

WASSA 2026

**15th Workshop on Computational Approaches to
Subjectivity, Sentiment and Social Media Analysis**

Proceedings of the Workshop

March 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

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

ISBN 979-8-89176-378-4

Introduction

This edition marks the 15th iteration of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), a crystal jubilee!

We can summarize the workshop as focusing on topics at the intersection of emotion, subjectivity, persuasion, and trust. Crystal is fragile, and many topics at this year's WASSA can be situated within a social or political context, reflecting the brittle state our world is in. Regarding methodologies, Large Language Models (LLMs) prevail: as tools for simplification or for simulating social media users, as synthetic data generators, and as an object of study. This year, WASSA also featured a Special Track on Multilinguality and Social Bridges between High- and Lesser-Resourced Languages and Communities. In the proceedings, one paper takes a truly multilingual approach to quantify social sentiment in hostels. Only three papers do not mention English data at all, but instead focus on subjective tasks in Farsi, emotion preservation when translating from Spanish to Basque, and assessing the capability of LLMs to predict the emotional ratings of Russian words. Other languages that are being studied alongside English are German, Danish, and Luxembourgish.

For the main workshop, we received 32 papers and 9 ARR commitments. From the direct submissions, we accepted 17 (53%), and from ARR, another 6 (66%), bringing the total to 23 accepted papers. The high quality of much of the work made the decision process difficult, as this year, WASSA is held as a half-day workshop, reducing the room we had for accepted papers.

For our jubilee, we decided to award both a Best Paper and a Best Theme Paper Award. Starting with the reviewer scores, the organizers of the workshop anonymously proposed candidates and then voted on them. The winner of the Best Paper Award of this year's WASSA is *The Impact of Highlighting Subjective Language on Perceived News Trustworthiness* by Mohammad Shokri, Vivek Sharma, Emily Klapper, Shweta Jain, Elena Filatova, and Sarah Ita Levitan. The winner of the Best Theme Paper Award is *Crowd-Based Evaluation of Emotion Intensity Preservation in Spanish–Basque Tweet Machine Translation* by Nora Aranberri. Congratulations!

On top of the main workshop, we are happy to have José Camacho-Collados from the Cardiff NLP group as our invited speaker. In his keynote he will discuss social media analysis in the language model era, and this from an interdisciplinary perspective.

We would like to thank the EACL Organizers and Workshop chairs for their help and support during the preparation. We also thank the OpenReview support team for their technical support. Finally, we thank the program committee for the time and effort they spent on reviewing, especially to those who were willing to perform emergency reviews. The quality of the program is a reflection of your hard work.

Finally, we would like to dedicate WASSA 2026 to Shabnam Tafreshi. Shab was one of the driving forces behind many previous iterations of WASSA and the shared tasks on emotion and empathy prediction.

Jeremy Barnes, Valentin Barriere, Orphée De Clercq, Roman Klinger, Debora Nozza, Célia Nouri & Pranaydeep Singh

WASSA 2026 Chairs

Acknowledgements

This work is partially supported by the Basque Government (IXA excellence research group IT1570-22) and the European Union under Horizon Europe (Project LUMINOUS, grant number 101135724). Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Valentin's work is partially funded by ANID grant Fondecyt de Iniciacion 11251024. Pranaydeep's work is supported by a BOF post-doctoral research fellowship (BOF.PDO.2025.0011.01) from Ghent University.

Organizing Committee

Program Chairs

Jeremy Barnes, University of the Basque Country UPV/EHU, Spain

Valentin Barriere, Universidad de Chile, Chile

Orphée De Clercq, Ghent University, Belgium

Roman Klinger, Otto-Friedrich Universität Bamberg, Germany

Debora Nozza, Bocconi University, Italy

Célia Nouri, INRIA and Sciences Po, France

Pranaydeep Singh, Ghent University, Belgium

Program Committee

Reviewers

Gavin Abercrombie, Wazir Ali, Liesbeth Allein

Murali Balusu, Sivaji Bandyopadhyay, Delphine Battistelli

Jose Camacho-Collados, Alessandra Teresa Cignarella

Luna De Bruyne, Loic De Langhe, Quanqi Du

Natalia Evgrafova

Neele Falk, Pascale Feldkamp, Elena Filatova

Gaël Guibon

Thomas Haider, Junko Hayashi, Jet Hoek, Andreas Hotho

Imezadelajara, Sekh Mainul Islam, Kazuhiro Ito, Gabriel Iturra-Bocaz

Aditya Joshi

Akbar Karimi, Yannis Karmim, Elma Kerz, Khalid Al Khatib, Marek Kubis, Kemal Kurniawan

Sofie Labat, Oier Lopez De Lacalle, Els Lefever, Natalia V Loukachevitch

Cinthia Sanchez Macias, Aaron Maladry, Eugenio Martínez-Cámara, Petter Mæhlum

Emerson Cabrera Paraiso

Tamara Quiroga

Ananya Raval, David Rossouw, Irene Russo, Egil Rønningstad

Hamidreza Saffari, Iñaki San Vicente, Elena Savinova, Mohammadamin Shafiei, Vivek Sharma, Mohammad Shokri, Phillip Smith, Efstathios Stamatatos, Benjamin Steel

Samia Touileb, Enrica Troiano

L. Alfonso Ureña

Cynthia Van Hee, Eva Maria Vecchi, Aswathy Velutharambath, David Vilares

Shoko Wakamiya, Jin Wang, Jingjing Wang, Michael Wiegand, Tak-Lam Wong

Zehong Yan

Lei Zhang

Keynote Talk

Social Media Analysis in the Language Model Era: An Interdisciplinary Perspective

Jose Camacho-Collados

Cardiff University, UK

29 March, 2026 – Time: 14:30 – 15:30 – Room: SALLE Le Chellah

Abstract: Understanding what is going on in social media is not a simple task, even after the significant progress in NLP in recent years. Given the specific nature of social media and the large amount of content generated, efficient and specialised solutions are often necessary.

At the same time, language models, and in particular large generative language models (LLMs) are currently ubiquitous across the NLP research landscape and beyond. However, it isn't clear how to effectively leverage them to process large amounts of unstructured data, such as the type usually found on social media, especially for non-expert users. There are also additional challenges when processing large amounts of text, such as ensuring transparency in the process and dealing with long context.

In this talk, I will explain my journey when it comes to social media analysis, from proposing efficient solutions for tasks such as sentiment analysis, hate speech detection or emotion detection, to sharing tips and lessons learned from interdisciplinary collaborations. I will try to revisit the importance of high-quality data and reliable evaluation benchmarks in the current LLM landscape, and some recent results suggesting promising areas for future research.

Bio: Jose Camacho-Collados is a UKRI Future Leaders Fellow and Professor at the School of Computer Science of Cardiff University, where he co-founded the Cardiff Natural Language Processing group (Cardiff NLP). Before joining Cardiff University, he completed his PhD in Sapienza University of Rome and was a Google AI PhD Fellow.

Jose has worked in multiple NLP areas with a particular focus on semantics, multilinguality and computational social science with an interdisciplinary perspective. In this area, he has been developing specialised and efficient NLP models for social media applications, such as TweetNLP and related efforts. His work has received several recognitions, including awards at top NLP conferences, or the 2023 AIJ Prominent Paper Award. He is also the co-author of the “Embeddings in Natural Language Processing” book.

Table of Contents

<i>Council of LLMs: Evaluating Capability of Large Language Models to Annotate Propaganda</i> Vivek Sharma, Shweta Jain, Mohammad Shokri, Sarah Ita Levitan and Elena Filatova	1
<i>Emoji Reactions on Telegram: Unreliable Indicators of Emotional Resonance</i> Serena Tardelli, Lorenzo Alvisi, Lorenzo Cima, Stefano Cresci and Maurizio Tesconi	13
<i>Quantifying Social Sentiment in Hostels Using A Domain-Specific Transformer Pipeline</i> Ian W. McMurry	24
<i>Predicting Convincingness in Political Speech: How Emotional Tone Shapes Persuasive Strength</i> Bhuvanesh Verma, Mounika Marreddy and Alexander Mehler	37
<i>Measuring LLMs’ Sensitivity to Paraphrased Opinion Prompts</i> Bushra Alhetelah and Irfan Ahmad	52
<i>The Impact of Highlighting Subjective Language on Perceived News Trustworthiness</i> Mohammad Shokri, Vivek Sharma, Emily Klapper, Shweta Jain, Elena Filatova and Sarah Ita Levitan	60
<i>Appraisal Trajectories in Narratives Reveal Distinct Patterns of Emotion Evocation</i> Johannes Schäfer, Janne Wagner and Roman Klinger	73
<i>Exploring Subjective Tasks in Farsi: A Survey Analysis and Evaluation of Language Model</i> Donya Rooein, Flor Miriam Plaza-del-Arco, Debora Nozza and Dirk Hovy	83
<i>Emotional Lexicons: How Large Language Models Predict Emotional Ratings of Russian Words</i> Polina V. Iaroshenko and Natalia V Loukachevitch	96
<i>Emotion-aware text simplification of user generated content using LLMs</i> Anastasiia Bezobrazova, Daria Sokova and Constantin Orasan	107
<i>Crowd-Based Evaluation of Emotion Intensity Preservation in Spanish–Basque Tweet Machine Translation</i> Nora Aranberri	123
<i>A Position Paper on Toxic Reasoning: Grounding Categories of Toxic Language in Implications and Attitudes</i> Stefan F. Schouten, Ilia Markov and Piek Vossen	134
<i>Is Sentiment Banana-Shaped? Exploring the Geometry and Portability of Sentiment Concept Vectors</i> Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer Nielbo and Kenneth Enevoldsen	146
<i>Disentangling Emotion Understanding and Generation in Large Language Models</i> Sadegh Jafari, Els Lefever and Veronique Hoste	161
<i>News Credibility Assessment by LLMs and Humans: Implications for Political Bias</i> Pia Wenzel Neves, Charlott Jakob and Vera Schmitt	172
<i>Towards Simulating Social Media Users with LLMs: Evaluating the Operational Validity of Conditioned Comment Prediction</i> Nils Schwager, Simon Münker, Alistair Plum and Achim Rettinger	208

<i>Label-Consistent Data Generation for Aspect-Based Sentiment Analysis Using LLM Agents</i> Mohammad Hossein Akbari Monfared, Lucie Flek and Akbar Karimi	222
<i>Antisocial Behavior Prediction: A Survey and Practical Guide</i> Anaïs Ollagnier	235
<i>Real-Time Mitigation of Negative Emotion in Customer Care Calls</i> Surupendu Gangopadhyay and Mahnoosh Mehrabani	252
<i>Says Who? Argument Convincingness and Reader Stance Are Correlated with Perceived Author Personality</i> Sabine Weber, Lynn Greschner and Roman Klinger	265
<i>A Transformer and Prototype-based Interpretable Model for Contextual Sarcasm Detection</i> Ximing Wen and Rezvaneh Rezapour	278
<i>Multimodal Claim Extraction for Fact-Checking</i> Joycelyn Teo, Rui Cao, Zhenyun Deng, Zifeng Ding, Michael Sejr Schlichtkrull and Andreas Vlachos	289
<i>A Multi-Aspect Evaluation Framework for Synthetic Data: Case Study on Irony and Sarcasm</i> Laura Majer, Ana Barić, Florijan Sandalj, Ivan Unković, Bojan Puvača and Jan Šnajder	305

