic pressure for Uniform Information Density in word order](#). *Transactions of the Association for Computational Linguistics*, 11:1048–1065.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Michael Roth, Talita Anthonio, and Anna Sauer. 2022. SemEval-2022 task 7: Identifying plausible clarifications of implicit and underspecified phrases in instructional texts. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1039–1049, Seattle, United States. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, London, UK. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pi  ric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Examples with Additional Context

Source: *Choreograph a Great Solo*

Section: Steps

Original:

2. Pick a music choice that agrees with your style choice. [So], if your dance is fast, pick a fast song.

Revised:

2. Pick a music choice that agrees with your style choice. [For example], if your dance is fast, pick a fast song.

Figure 1: Example of connective replacement

Article: *Make a Summer Dress out of a Bedsheet*

Section: Steps

Original:

9. Cut the four shirt pieces out of the sheet material. [Then] place those pieces on the lining material, face side up, and pin them into place before cutting them out.

Revised:

9. Cut the four shirt pieces out of the sheet material. Place those pieces on the lining material, face side up, and pin them into place before cutting them out.

Figure 2: Example of connective deletion in revision

Source: *Chat Using Facebook Messenger App on iOS*

Section: Logging into Messenger

Original:

1. Sign in. If you’re already using the Facebook app and have it installed and running on your mobile device, just tap on the blue “Continue as...” button upon launch. [Otherwise], input your email and password for your Facebook account.

Revised:

1. Sign in. If you’re already using the Facebook app and have it installed and running on your mobile device, just tap on the blue “Continue as...” button upon launch. Input your email and password for your Facebook account.

Figure 3: Example of misleading connective deletion.

Source: *Use Every Nikon Digital SLR*

Original:

[While] there are enough similarities between all Nikon digital SLRs.

These categorisations are used here for convenience’s sake and have nothing to do with image quality.

Revised:

There are enough similarities between all Nikon digital SLRs.

These categorisations are used here for convenience’s sake and have nothing to do with image quality.

Figure 4: Example of connective deletion causing ungrammatical output.

B Information-Theoretic Metrics on the Full Dataset

We analyze the whole dataset using the metrics described in §5.1 and present the results in this section.

Average token surprisal: Connective insertions and replacements lead to a statistically significant increase in average surprisal, while deletions cause a significant decrease (Figure 5). An increase in average surprisal suggests that the resulting text became less predictable overall, whereas a decrease reflects a shift toward greater predictability.

Interestingly, the reverse of insertions (i.e., removing connectives that had been inserted by humans, labeled as `insertion_rev`) also leads to a decrease in average surprisal, mirroring the effect of deletions. The difference between deletion and `insertion_rev` is not statistically significant, suggesting that these two operations have symmetrical effects on average surprisal despite differing in context.

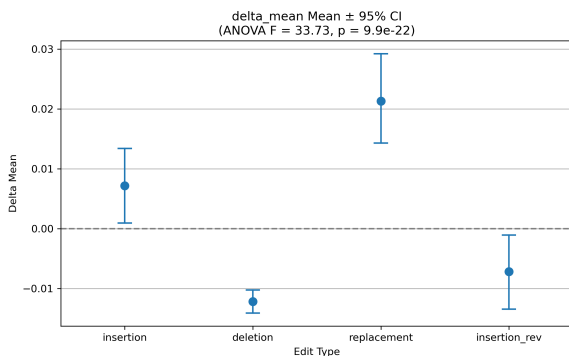


Figure 5: Mean change in surprisal (Δ Mean) across edit types. Error bars represent 95% confidence intervals. ANOVA indicates a significant effect of edit type on mean surprisal ($F = 59.76$, $p = 1.4e-26$).

Variance: As shown in Figure 6, both deletion and replacement significantly reduce surprisal variance, making the text more uniformly predictable. In contrast, insertions and therefore their reverse functions `insertion_rev` do not significantly alter variance.