Council of LLMs: Evaluating Capability of Large Language Models to Annotate Propaganda

Vivek Sharma¹, Shweta Jain^{1,2}, Mohammad Mahdi Shokri¹,
Sarah Ita Levitan^{3,1}, Elena Filatova^{4,1}

¹The Graduate Center, CUNY, ²John Jay College of Criminal Justice, CUNY,
³Hunter College, CUNY, ⁴New York City College of Technology, CUNY,

Abstract

Data annotation is essential for supervised natural language processing tasks but remains labor-intensive and expensive. Large language models (LLMs) have emerged as promising alternatives, capable of generating high-quality annotations either autonomously or in collaboration with human annotators. However their use in autonomous annotations is often questioned for their ethical take on subjective matters. This study investigates the effectiveness of LLMs in an autonomous, and hybrid annotation setups in propaganda detection. We evaluate GPT and open-source models on two datasets from different domains, namely, Propaganda Techniques Corpus (PTC) for news articles and the Journalist Media Bias on X (JMBX) for social media. Our results show that LLMs, in general, exhibit high recall but lower precision in detecting propaganda, often over-predicting persuasive content. Multi-annotator setups did not outperform the best models in single-annotator setting although it helped reasoning models boost their performance. Hybrid annotation, combining LLMs and human input, achieved the highest overall accuracy than LLM-only settings. We further analyze misclassifications and found that LLM have higher sensitivity towards certain propaganda techniques like loaded language, name calling, and doubt. Finally, using error typology analysis, we explore the reasoning provided on misclassifications by the LLM. Our result shows that although some studies report LLM outperforming manual annotations and it could prove useful in hybrid annotation, its incorporation in the human annotation pipeline must be implemented with caution.

1 Introduction

Data annotation is a foundational task in natural language processing (NLP), providing the ground truth needed to train and evaluate supervised models. Traditionally, this task is performed by subject matter experts (SMEs) who follow detailed

guidelines to ensure consistency and accuracy. Due to cost and time consideration the annotation process is often scaled using crowdsourcing platforms like Amazon Mechanical Turk and Prolific [Fort et al.](#), [Sabou et al.](#). These labels are typically filtered and validated to produce high-quality “gold” annotations to train NLP models.

Despite its importance, manual annotation remains slow, expensive, and susceptible to annotator fatigue, especially with large datasets. This can lead to inconsistencies that degrade model performance. In response, recent work has explored using large language models (LLMs), such as GPT-4 and GPT-4o, for annotation. These models have shown promise in generating high-quality labels in both hybrid settings, where human annotators review LLM-generated suggestions and autonomous settings, where models label data without human intervention.

Hybrid annotation seeks to reduce human workload while preserving annotation quality. In contrast, fully automated annotation offers scalability but raises concerns about bias, hallucination, and the propagation of model-generated errors. While LLMs continue to improve in performance and alignment, questions remain about their reliability and accountability in annotation workflows.

Though LLMs have been evaluated for tasks like classification, summarization, and question answering, their use as a multi-annotator in subjective tasks like propaganda detection remains underexplored. Moreover, there is a lack of in-depth exploration in their reasoning which can help us understand the black-box nature of their analysis. This paper addresses the gap in the context of propaganda detection, a domain where language is often subtle and context-dependent. It is easy for annotators’ biases to impact labels in the data. In contrast, LLMs can be engineered to act neutrally and adhere to guidelines more strictly. We study the difference in annotation accuracy compared to

human-generated labels by measuring LLMs’ capability in single-annotator, multi-annotator, and hybrid settings in an exhaustive comparison of the latest GPT [Achiam et al.](#) and open-source models. We evaluate LLMs using two datasets: the Propaganda Techniques Corpus (PTC) ([Martino et al., 2020](#)) for news articles and Journalist Media Bias on X (JMBX) ([Sharma et al., 2025a](#)) for social media content.

Through this study, we investigate the potential of LLMs to perform annotation without human supervision and evaluate the performance of multi-annotator collaboration among LLMs, both quantitatively and qualitatively. Specifically, we explore the following research questions:

RQ1: Does hybrid annotation (human-LLM) perform better in detecting propaganda than automated LLM annotations? To what extent does hybrid annotation (human-LLM) outperform fully automated LLM annotation in detecting propaganda?

RQ2: Do human-LLM disagreements cluster around specific propaganda techniques, and if so, which ones?

RQ3: Under what conditions, if any, does consolidation in a multi-annotator setting match or exceed single-annotator accuracy across text domains (news vs. social media)?

Our results underscore shortcomings of multi-annotator models and how consolidation performance in adjudication depends on individual annotators. Qualitatively, we find that while proprietary and open-source LLMs often detect valid propaganda patterns, they overgeneralize—especially in informal social media contexts. Their behavior is shaped by linguistic features as well as formatting and tone, which can cause systematic mislabeling without strong contextual grounding.

2 Related Work

LLM perform well on domain-specific annotation tasks, including software engineering ([Ahmed et al., 2025](#)). Beyond technical domains, [Törnberg](#) shows GPT models can surpass expert annotators and supervised classifiers in labeling political social media. These results suggest LLM annotations can closely match human work. However, several scholars warn of risks in social science annotation workflows. LLMs are often viewed as black boxes ([Kristensen-McLachlan et al., 2023](#); [Bender](#)

[et al., 2021](#)), making their output difficult to interpret and reproduce ([Ollion et al., 2024](#)). Moreover, their training data can embed and amplify existing biases, such as gender stereotypes ([Lucy and Bamman, 2021](#)). By contrast, human annotation follows guidelines and quality controls to mitigate such biases. Consequently, calls have been made for standards for LLM-based annotations to ensure reliability and transparency ([Törnberg, 2024a,b](#)).

Technical limitations constrain the efficacy of LLMs by being prone to hallucination ([Lin and Zhang, 2025](#)), over-confidence ([Xiong et al., 2023](#)), and when annotations are wrong, often produce explanations lacking coherence and sufficiency ([Wang et al., 2024](#)). Prompt design can improve annotation quality ([Reiss, 2023](#)), though others find only marginal gains ([Rytting et al., 2023](#); [Murugadoss et al., 2025](#)). Societal risks persist: removing human annotators can erase interpretive insight and reinforce systemic bias ([Das et al., 2024](#); [Abid et al., 2021](#); [Wang et al., 2023](#)), leading to unfair outcomes and broader disparities ([Dillion et al., 2023](#)).

To balance risks and benefits, hybrid workflows let LLMs and humans collaborate across annotation stages. LLMs may pre-process data, propose labels, or validate human work. [Lin and Zhang \(2025\)](#) position the LLM either as the main annotator for large-scale tasks with humans verifying output or as a secondary assistant when dataset sizes are smaller. Several studies have experimented with different hybrid configurations. [Heseltine and Clemm von Hohenberg \(2024\)](#) use a three-phase pipeline where the LLM labels twice, discrepancies are reconciled, then a final model is trained. [Hamilton et al. \(2024\)](#) introduce RhetAnn, a web tool that aids annotators with LLM-generated explanations for rhetorical techniques. [Wang et al. \(2024\)](#) adds LLM explanations as supplementary guidance, accelerating work while keeping humans in the loop. [Hasanain et al. \(2023\)](#) shows LLMs can serve as consolidators, roles usually for experts, achieving strong propaganda span detection with detailed rationales. [Kim et al. \(2024\)](#) showed hybrid frameworks improves quality and address scalability and fatigue. [Rouzegar and Makrehchi \(2024\)](#) integrate LLM uncertainty into active learning, switching between human and machine annotators by confidence thresholds.

As hybrids mature, fully automated annotation is also examined. [Hamilton et al. \(2024\)](#) report GPT-4 matching human quality at one-tenth the cost with

minimal manual labels. [Törnberg \(2023\)](#) finds that ChatGPT outperforms crowdsourced annotators on tweets. Not all agree: [Gu et al. \(2025\)](#) show automated LLMs beat TF-IDF methods yet remain less reliable than experts. [Golazizian et al. \(2024\)](#) combine RoBERTa-Large and LLama-3, gaining 4% F1 on hate speech with 50% of the budget and 2% on moral sentiment with 25% of costs. [Subramaniam et al. \(2025\)](#) proposed a self-refining multi-annotator system that preserves diverse reasoning.

These studies show multi-annotator LLMs and hybrid annotations can handle varied tasks with mixed performance. However, none directly compare single-annotator and multi-annotator models on subjective tasks like propaganda detection. Such a comparison helps assess performance differences and reveal how LLMs behave in multi-annotator settings. Grasping classification nuances is necessary to interpret LLM reasoning, see where it diverges from human annotation, and analyze that process. This, in turn, helps prompt engineers refine prompts and adjust model parameters for subjective tasks. Our aim is not peak accuracy, but to explore LLM performance across scenarios and domains.

3 Experiment and Result

To evaluate LLM annotation in single-annotator, multi-annotator, and hybrid setups, we run experiments on two datasets: the Propaganda Techniques Corpus (PTC) [Martino et al. \(2020\)](#) and the Journalist Media Bias on X (JMBX) corpus [Sharma et al. \(2025b\)](#). PTC provides sentence-level binary labels for news articles, while JMBX contains annotated tweets from journalists’ accounts associated with biased outlets on X (formerly Twitter). From each dataset, we sample 200 sentences with equal numbers of propaganda and non-propaganda instances. The propaganda subset follows the natural distribution of techniques in the original datasets. All models, both GPT and open-source LLMs, are evaluated on the same sentences for consistency and the reported performance is the average of three runs on each setting.

In standalone setting, each model independently performs binary classification (propaganda vs. non-propaganda) at the sentence level. Using a prompt that includes technique definitions by [Martino et al. \(2020\)](#) (see Appendix B), the model is asked to identify any technique(s) present. If a technique is

returned, the sentence is labeled as propaganda; if none is returned, it is labeled as non-propaganda. In a multi-annotator setting, each model first operates as an independent annotator. In this stage, the model labels every sentence with one or more propaganda techniques and provides a brief justification for each decision. These initial annotations are generated in isolation, so no model has access to the labels or rationales produced by any other model.