This asymmetry between deletions and `insertion_rev`, despite their functional similarity, suggests that while the mean surprisal remains stable, the local variance depends more on the editing context.

Smoothness: Figure 7 shows that insertions, deletions, and replacements each lead to small

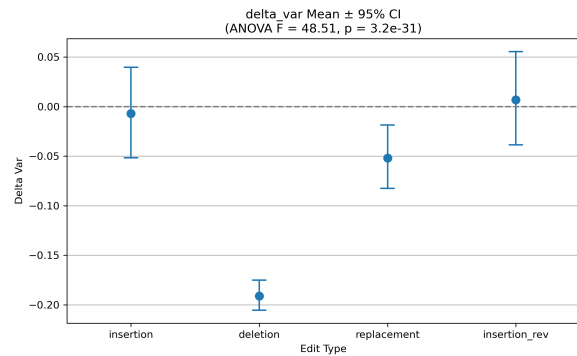


Figure 6: Mean change in variance (Δ Var) across edit types. Error bars represent 95% confidence intervals. ANOVA indicates a significant effect of edit type on surprisal variance ($F = 55.11$, $p = 1.4e-24$).

but statistically significant decreases in surprisal smoothness (i.e., negative values indicate that revisions are less smooth). However, the differences across edit types are not statistically significant, suggesting that all three introduce increases in local unpredictability.

There is a clear contrast between deletions and `insertion_rev` in terms of smoothness. Removing human-inserted connectives appears to make the surprisal values change more gradually, yielding a smoother predictability pattern.

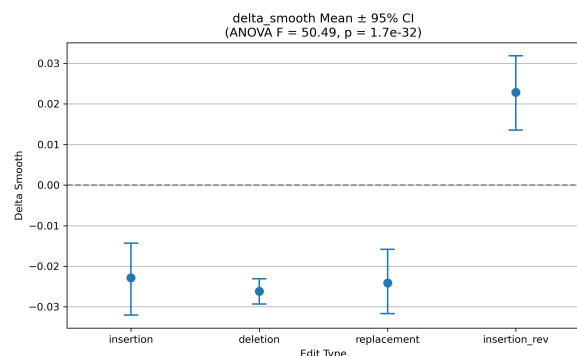


Figure 7: Mean change in smoothness (Δ Smooth) across edit types. Error bars represent 95% confidence intervals. ANOVA does not indicate a significant effect of edit type on smoothness ($F = 0.42$, $p = 0.66$).

Discussion These results show that while deletion and `insertion_rev` have similar effects on average surprisal, they differ notably in their impact on surprisal variance and smoothness. This suggests that although a connective’s overall informativeness may remain stable, its influence on local predictability (i.e., whether it smooths or disrupts the flow) depends on context.

Overall, the surprisal-based metrics offer a nuanced view of how connective edits affect a language model’s expectations (as measured by GPT-2). Interestingly, connective insertions and replacements increase average surprisal and reduce smoothness, indicating that these edits do not improve predictability as one might expect with explicit connectives. Since some of these findings run counter to intuition and we do not yet provide a deeper analysis, we consider them inconclusive.

C Prompts

Prompt Example:

TASK: Insert an appropriate discourse connective into the gap marked by <...> in the following text.

Text:

Steps 1. To easily interpret your quotation and make a plausible argument and analysis, ask yourself questions. For example, why was this said? What possible situation made this quotation significant? 2. For your main points, ask yourself questions based on your interpretation. Always remember why, what, and how. 3. <...> you have your main points, come up with solid examples. Books, movies, politics, current events, music, art, culture, history, and any other category will work. Just remember to not use all of one category! Mix them up for a more solid analysis.