In the second stage, each model is then anonymously presented with the full set of sentence-level labels and justifications produced in the first stage by all models, including its own. Using a consolidation prompt that includes the definitions of all propaganda techniques, the model is instructed to review these candidate labels and select a single final label together with its accompanying justification. Importantly, this consolidation step is framed as a re-annotation and not as an aggregation procedure

We replicate the same prompts and sentence sets across proprietary GPT models and open-source models on both PTC and JMBX to compare performance across models and datasets.

The performance of LLMs in various configuration is reported using standard metrics ([Sokolova and Lapalme, 2009](#)): F1-score, precision and recall generally used in the studies. This analysis is followed by qualitative analysis understanding understanding the reasoning on misclassifications.

3.1 Automated annotations by LLM

3.1.1 GPT models

First, we evaluate the annotation performance of three GPT models: GPT-4o (gpt-4o-2024-08-06), GPT-4.1 (gpt-4.1-2025-04-14), and GPT-o3 (o3-2025-04-16). These models are selected for their diverse capabilities. GPT-4.1 for its proficiency in complex reasoning, GPT-4o for its speed and balanced performance, and GPT-o3 for its specialized reasoning strengths simulating varied annotator expertise.¹

Table 1 and Table 2 report performance on PTC and JMBX datasets, respectively.

¹<https://platform.openai.com/docs/models>

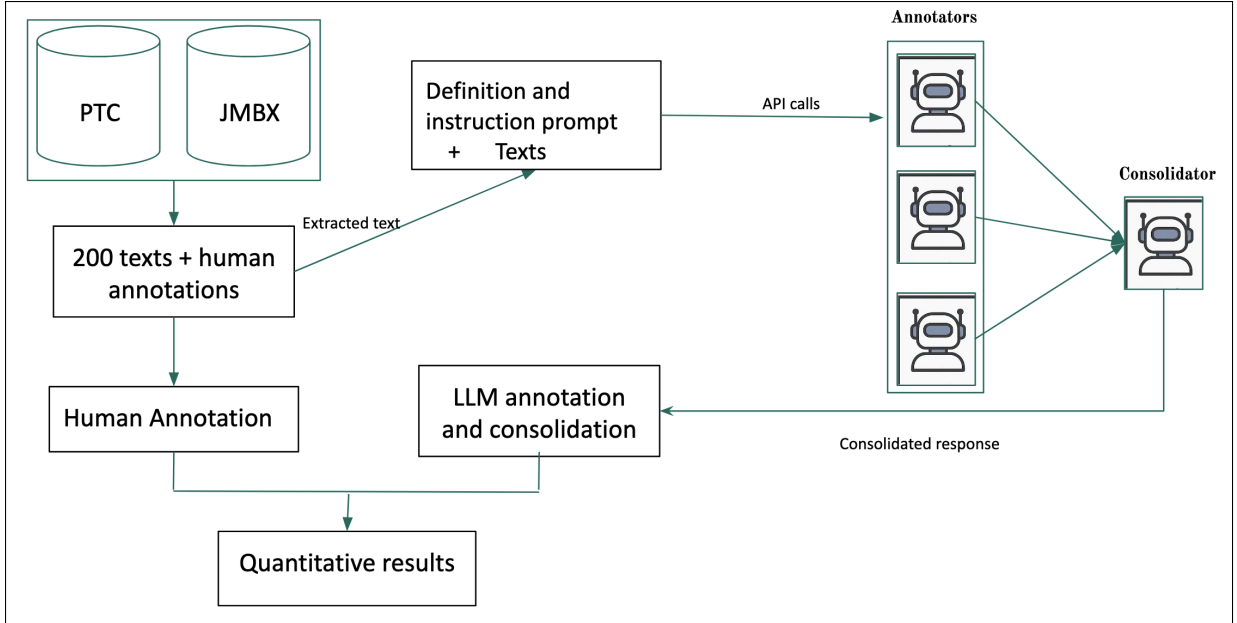


Figure 1: Experiment to assess LLMs performance in annotating propaganda in an automated setting

Model	P/R/F	Consolidation P/R/F
GPT 4.1	0.77/0.74/0.74	0.78/0.71/0.70
GPT 4o	0.79/0.74/0.73	0.78/0.70/0.68
GPT o3	0.78/0.70/0.68	0.78/0.71/0.69

Table 1: Performance of GPT models in annotating PTC dataset as a standalone system and as consolidator in a multi-annotator setting

Model	P/R/F	Consolidation P/R/F
GPT 4.1	0.77/0.77/0.76	0.74/0.72/0.71
GPT 4o	0.77/0.77/0.76	0.74/0.71/0.71
GPT o3	0.73/0.71/0.70	0.77/0.74/0.73

Table 2: Performance of GPT models in classifying JMBX dataset as a standalone system and as consolidator in a multi-annotator setting

Our analysis shows that the flagship GPT models performed well in a single-annotator setup on both datasets compared to the reasoning models. However, their performance declined in the multi-annotator setup, while the reasoning model(o3) improved. This highlights that performance in a multi-annotator setup depends on annotator quality: when annotators perform poorly, they reduce the consolidator’s effectiveness, even when the consolidator is instructed not to aggregate results but to decide after careful analysis.

3.1.2 Open-source LLMs

We repeat the same experiments using open-source LLMs: Qwen3-32B, QwQ-32B, and DeepSeek-R1

Llama 3.3 70B. The performance in standalone and consolidator configurations is shown in Table 3 (PTC) and Table 4 (JMBX). The result shows a similar pattern to the proprietary models. The reasoning model here, QwQ, performed well in the multi-annotator setting, whereas the other two models performed better in the single-annotator setting and the performance dipped due to a poor performance of the reasoning model. This shows that consolidation is only as good as the individual annotations in a multi-annotator setting.

Model	P/R/F	Consolidation P/R/F
Qwen3-32B	0.78/0.76/0.75	0.79/0.73/0.72
QwQ-32B	0.79/0.70/0.68	0.79/0.72/0.71
DeepSeek R1*	0.80/0.76/0.75	0.79/0.73/0.71

Table 3: Performance of open-source LLMs in annotating PTC dataset as a standalone system and as consolidator.*DeepSeek-R1-Distill-Llama-70B

Model	P/R/F	Consolidation P/R/F
Qwen3-32B	0.70/0.70/0.70	0.71/0.69/0.69
QwQ-32B	0.70/0.65/0.63	0.72/0.69/0.67
DeepSeek R1*	0.70/0.69/0.68	0.68/0.66/0.65

Table 4: Performance of open-source LLMs in annotating JMBX dataset as a standalone system and as consolidator.*DeepSeek-R1-Distill-Llama-70B

Overall, open-source LLMs performed better on news articles than on social media texts.

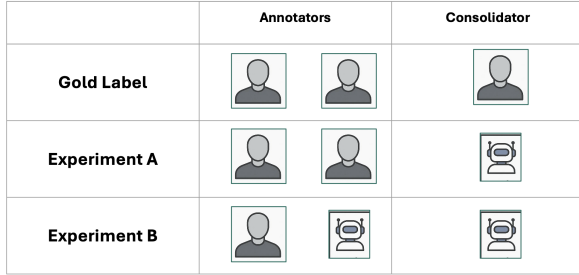


Figure 2: Hybrid Annotation experiments

Among these, DeepSeek-R1 showed consistently strong performance across tasks. Interestingly, the reasoning-focused models (GPT-o3 and QwQ-32B) had the weakest performance when used alone; however, their scores improved in multi-annotator setups, likely benefiting from the additional contextual input provided by other better-performing models.

3.2 Hybrid annotation

To complete our evaluation, we conducted Experiment 3, involving a hybrid setup with both human and LLM collaborators. Because the JMBX dataset includes annotations from two human annotators and one human consolidator, it enabled us to simulate and compare mixed human–AI annotation workflows.

In the first hybrid configuration, annotations by human annotators were passed to LLM models for final consolidation. The results are shown in Table 5, revealing that hybrid setups outperformed fully automated configurations.

Model	Consolidation P/R/F	Class-wise P/R P(prop), R(prop), P(non-prop), R(non-prop)
GPT 4.1	0.79/0.79/0.78	0.76/0.83/0.81/0.74
GPT 4o	0.85/0.84/0.84	0.86/0.83/0.83/0.86
GPT o3	0.78/0.76/0.75	0.70/0.90/0.86/0.61

Table 5: Performance of hybrid setup (experiment A) with GPT models as consolidators on the JMBX dataset. P/R/F are weighted-averaged. The last column lists class-wise P/R as (*prop*, *non-prop*).

Model	Consolidation P/R/F	Class-wise P/R P(prop), R(prop), P(non-prop), R(non-prop)
Qwen3-32B	0.76/0.76/0.75	0.74/0.79/0.77/0.72
QwQ-32B	0.75/0.69/0.66	0.62/0.94/0.88/0.43
DeepSeek R1*	0.74/0.70/0.69	0.65/0.89/0.83/0.52

Table 6: Performance of hybrid setup (experiment A) with open-source models as consolidators on the JMBX dataset. *DeepSeek-R1-Distill-Llama-70B P/R/F are weighted-averaged across classes. The last column lists class-wise P/R as (*prop*, *non-prop*).

In a follow-up scenario, we replaced annotations by one human annotator and human consolidator by LLMs, keeping only one human annotation in the loop. GPT-4o acted as the consolidator and GPT-4.1 as the other annotator, while, in a parallel open-source configuration, DeepSeek R1 acted as the consolidator and Qwen3 as the annotator. GPT-4o’s performance with precision/recall/F1 = 0.79/0.79/0.79, with class-specific scores of 0.70/0.90 for propaganda and 0.86/0.61 for non-propaganda. As expected, the performance dropped slightly compared to the previous hybrid setup, but it still exceeded the GPT-4o model’s performance in its standalone and consolidation roles (see Table 2). A similar trend was observed with the open-source models, where precision/recall/F1 was 0.75/0.74/0.74, with class-specific scores of 0.74/0.76 for propaganda and 0.75/0.73 for non-propaganda.

4 Performance Analysis

To better understand model behavior and failure points, we analyze misclassifications across both the propaganda and non-propaganda classes. Specifically, we examine class-wise precision and recall to identify which classes the models over-predict or struggle to detect accurately. These detailed metrics are shown in Appendix C Tables 7 through 10.