Answer:

D Mistral-7B’s Connective Predictions on Annotated Data

Human Majority	Equal	Orig	Revi
Either_way	31	2	3
Original_better	10	0	0
Original_should_stay	5	2	1
Revision_better	27	6	2
Revision_necessary	10	0	1

Table 5: Mistral-7B’s match breakdown by human majority class

E Frequency Distribution of Connective Deletions and Replacements

Connective	Count	Normalized (%)
Then	882	20.5%
For example	669	15.5%
However	558	13.0%
Also	425	9.9%
But	237	5.5%
Or	185	4.3%
And	177	4.1%
So	174	4.0%
For instance	139	3.2%
If	121	2.8%
Finally	118	2.7%
Instead	106	2.5%
In addition	70	1.6%
Otherwise	42	1.0%
In fact	35	0.8%

Table 6: Most frequent connective insertions with raw counts and normalized percentages.

Connective	Count	Percentage (%)
Then	4287	31.5%
Also	1690	12.4%
And	1600	11.8%
So	971	7.1%
But	960	7.1%
However	917	6.7%
Finally	851	6.3%
Or	414	3.0%
For example	374	2.8%
Instead	147	1.1%
Because	120	0.9%
In order	114	0.8%
Therefore	99	0.7%
If	97	0.7%
For instance	88	0.6%

Table 7: Most frequent connective deletions with raw counts and normalized percentages.

Replacement	Count	Percentage (%)
But → However	314	17.1
When → If	106	5.8
If → When	80	4.3
And → Also	65	3.5
Because → Since	39	2.1
However → But	35	1.9
If → Even if	31	1.7
Also → In addition	29	1.6
When → Once	23	1.2
And → In addition	22	1.2
Once → When	18	1.0
Even if → If	18	1.0
While → Although	18	1.0
Then → Finally	18	1.0
So → Therefore	18	1.0

Table 8: Most frequent connective replacements with raw counts and normalized percentages.

Author Index

- Adhikari, Surabhi, 120
Ahmad, Mahmoud, 54
Aktas, Berfin, 228
Alshomary, Milad, 166
Alta, Marijn, 27
Anagnostopoulou, Aliko, 166
Apishev, Murat, 1
Averkiev, Sergei, 1
- Bhattacharya, Debasmita, 64
Bloem, Jelke, 27
Bulatov, Aydar, 1
- Carenini, Giuseppe, 148
Chistova, Elena, 197
Churin, Igor, 1
- Ding, Siying, 64
- Feldhus, Nils, 166
Fenogenova, Alena, 1
Fuchs, Anna, 130
- Guan, Wenwen, 27
- Haensch, Anna-Carolina, 130
Hardmeier, Christian, 157
Hirschberg, Julia, 64
Hsu, Yi-Sheng, 166
- Jeon, Sungho, 42
- Kakudi, Habeebah, 54
Kuratov, Yuri, 1
- Latusek, Anna, 182
Lewandowska, Martyna, 182
Li, Chuyuan, 148
Liu, Guifu, 81
Liu, Wei, 96
- Ma, Bolei, 130
Mehri, Shuhaib, 148
Murzaku, John, 107
- Naseem, Usman, 120
Noltenius, Elisa, 130
- Ogrodniczuk, Maciej, 182
Okrański, Adam, 182
- Rambow, Owen, 107
Rauniyar, Kritesh, 120
Razzak, Imran, 120
Rohde, Hannah, 81
Rosalska, Paulina, 182
Roth, Michael, 228
Ruths, Derek, 209
- Saputa, Karolina, 182
Sheikh, Amna, 157
Shevelev, Denis, 1
Sonntag, Daniel, 166
Steel, Benjamin, 209
Strube, Michael, 42, 96
- Thapa, Surendrabikram, 120
Tikhonova, Maria, 1
Tomaszewska, Aleksandra, 182
- Veeramani, Hariram, 120
- Wachsmuth, Henning, 166
Wan, Stephen, 96
Webber, Bonnie, 81
Weinzierl, Caroline, 130
Wróblewska, Alina, 182
Wu, Jingni, 14
- Yi, Anxin, 64
- Zeldes, Amir, 14
Ziembicki, Daniel, 182
- Śliwicka, Anna, 182
- Żuk, Bartosz, 182
Żurowski, Sebastian, 182