A consistent pattern across models is low precision and high recall for the propaganda class. This indicates that models tend to over-predict propaganda, often labeling neutral sentences as propagandistic. Among the GPT models, performance is generally stronger on social media text than on news articles. Notably, GPT-4o performs particularly well on JMBX with balanced results: 0.78 precision and 0.73 recall for propaganda, and 0.75 precision and 0.80 recall for non-propaganda. This suggests GPT-4o may be more attuned to the infor-

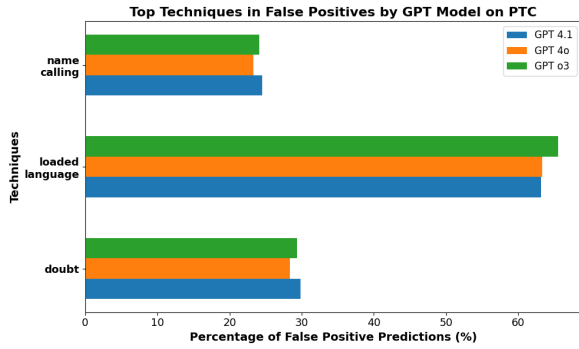


Figure 3: Top Propaganda Techniques in false positive by GPT models on PTC dataset

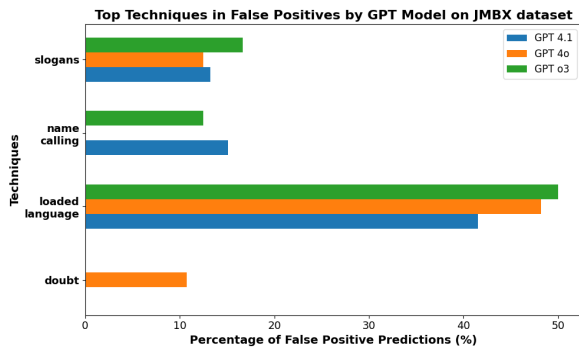


Figure 4: Top Propaganda Techniques in false positive by GPT models on JMBX dataset

mal language styles common in social media.

Open-source LLMs show high recall but relatively low precision for propaganda, mirroring the aggressive classification pattern seen in GPTs. Qwen3-32B, however, stands out with the most balanced class-wise performance across both datasets. For PTC, it achieves a precision of 0.70 and recall of 0.90 for propaganda, while also maintaining relatively high scores for non-propaganda (0.86 precision, 0.61 recall). On the JMBX dataset, Qwen3-32B remains strong and balanced, outperforming other local models in non-propaganda recall.

Given the high false positive rate across models, we examine the propaganda techniques where human-LLM disagreement was greater. Although a full fine-grained analysis is beyond the scope of this work, we used technique labels assigned by the models in false-positive cases to inspect which rhetorical patterns were overused.

As shown in Figures 3 and 4, “Loaded Language” was the most common technique in misclassifications by GPT models. This suggests that GPTs are highly sensitive to emotionally charged or emphatic wording, even in non-propagandistic sentences. The second most common techniques

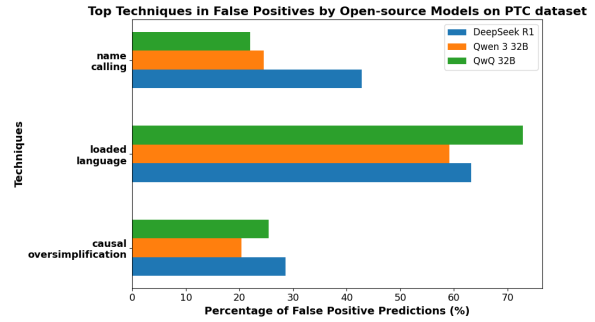


Figure 5: Top Propaganda Techniques in false positive by open source models on PTC dataset

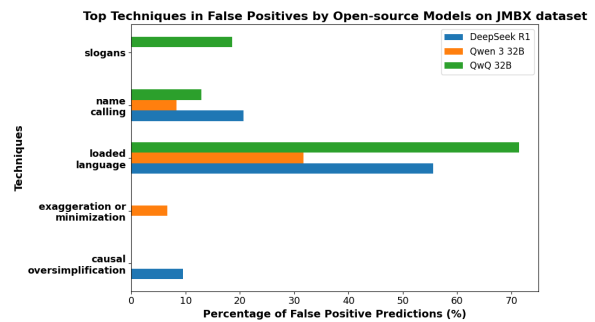


Figure 6: Top Propaganda Techniques in false positive by open source models on JMBX dataset

were “Name Calling” and “Doubt,” which may indicate a bias toward detecting adversarial or skeptical tones [Martino et al.](#). On the JMBX dataset specifically, false positives frequently included “Slogans,” likely triggered by tweet-formatting conventions such as all caps or punchy phrases.

Open-source LLMs demonstrated similar trends. As shown in Figures 5 and 6, “Loaded Language” again appeared as the dominant technique in false positives. However, these models also frequently mislabeled sentences with “Causal Oversimplification” and “Exaggeration or minimization” indicating error patterns that differ from those of the GPT models. These techniques often involve broad generalizations or emotionally overstated claims, which may be harder to contextualize without deeper sentence-level understanding.

5 Qualitative Analysis of human-LLM disagreements

Quantitative metrics are necessary but insufficient when interpretive ambiguity and reasoning quality matter: a model may diverge from a gold label yet offer a defensible rationale. Recent work by [Belinkov and Glass \(2019\)](#) calls for qualitative analyzes to expose such nuances and complement ag-

gregate scores. Error-typology analysis, in particular, groups mismatches into cohorts to reveal failure patterns, overreach, and reasonable disagreement, and prior studies urge looking beyond fluent outputs to the quality of underlying reasoning (Narang et al., 2020).

In this section, we run a structured qualitative analysis of LLM–human disagreements, focusing on prevalent false positives where GPT predictions diverge from human gold labels across both datasets. Our approach draws on behavioral testing and cohort-based diagnostics (e.g., GLUE diagnostics, NLI stress tests), where naming reproducible error types supports targeted evaluation and debugging.

Methodologically, we treat disagreement as signal, consistent with evidence that some conflicts reflect genuine ambiguity rather than annotator error (Pavlick and Kwiatkowski, 2019). We group mismatches into three categories by the interpretive value of the LLM’s explanation: (i) Defensible Divergences (reasonable alternate readings), (ii) Overreaching Reasoning (technically plausible but contextually weak), and (iii) Unconvincing/Unsupported Reasoning (clear failures). This reason-aware framing complements standard metrics and helps separate true model errors from interpretive disagreements.

5.1 Defensible Divergence

In this case, the LLM prediction does not match the gold label, but the reasoning depicts rhetorical awareness and interpretive depth that makes the label plausible. These are not errors in the conventional sense, but alternative valid readings of rhetorical cues. This category highlights a key limitation of rigid gold labels and bias of human subjects in subjective tasks.

Text: *"What could possibly go wrong at Norwich?"*

- **LLM annotation:** Propaganda(Doubt)
- **Justification:** The sentence ‘what could possibly go wrong at Norwich?’ is utilizing the technique of doubt by questioning the credibility or reliability of the situation or event occurring at Norwich. It implies skepticism about the outcome without providing specific details or evidence, leading the audience to question or doubt what might occur there.

The above example suggest that strict disagreement with gold labels may underestimate the in-

terpretive capacity of LLMs, especially when the model is attuned to subtle rhetorical stimulation.

5.2 Reasonable but Overreaching

In the following example, the LLM’s reasoning follows the formal definition of a propaganda technique, but its application feels forced or contextually inappropriate. These mismatches reflect overgeneralization where the model applies technique labels to rhetorical elements that are technically present but not persuasive or manipulative in intent.

Text: *"As a firearms expert, he questioned the lack of flashes coming from the hotel windows where the shooter was supposedly firing from."*

- **LLM annotation:** Propaganda(Appeal to Authority, Doubt)
- **Human annotation:** Non-propaganda
- **Justification:** The sentence uses ‘as a firearms expert’ to give credibility to the speaker’s claim; it also includes ‘questioned the lack of flashes’, which fits the doubt technique.

The example highlight how surface-level match to technique definition can mislead models when they are not grounded in pragmatic or discourse-aware understanding.

5.3 Unconvincing or Unsupported Reasoning

In the final category, LLM neither aligns with the gold label nor offers a compelling justification for its choice. The model’s explanations rely heavily on partial lexical cues or structural repetition, without demonstrating real understanding of rhetorical function or context. Here’s an example of model failure.

Text: *"Completing this poll grants you access to Freedom Outpost updates free of charge."*

- **LLM annotation:** Propaganda(Loaded Language)
- **Justification:** The term ‘freedom’ in ‘Freedom Outpost’ carries a strong positive emotional connotation, appealing to readers’ favorable feelings toward the concept of freedom.

This qualitative analysis reveals that not all mismatches are same. While some represent genuine failures, others demonstrate defensible divergence

or overgeneralization, both of which carry implications for how we should evaluate LLM performance and question over-reliance on human annotation as gold label for subjective task. These divergences can prove its usefulness to extend human-insights or question human’s decision in their annotation. Through this we aim to show the nuance in model’s performance and show importance of holistic perception in such task.

6 Conclusion

This study investigates LLMs for propaganda annotation in single annotator, multi-annotator, and hybrid configurations. We test multiple proprietary GPT models and open-source models (e.g., Qwen, DeepSeek R1) on news and social media datasets to evaluate annotation capability and identify failure points.

LLM performance falls short of human accuracy. Human–LLM hybrids consolidated by GPT-4o come closest to experts, followed by single-annotator LLM. Multi-annotator LLMs perform worst, with outcomes tied to the accuracy of the consolidated individual annotations.

Most models show low recall for the non-propaganda class, yielding many false positives. To probe these errors, we analyzed techniques assigned in false positives. “Loaded Language” was most overused, followed by “Name Calling” and “Doubt”; open source models sometimes mapped non-propagandistic content to “Causal Oversimplification.” This indicates high sensitivity to certain techniques, so neutral content is often labeled accordingly.

Using an error typology, we categorized human–LLM disagreements into three groups based on LLMs’ stated reasoning. These categories show misclassifications are not equal, informing debates on LLMs in annotation workflows and highlighting both potential and limitations.

7 Limitations

While our experiments offer useful insights into LLM capabilities for propaganda detection in standalone and multi-annotator settings, several limitations remain. Our analysis centers on binary classification. Although we include exploratory technique-level analysis in our qualitative evaluations, we do not perform full multi-label classification. Consequently, the models’ ability to identify multiple propaganda spans within a single sen-

tence is underexplored. Additionally, the number of sentences included in our evaluation is relatively small due to budgetary and token-usage constraints, which limited the scale of our experimental setup. These limitations suggest several future directions: larger-scale evaluations, fine-grained multi-label annotation, iterative multi-annotator collaboration, and human-in-the-loop verification.

8 Ethical Considerations

This study is limited to propaganda detection using fixed definitions and task-specific prompts. A key concern is bias in LLM training data, which may influence annotation outcomes despite objective definitions. While we did not evaluate bias directly, its impact on reliability, especially in politically or culturally sensitive contexts, warrants further investigation.

We also caution against unsupervised pipelines that use LLM-generated annotations as training data. LLMs can hallucinate or misclassify ambiguous inputs, which is problematic in high-stakes domains. Mislabeling satire, dissent, or stylistic variation as propaganda can have ethical consequences in real-world deployments. Finally, while LLMs can support or automate annotation, human oversight remains essential, especially for consolidating judgments, handling edge cases, and ensuring transparency and accountability. LLMs can benefit further from human–LLM collaboration via iterative feedback.

9 Acknowledgment

This work used computational resources provided by the National Artificial Intelligence Research Resource (NAIRR) Pilot under Allocation NAIRR240383.

References

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Toufique Ahmed, Premkumar Devanbu, Christoph Treude, and Michael Pradel. 2025. Can LLMs replace manual annotation of software engineering ar-

- tifacts? In *Proceedings of the 22nd IEEE/ACM International Conference on Mining Software Repositories (MSR)*, pages 526–538. IEEE.
- Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, and 1 others. 2024. Investigating annotator bias in large language models for hate speech detection. In *Neurips Safe Generative AI Workshop 2024*.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.
- Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.
- Preni Golazizian, Alireza Salkhordeh Ziabari, Ali Omrani, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3474–3491.
- Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.
- Kyle Hamilton, Luca Longo, and Bojan Bozic. 2024. Gpt assisted annotation of rhetorical and linguistic features for interpretable propaganda technique detection in news text. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1431–1440.
- Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2023. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.
- Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.
- Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.
- Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kardos, Mia Jacobsen, and Lene Aarøe. 2023. Chatbots are not reliable text annotators. *arXiv preprint arXiv:2311.05769*.
- Hao Lin and Yongjun Zhang. 2025. The risks of using large language models for text annotation in social science research. *arXiv preprint arXiv:2503.22040*.
- Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pages 48–55.
- G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.
- Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the evaluator: Measuring llms’ adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19589–19597.
- Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.
- Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, 6(1):4–5.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.
- Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through llm-driven active learning and human annotation. *arXiv preprint arXiv:2406.12114*.
- Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. Towards coding social science datasets with language models. *arXiv preprint arXiv:2306.02177*.
- Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.
- Vivek Sharma, Mohammad Shokri, Sarah Levitan, Elena Filatova, and Shweta Jain. 2025a. *Analysis of propaganda in tweets from politically biased sources*.

The International FLAIRS Conference Proceedings, 38.

Vivek Sharma, Mohammad Mahdi Shokri, Sarah Ita Levitan, Elena Filatova, and Shweta Jain. 2025b. Analysis of propaganda in tweets from politically biased sources.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Petter Törnberg. 2024a. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Petter Törnberg. 2024b. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, page 08944393241286471.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

A More Quantitative Analysis Examples

A.1 Defensible Divergence

Some other examples of defensible divergence in the dataset are

- **Text:** *"The unfortunate Dowden suffers from memory loss."*
 - **LLM annotation:** Propaganda(Name Calling / Labeling)

- **Justification:** The reasoning for selecting 'name calling or labeling' is that the sentence labels Dowden as 'unfortunate' and suffering from 'memory loss'. The use of 'unfortunate' implies a negative connotation, thus labeling Dowden in a potentially undesirable way to influence perceptions.

- **Text:** *"Sorry, man—we're not giving up cake no matter the nuance."*

- **LLM annotation:** Propaganda(Thought-Terminating Cliché)

- **Justification:** The phrase 'no matter the nuance' suggests a refusal to consider any complexities or nuances related to the idea of giving up cake. This fits with the definition of a thought-terminating cliché, as it dismisses further discussion or critical thought on the topic by simplifying the decision to an absolute dismissal.

A.2 Reasonable but Overreaching

More examples of reasonable but overreaching ones are

- **Text:** *"While Democratize SSMU later apologized for the language and deleted it from its site, the damage had already been done, Boudreau concluded."*

- **LLM annotation:** Propaganda(Minimization or Exaggeration)

- **Justification:** The phrase 'the damage had already been done' uses minimization or exaggeration to present the consequences of the original language (which was deleted and apologized for) as significant and irreversible, possibly overstating the impact without providing evidence for the extent of the damage.

- **Text:** *"For those heavy into it, that's their thing, but it ain't my thing."*

- **LLM annotation:** Propaganda(Thought-Terminating Cliché)

- **Justification:** The phrase is a generic statement that simplifies the situation and shuts down further discussion by accepting others' involvement without engaging in deeper exploration or critique.

A.3 Unconvincing or Unsupported Reasoning

- **Text:** *"All they have to do is delay the confirmation until after the election."*
 - **LLM annotation:** Propaganda(Oversimplification)
 - **Justification:** The sentence implies that a single, simple action—delaying the confirmation—is sufficient to achieve the desired outcome, thereby reducing a complex political process with multiple variables to one straightforward cause-and-effect step
- **Text:** *"Each Carmel has its own spirit, its own tempo, its own feel."*
 - **LLM annotation:** Propaganda(Loaded Language, Repetition)
 - **Justification:** Words such as 'spirit', 'tempo', 'feel', and the emotive notion of a personal 'call to a specific Carmel' carry spiritually charged connotations meant to inspire or move the reader. The phrase 'its own' is repeated three times to reinforce uniqueness.

B Prompt

Annotation prompt: "You are an annotator for a neutral corporation that provides unbiased annotation services. You remain impartial to any political or ideological stance, such as left, right, or center. Your task is to detect text-based propaganda objectively, strictly adhering to the provided definition without introducing any personal or ideological bias."

Consolidation instruction: "You are a consolidator for a neutral corporation that provides unbiased annotation after reading annotations and reasoning by three annotators and definition of propaganda techniques. You remain impartial to any political or ideological stance, such as left, right, or center. As a subject matter expert, your job is to detect text-based propaganda, objectively, without introducing any personal or ideological bias. As an expert you are free to add or remove relevant or irrelevant techniques respectively based on the definitions. The annotators have worked individually based on the same definitions. Try to understand the view of annotators but be critical in your judgment before making the decision."

Definitions of rhetorical techniques: "Below is a list of rhetorical techniques with their definitions:

- loaded language: Using words/phrases with strong emotional implications (positive or negative) to influence an audience.
- name calling or labeling: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises.
- repetition: Repeating the same message over and over again, so that the audience will eventually accept it.
- exaggeration or minimization: Either representing something in an excessive manner: making things larger, better, worse.
- doubt: Questioning the credibility of someone or something. - appeal to fear/prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments.
- flag-waving: Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea.
- causal oversimplification: Assuming one cause when there are multiple causes behind an issue. We include scapegoating as well which is defined as the transfer of the blame to one person or group of people without investigating the complexities of an issue.
- slogans: A brief and striking phrase that may include labeling and stereotyping.
- appeal to authority: Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence. Include the special case where the reference is not an authority/expert, although it is referred to as testimonial in the literature.
- black-and-white fallacy: Presenting two alternative options as the only possibilities, when in fact more possibilities exist, eliminating any other possible choice. and as an extreme telling the audience exactly what actions to take, which is also called as dictatorship.
- thought-terminating cliché: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are

typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought.

- whataboutism: Discredit an opponent’s position by charging them with hypocrisy without directly disproving their argument.
- reductio ad hitlerum: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.
- red herring: Introducing irrelevant material to the issue being discussed, so that everyone’s attention is diverted away from the points made.
- bandwagon: Attempting to persuade the target audience to join in and take the course of action because “everyone else is taking the same action”.
- obfuscation/intentional vagueness/confusion: Using deliberately unclear words, so that the audience may have its own interpretation."

C Performance Analysis - Precision/Recall

The following table shows the class-wise precision and recall of both propaganda and non propaganda class as Precision(propaganda), recall (propaganda), precision (non-propaganda), and recall(non-propaganda). These tables shows the per-class performance of P/R/F results shown in Table 1– 4.

Table 7

Model	PTC	JMBX
	P/R (prop, non-prop)	P/R (prop, non-prop)
GPT 4.1	0.68/0.91/0.87/0.58	0.75/0.79/0.78/0.74
GPT 4o	0.68/0.94/0.90/0.55	0.78/0.73/0.75/0.80
GPT o3	0.64/0.96/0.92/0.45	0.66/0.86/0.80/0.56

Table 7: GPT model performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as (*prop, non-prop*).

The following table 8 shows per-class precision and recall for GPT models as **consolidators** in a multi-annotator setting.

Model	PTC	JMBX
	P/R (prop, non-prop)	P/R (prop, non-prop)
GPT 4.1	0.64/0.96/0.92/0.47	0.67/0.86/0.81/0.58
GPT 4o	0.63/0.97/0.93/0.43	0.66/0.88/0.82/0.55
GPT o3	0.64/0.96/0.92/0.46	0.68/0.90/0.85/0.58

Table 8: GPT model performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as (*prop, non-prop*).

The following table for open source model per class precision and recall.

Model	PTC	JMBX
	P/R (prop, non-prop)	P/R (prop, non-prop)
Qwen3-32B	0.70/0.90/0.86/0.61	0.72/0.67/0.69/0.73
QwQ-32B	0.63/0.97/0.94/0.44	0.60/0.90/0.80/0.40
DeepSeek R1	0.68/0.95/0.92/0.56	0.65/0.81/0.75/0.56

Table 9: Open-source LLMs performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as (*prop, non-prop*).

The following table 8 shows per-class precision and recall for GPT models as **consolidators** in a multi-annotator setting.

Model	PTC	JMBX
	P/R (prop, non-prop)	P/R (prop, non-prop)
Qwen3-32B(c)	0.66/0.95/0.91/0.52	0.66/0.81/0.75/0.58
QwQ-32B(c)	0.65/0.96/0.92/0.49	0.63/0.89/0.81/0.48
DeepSeek R1(c)	0.66/0.96/0.93/0.50	0.62/0.83/0.74/0.49

Table 10: Open-source LLMs performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as (*prop, non-prop*).

Emoji Reactions on Telegram: Unreliable Indicators of Emotional Resonance

Serena Tardelli¹, Lorenzo Alvisi^{1,2}, Lorenzo Cima^{1,3}, Stefano Cresci¹, Maurizio Tesconi¹

¹Institute of Informatics and Telematics, National Research Council, Italy

²IMT School for Advanced Studies Lucca, Italy

³Dept. of Information Engineering, University of Pisa, Italy

{firstname.lastname}@iit.cnr.it

Abstract

Emoji reactions are a frequently used feature of messaging platforms, yet their communicative role remains understudied. Prior work on emojis has focused predominantly on in-text usage, showing that emojis embedded in messages tend to amplify and mirror the author’s affective tone. This evidence has often been extended to emoji reactions, treating them as indicators of emotional resonance or user sentiment. However, they may reflect broader social dynamics. Here, we investigate the communicative function of emoji reactions on Telegram. We analyze over 650k crypto-related messages that received at least one reaction, annotating each with sentiment, emotion, persuasion strategy, and speech act labels, and inferring the sentiment and emotion of emoji reactions using both lexicons and LLMs. We uncover a systematic mismatch between message and reaction sentiment, with positive reactions dominating even for neutral or negative content. This pattern persists across rhetorical strategies and emotional tones, indicating that emojis used as reactions do not reliably function as indicators of emotional mirroring or resonance of the content, in contrast to findings reported for in-text emojis. Finally, we identify the features that most predict emoji engagement. Overall, our findings caution against treating emoji reactions as sentiment labels, highlighting the need for more nuanced approaches in sentiment and engagement analysis.

1 Introduction

Emoji reactions are a central feature of digital platforms, enabling users to respond quickly and visually to content. While much research has focused on in-text emojis (i.e., linguistic elements embedded within authored messages), emoji reactions (i.e., reactions appended by readers as lightweight feedback mechanisms), remain comparatively underexplored. Yet they are equally revealing, as they

are instantaneous, costless, and a quick way to interact with content without composing a reply. Moreover, while prior work has examined in-text emojis, showing they amplify sentiment (Shiha and Ayvaz, 2017), improve sentiment classification (Liu et al., 2021), and provide emotional nuance (Lou et al., 2024; Khan et al., 2025), this understanding has also been implicitly extended to emoji reactions, which have been used as proxies for emotional labels in distant supervision settings (Pool and Nissim, 2016; Plaza Del Arco et al., 2018; Raad et al., 2018). However, recent work suggests that emoji reactions may instead express pragmatic or social meanings, such as to approve or normalize toxic discourse (Morales et al., 2025).

Building on this idea, we provide the first large-scale study examining whether emoji reactions actually reflect emotional resonance, an assumption often implicitly adopted when reactions are used as proxy labels in sentiment analysis. To do so, we investigate the communicative function of emoji reactions on Telegram, a messaging platform widely used for news, politics, and activism. On Telegram, emoji reactions are publicly visible and frequently used, making them a rich signal of collective response. In detail, we analyze over 650k crypto-related Telegram messages and their associated emoji reactions. We combine large-scale emotion classification with rhetorical analysis to examine the alignment between message tone and reaction type. In particular, we address three research questions:

- **RQ1:** Do emoji reactions align with the sentiment expressed in the original message?
- **RQ2:** Do emoji reactions reflect the emotional tone of the original message, creating emotional resonance?
- **RQ3:** Which message features predict emoji reactions?

We adopt a mixed-methods approach to examine messages and their reactions across multiple dimensions, including sentiment, emotional tone, persuasion strategy, and speech act. We find that positive emoji reactions dominate, even in response to messages that are emotionally neutral or negative. This consistent mismatch holds across rhetorical strategies and emotional categories, suggesting a pragmatic function for some emoji reactions, more closely tied to social approval or support than to shared emotional resonance. Finally, we show that certain rhetorical strategies are strong predictors of emoji engagement, and that emotionally neutral or negative messages tend to attract more emoji reactions overall. Our contributions are manifold:

- We provide the first large-scale analysis of emoji reactions on Telegram to determine whether they distort or amplify emotional alignment.
- We conduct our analysis on a large dataset of 650k crypto-related Telegram channels, a domain characterized by strong community identity and high engagement, where in-text emojis are known to play a meaningful role in shaping financial sentiment (Mahrous et al., 2023; Zuo et al., 2024), while emoji reactions remain largely unexplored.
- We show a consistent misalignment between message sentiment and emoji reactions, indicating that, unlike in-text emojis, reactions do not reliably act as emotional mirroring.
- We identify specific message features, such as emotional framing or persuasive tactics, that predict user engagement in terms of emoji reactions.

Our findings suggest that emoji reactions are not reliable proxies for sentiment content, as they weakly mirror or resonate with the emotion evoked by the message. Together, these results challenge the assumption that emoji reactions reliably reflect emotional response, as commonly observed for in-text emojis. These patterns should also be interpreted in light of Telegram’s specific communicative context, where community norms may shape reaction behavior. More broadly, these findings have important implications for research on misinformation, particularly in political and financial contexts, where reaction-based signals can be in-

terpreted as indicators of emotional resonance or public sentiment.

2 Related work

Emojis have been extensively studied as markers of affective communication in digital discourse. Early work recognized their role as nonverbal cues that supplement text with emotional tone (Miller et al., 2016; Kralj Novak et al., 2015), and many researchers treated them as reliable proxies for emotion. Subsequent studies showed that emojis can amplify sentiment in social media messages (Shiha and Ayvaz, 2017). Other research found that incorporating emojis as features can significantly improve sentiment classification accuracy (Liu et al., 2021). More recent work highlighted how emojis provide emotional context in posts highlighting the role of emojis as emotional indicators (Lou et al., 2024; Zhou et al., 2025). Similar work showed that emojis play a vital role in sentiment expression, often conveying more explicit and nuanced information than the text alone (Khan et al., 2025). As a result, datasets such as the one introduced by Kralj Novak et al. (2015), which assign sentiment scores to individual emojis based on crowd-sourced annotations, have been widely used for downstream sentiment classification. Similarly, the DeepMoji model (Felbo et al., 2017) was trained on large-scale Twitter data labeled with emojis in order to learn emotional representations, relying on the assumption that emoji usage reflects the author’s emotional intent.

This assumption has been extended to emoji reactions, i.e., predefined emoji responses users can select to express a reaction to content on platforms such as Facebook. These reactions have often been interpreted as ground truth labels for emotion classification tasks (Pool and Nissim, 2016; Plaza Del Arco et al., 2018; Raad et al., 2018), enabling models to learn emotion categories from reaction distributions. However, more recent studies caution against this approach. Graziani et al. (2019) observed that reactions can be semantically ambiguous or perform multiple pragmatic functions. Paolillo (2023) noted that reactions can operate not only as affective signals but as interactional cues, marking approval, affiliation, or irony depending on context.

Our work extends these critiques by examining emoji reactions on Telegram, a platform where reactions are public, optional, and highly community-

driven. Unlike prior research that treats emoji reactions as emotional response, we find systematic mismatch between the sentiment expressed in a message and the sentiment inferred from emoji reactions. Messages with negative or neutral tone frequently receive positive emoji reactions, such as the thumb up or love emoji. This suggests that reactions function more as social endorsement than as expressions of shared affect. The result challenges the methodological practice of using reaction emojis as emotional supervision signals, at least in environments where group identity and approval play a stronger role than emotion mirroring.

On a similar line, recent work (Morales et al., 2025) explored how emoji reactions are used in toxic versus non-toxic contexts on Telegram, showing that even toxic comments often receive positive emoji reactions, signaling approval or normalization of harmful content. Our findings complement and generalize this pattern: while they focused specifically on toxicity, we observe that positive reaction bias holds across a broader range of communicative intentions, including persuasive tactics, emotional tone, and speech acts. Moreover, while our results do not exclude the possibility of emotional mirroring, they suggest that social signaling frequently overrides it in practice.

Further, prior work on emotional dynamics in engagement showed that emotionally charged content tends to propagate faster (Vosoughi et al., 2018), and that moral-emotional language increases virality within ideologically homogeneous groups (Brady et al., 2017). Our findings complement this literature by showing that the type of emotional response conveyed through reactions is not always aligned with message tone, and instead may reinforce in-group signaling mechanisms. This distinction between emotion expression and interactional feedback is especially relevant for studying engagement in polarized or activist spaces, where affective reactions might serve to signal allegiance more than to reflect emotional resonance.

By focusing on Telegram, which remains comparatively understudied despite its growing role in political and financial discourse and in information diffusion, we contribute new evidence on communication dynamics within this platform. In fact, while recent studies have examined toxicity (Alvisi et al., 2025b,a) and misinformation (Urman and Katz, 2022; Alvisi et al., 2024) on Telegram, as well as their association with emoji usage (Morales et al., 2025), the communicative function of emoji

reactions remains largely unexplored, a gap that we address in the present study.

3 Dataset

We construct our dataset by collecting Telegram channels listed on TGStat¹, a third-party analytics platform that indexes and categorizes public channels by topic, language, and popularity. From this source, we select channels related to the cryptocurrency topic, as it is an area where emojis are known to play a meaningful role in shaping financial sentiment (Mahrous et al., 2023; Zuo et al., 2024). We collect those English-language channels with at least 5,000 subscribers, resulting in an initial sample of 20,695 channels. For each channel, we retrieve all publicly available messages posted between January and December 2024. The distribution of messages follows a long-tail pattern consistent with a power-law distribution, where a small number of channels account for the majority of messages.

Since our analysis focuses on messages that received at least one emoji reaction, we further filter the dataset accordingly. In particular, we retain only messages that include at least one standard (default) emoji reaction, excluding custom emojis, which may be enabled depending on channel settings. After this filtering, the final dataset consists of 647,879 messages from 993 channels.

4 Methods

We analyze Telegram messages using a multi-stage pipeline that combines NLP-based linguistic annotation with behavioral cues derived from user reaction data (i.e., emoji-based reactions). First, we automatically annotate each message with linguistic features, including sentiment, persuasion strategies, emotion, and speech act categories. Then, we automatically label the emoji reactions to infer both their sentiment polarity and emotional content. This allows capturing both linguistic content and social responses.

Message Annotation We enrich each Telegram message with linguistic and psychological features through automatic labeling using multiple classification layers. For each annotation task, although multiple labels may co-occur within a single message, we adopt a single-label (multiclass) formu-

¹<https://tgstat.com/>

lation and retain only the highest-scoring label returned by each classifier.

- *Persuasion strategies*: We detect rhetorical appeals of each message (e.g., emotion, authority, scarcity, social proof, fear, popularity, logic, reciprocity, fear of missing out) plus a neutral and a no persuasion categories using the *valhalla/distilbart-mnli-12-1* zero-shot model (Valhalla, 2020) and the taxonomy by Qachfar and Verma (2023).
- *Emotion analysis*: We assign each message an emotion from (Plutchik, 1994)’s taxonomy using the *bhadresh-savani/bert-base-uncased-emotion* model (Savani, 2021).
- *Sentiment analysis*: We label messages as positive, neutral, or negative using *cardiffnlp/twitter-roberta-base-sentiment* (CardiffNLP, 2021), widely adopted in computational social science (Jahin et al., 2024).
- *Speech acts*: We classify messages as assertives, directives, commissives, expressives, and declarations based on Searle (1969) taxonomy (Jegade, 2024; Saha et al., 2020) using the *valhalla/distilbart-mnli-12-1* zero-shot model (Valhalla, 2020).

Emoji Reaction Interpretation We infer the emotional and affective content of reactions using two complementary methods:

- *Emoji sentiment score*: We use the emoji sentiment lexicon (Kralj Novak et al., 2015)², which assigns each emoji a polarity (−1, 0, +1). For each message, we compute a weighted average score (−1 to +1) based on reaction frequencies, then map it to positive (≥ 0.2), negative (≤ -0.2), or neutral, as done in prior work (Chifu et al., 2015; Chaithra et al., 2019) showing that defining a neutral range around zero reduces label noise and improves the robustness of downstream sentiment analyses. This mapping allows us to assign a discrete sentiment label to the emoji reaction set for each message.
- *LLM-based emoji emotion*: We prompt OpenAI GPT-4o model with the list of reaction

²<https://www.clarin.si/repository/xmlui/handle/11356/1048>

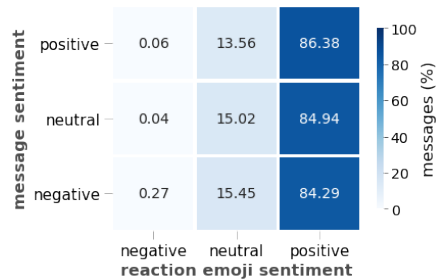


Figure 1: Message sentiment vs emoji reactions sentiment. Positive emoji reactions dominate independently of message sentiment.

emojis to assign an emotion from Plutchik (1994)’s taxonomy (coarse analysis) or a fine-grained emotional label, allowing for more nuanced and free interpretation beyond the core categories. In this way, we can also assign each message the emotion of its most frequent reaction emoji.

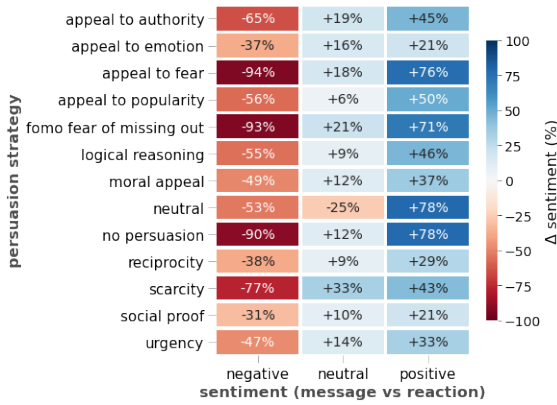
Predictive Modeling and Feature Interpretation

We investigate which aspects of a message predict its likelihood of receiving a large number of emoji reactions. We frame the task as a binary classification problem. We train a binary Random Forest classifier (Breiman, 2001) with 28 features to distinguish between “high engagement” posts and “low engagement” ones, using the median number of emoji reactions as threshold. The 28 features correspond to the automatically inferred linguistic attributes described above (i.e., sentiment, emotion, persuasion strategy, and speech act), encoded as categorical features. While the classification task is not the core goal, this predictive setup allows us to quantify the impact of individual message-level features. To understand the contribution of each feature to engagement, we leverage SHAP, a game-theoretic method that has become standard in explainable AI (Lundberg and Lee, 2017). SHAP assigns each feature a local importance score for every prediction, allowing us to rank linguistic and emotional cues according to their predictive power, as done in multiple related social media analysis tasks (Gambini et al., 2024; Tessa et al., 2025).

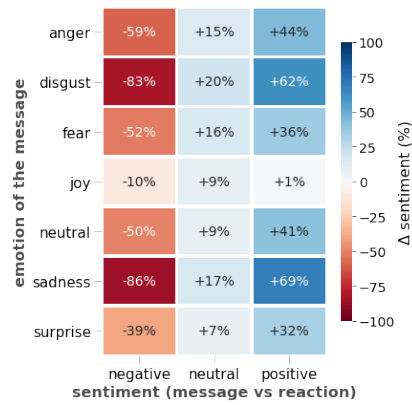
5 Results

5.1 RQ1 - Alignment of emoji reactions with sentiment

In Figure 1 we compare the sentiment conveyed in the message with the sentiment of the emoji reactions it receives. The majority of messages, regard-



(a) Message sentiment vs emoji reactions sentiment, for different persuasive strategies. Most strategies show a consistent shift toward more positive sentiment in reactions. Sentiment distributions in messages and reactions differ significantly across all persuasion strategies (Chi-squared test, $p < 0.001$), revealing a systematic misalignment.



(b) Message sentiment vs emoji reactions sentiment, for different emotions. Negative emotions (e.g., sadness, disgust) receive positive reactions. Sentiment distributions in messages and reactions differ significantly across all emotions (Chi-squared test, $p < 0.001$).

Figure 2: Sentiment shift from messages to reactions by persuasion strategy (left) and emotion category (right).

less of whether they express positive, neutral, or negative sentiment, receive predominantly positive reactions. For instance, over 84% of negative and neutral messages are met with positive emoji sentiment, with less than 0.3% of negative messages receiving negative emoji feedback. This suggests a systematic mismatch between the emotional tone of the message and the sentiment expressed in the reactions, an inconsistency in line with previous work on other platforms (Wang et al., 2023).

To further explore this phenomenon, in Figure 2 we break down the analysis by persuasion strategy and emotion detected in the message. In both cases, we plot the delta in sentiment, that is, the sentiment shift computed as the difference between the sentiment score of the original message and the weighted average sentiment score of the emoji reactions it received. These differential heatmaps reinforce the pattern we previously observed. While the original sentiment expressed in messages spans a wide range, the received sentiment, as inferred from emojis, is skewed toward positive or neutral categories. This suggests a general positivity bias in emoji usage, potentially reflecting users’ preference for agreement, support, or acknowledgment over dissent.

In Figure 2a, this effect holds across most persuasive strategies. Techniques like appeal to fear, fear of missing out, and scarcity, which typically rely on negatively framed content, still elicit predominantly positive emoji reactions, with net sentiment shifts exceeding +70%. However, the magnitude

of this shift varies across strategies. For instance, techniques such as reciprocity, fear of missing out, or appeal to emotion show more balanced distributions, with a smaller difference between message tone and reaction sentiment. This suggests that not all forms of persuasion equally trigger endorsement, and that emotional proximity between message and reaction may play a role in moderating the shift. A similar pattern emerges in Figure 2b, where messages expressing sadness, disgust, or fear are still met with disproportionately positive reactions. However, the shift is not uniform, as messages labeled with joy or surprise show only modest changes, indicating that positive messages are not necessarily amplified through emoji reactions. Interestingly, this tendency is especially evident in response to emotionally neutral content, where reactions may reflect a desire to affirm or engage with others, convey presence, support, or as a form of interpersonal acknowledgment.

5.2 RQ2 - Emoji reactions and emotional resonance

While RQ1 revealed a tendency for messages to receive positive emoji reactions regardless of their sentiment, the analysis was based on a coarse-grained polarity framework (positive, neutral, negative). To better understand which emotions drive this misalignment, RQ2 shifts focus from sentiment to discrete emotion categories. We ask whether certain emoji reactions systematically distort alignment with the emotional tone of the message, and

whether removing them improves interpretability. Specifically, we hypothesize that emoji reactions are not always used to mirror the emotional tone of a message, but instead often serve as signals of social approval, engagement, or banter. To evaluate this, we measure how often the emotion of the message appears among the top-ranked emoji reactions. In other words, we compute the percentage of messages where the top-ranked emoji reaction shares the same emotion category as the message itself. To assess the impact of individual emoji reactions, we perform an ablation analysis. We remove one emotion at a time from the reaction distribution and we re-calculate the top-ranked emoji reaction. If removing a specific emotion significantly increases the number of matches (i.e., if it makes the dominant emoji reaction more likely to align with the emotion of the message), we interpret that category as disruptive to emotional alignment (i.e., the removed reaction was often selected even when it did not match the emotional tone of the message). Figure 3 presents the relative change in alignment match after removing each emoji reaction category, grouped by the emotion label of the message. Each cell shows the percentage change in the proportion of messages where the emotion of the top-ranked emoji matches the emotion of the message. Statistically significant changes, defined as those with a standardized effect size (Cohen’s h) ≥ 0.2 , are marked with an asterisk. Cohen’s h quantifies the magnitude of difference between proportions before and after ablation, providing an interpretable measure of effect strength beyond percentages.

As expected, the lowest scores appear along the diagonal, where the removed emoji category matches the labeled emotion of the message. In fact, removing the correct category naturally lowers alignment, simply because it eliminates true matches. In these cases, the drop does not indicate interference from other emotions, but simply reflects the expected loss in alignment from removing the correct category. The main takeaway from our results is that removing joy consistently leads to large and statistically significant improvements in alignment, especially for messages labeled as sadness, neutral, surprise, and fear. In contrast, removing other reaction emotions has minimal impact. For instance, while removing surprise reactions results in a large relative increase in matches for disgust messages (i.e., +125%), the number of messages labeled with disgust is very small, making this effect statistically negligible.

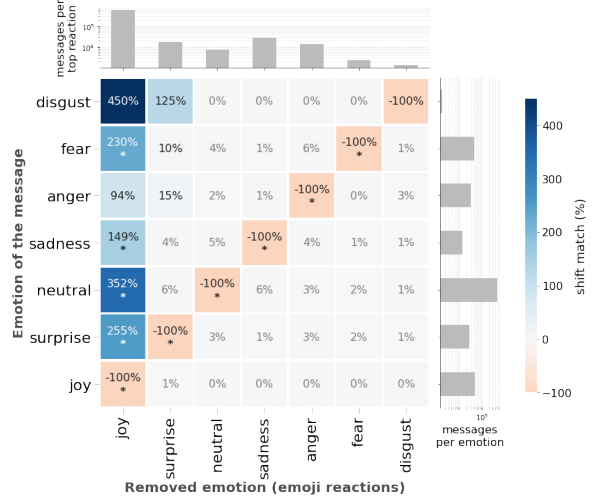


Figure 3: Percentage change in alignment between message emotions and top emoji reactions after removing each emotion. Removing joy leads to large and statistically significant improvements.

The consistent improvement in alignment when joy reactions are removed suggests that these emojis are often used even when they do not match the emotional tone of the message. Unlike other categories, joy appears to systematically distort emotional alignment, pointing to a distinct functional role. Rather than reflecting shared affect, joy reactions may serve pragmatic or social purposes, such as signaling encouragement, approval, banter, or support, regardless of the emotional content of the message. The disruptive effect of joy, revealed through targeted ablation, highlights its unique communicative function in digital contexts, one that is orthogonal to emotional resonance.

Our previous analysis relied a binary metric: whether the top-ranked emoji reaction matched the emotion expressed in the message. While this provides an intuitive signal of alignment, it reduces the full distribution of emoji reactions to a binary outcome. To capture a more nuanced and robust measure of alignment, we compute the cosine similarity between the distribution of emoji reaction emotions and the message emotion represented as a one-hot vector. This allows us to detect subtle changes in alignment, not just categorical shifts. Unlike the previous binary approach, cosine similarity reflects the degree to which the overall emotional tone of emoji reactions matches the intended emotion of the message, making it more sensitive and interpretable. This is particularly important when testing whether the presence of specific emoji emotions obscures rather than reflects affective res-

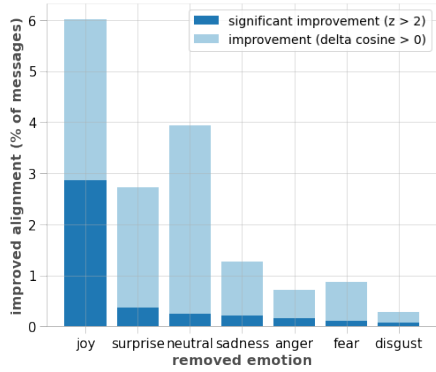


Figure 4: Percentage of messages with improved alignment after removing each emoji reaction emotion. Joy most frequently obscures alignment between message and reaction tone.

onance. To identify which emotions most affect this alignment, we systematically removed each emotion category from the emoji distribution and measured the resulting change in cosine similarity (Δ). A negative effect indicates that removing the emotion improves alignment, suggesting that its presence may distort the emotional signal conveyed by emoji reactions.

Among all categories, joy produced the most pronounced negative effect ($\Delta = -0.107$), followed by surprise, with a much smaller change ($\Delta = -0.006$). Other emotions, such as disgust ($\Delta = -0.0007$), had negligible impact. This suggests that joy-related emojis, more than any other category, systematically misalign with the emotional tone of the message. We then examined, for each emotion removed from the emoji reaction distribution, how often this ablation led to improved alignment with the emotion of the message. Figure 4 shows the results. While all emotions showed low rates of improvement in cosine similarity (i.e., $< 4\%$, with most $< 1\%$), removing joy led to a substantial increase in alignment in over 6% of the messages. Moreover, joy accounted for over 17,000 statistically significant improvements (z-score > 2), surpassing any other emotion. This pattern is not explained simply by frequency. While joy appears in over 95% of emoji reactions, its removal leads to disproportionately many and significantly large alignment improvements, suggesting that joy plays a qualitatively different role. In fact, while sadness appears in over 26% of emoji reactions, removing it leads to improved alignment in just 1.27% of messages, compared to 6.02% for joy. No other emotion, regardless of frequency,

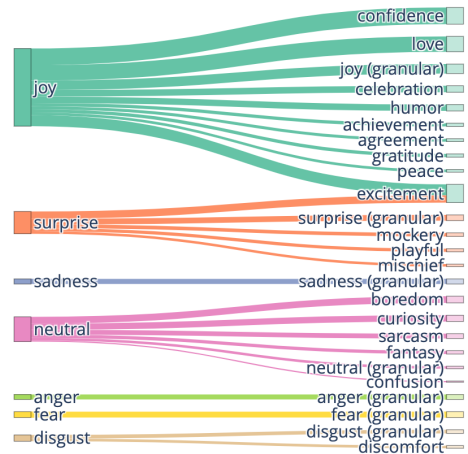


Figure 5: Mapping of coarse-grained emotions (left) to their finer-grained subcategories (right). Even when emojis reflect joy or neutrality, they frequently encode more specific sentiments of approval like agreement, gratitude, and excitement.

produces such a consistently large and disproportionate improvement in alignment when removed. These findings reinforce the idea that joy reactions are not merely emotional mirroring, but often mask or override the underlying sentiment of a message.

To shed light on why joy behaves differently from other emotions, we examine the fine-grained emotional subtypes associated with each coarse emoji emotion category. Figure 5 shows a Sankey diagram mapping each coarse emoji emotion to its fine-grained variants. Notably, while categories like anger or disgust map to narrowly defined emotional expressions, joy includes a wide range of sub-emotions, such as confidence, love, celebration, agreement, and gratitude. These variants often signal social approval, support, or encouragement, rather than shared affect. This broader functional range helps explain why joy reactions are frequently misaligned with the emotional content of messages, suggesting that they do not function as an emotional response but rather as a vehicle for social feedback. The mismatch becomes evident in real examples from the dataset, shown in Figure 6, where joy reactions are frequently used in response to emotionally negative messages such as those expressing sadness, fear, or anger. For example, reacting with a heart emoji to a sad post might mean “I’m here for you” more than “I feel love,” reflecting group dynamics and social norms.

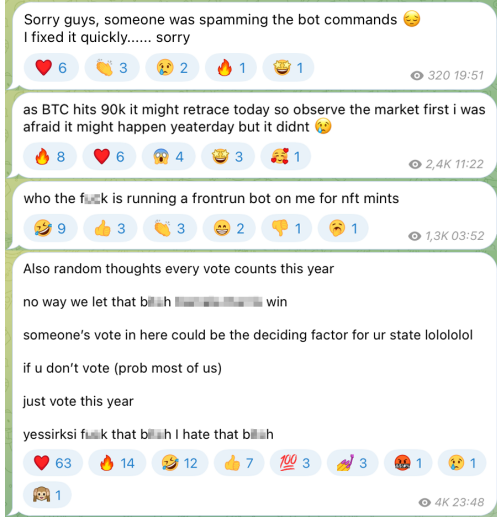


Figure 6: Examples from our Telegram dataset showing that messages expressing negative or neutral emotions receive predominantly positive reactions.

5.3 RQ3 - Prediction of emoji reactions

This research question investigates which message-level features predict the presence of emoji reactions, with the goal of understanding how rhetorical and linguistic strategies influence the engagement of the users through emoji reactions.

To this end, we formulate the task as a binary classification problem, where the model predicts whether a message elicits a high level of emoji reactions based on its textual, emotional, and rhetorical features. We employ a Random Forest classifier (Breiman, 2001) with class balancing, achieving a moderate yet consistent performance, with accuracy = 0.71 and F1 = 0.71. This provides a solid base for further interpreting results through SHAP, which estimates the marginal impact of each input feature on the predictions of the model.

Figure 7 shows the top 15 contributing features (red-colored points represent high, while blue low) and their influence on the prediction toward high engagement. The top predictors of emoji reactions include Emotion Neutral, Sentiment Negative, and (to a lesser extent) Sentiment Positive, suggesting that the emotional valence of a message, whether positive or negative, plays a key role in shaping reaction dynamics. Interestingly, Emotion Neutral emerges as the strongest individual predictor, indicating that even emotionally flat messages can elicit strong social responses. Notably, stronger negative sentiment correlates with higher reaction rates, while less intensely positive sentiment is also linked to higher reactions. This pattern suggests

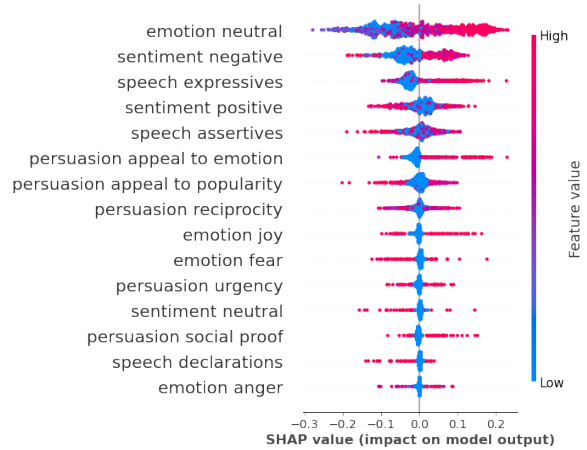


Figure 7: Impact of linguistic features on predicting high emoji reaction engagement.

that both negative emotional content and muted positivity may drive greater engagement, possibly because they elicit stronger social responses or invite interpretation. These findings align with prior research on negativity bias in online platforms, which shows that negatively valenced messages tend to attract more interaction on social platforms (Schöne et al., 2023; Watson et al., 2024). Our novel results indicate that this dynamic extends to emoji reactions as well.

Linguistic categories also contribute. For example, Speech Expressives and Speech Assertives are among the most impactful predictors, suggesting that stylistic elements of the message may influence emotional engagement beyond what is said. Several persuasion strategies such as Appeal to Emotion, Appeal to Popularity, and Reciprocity also emerge as relevant predictors. In contrast, fine-grained emotion labels such as joy, anger, or fear showed comparatively modest influence.

6 Discussion and Conclusions

Our analysis reveals that emoji reactions on Telegram do not reliably act as emotional mirroring and may instead signal social support. While previous studies have largely examined emojis as elements embedded within text, demonstrating that they can intensify expressed sentiment (Shiha and Ayvaz, 2017), improve the performance of sentiment analysis models (Liu et al., 2021), and convey additional affective information (Lou et al., 2024; Khan et al., 2025), this perspective has subsequently been extended, often implicitly, to emojis used as reactions as well, treating them as proxies for sentiment (Pool and Nissim, 2016; Plaza Del Arco et al.,

2018; Raad et al., 2018).

Our findings challenge this assumption. In particular, through ablation analysis, we show that joy is the most disruptive emotion in terms of alignment. Unlike other categories, joy reactions frequently appear in response to emotionally negative or neutral messages. This suggests that joy serves a pragmatic role, signaling approval, support, or social bonding. These patterns indicate that emoji reactions on Telegram may operate more as social cues following human biases, than as emotional reflections. This urges caution against using emoji reactions as ground truth for sentiment analysis and raises concerns about AI models that rely on such signals to estimate public opinion. Our results also help explain why negative or controversial content online often receives higher engagement, as reactions may reflect participation or support for the person posting, complementing work by Morales et al. (2025), where authors highlighted positive emoji reactions in the context of toxic discourse as approval signal. Moreover, we show that this bias toward positive reactions holds across a wider range of communicative intents and emotional tones. Finally, we show that stronger negative sentiment and muted positivity are linked to higher emoji reaction rates, extending the negativity bias observed in online platforms (Schöne et al., 2023; Watson et al., 2024) to emoji-based interactions. Overall, since our analysis focuses on crypto-related messages, a domain where emoji reactions are highly visible and central to community interaction and where in-text emojis are known to shape financial sentiment (Mahrous et al., 2023; Zuo et al., 2024), our findings shed light on a largely unexplored aspect of communication, namely the use of emoji reactions, while also highlighting the need to assess whether these patterns generalize beyond this context.

In conclusion, we provide insights into the social meaning of emoji reactions on Telegram, suggesting that they serve a distinct communicative purpose compared to in-text emojis. Recognizing this distinction is especially important for research on online engagement and misinformation, not only in financial discourse but also in political (Tretakov et al., 2025) and infodemic (Calamusa et al., 2020; Ng and Loke, 2020) contexts, as well as in moderation settings (Zhou et al., 2025; Cima et al., 2025; Tessa et al., 2025), where emoji reactions may be used as signals or indicators of emotional resonance or public sentiment.

7 Limitations

First, our findings rely on automated classifiers, including transformer-based models (e.g., BERT, RoBERTa, DistilBART) and a large language model (GPT-4o), which were not specifically fine-tuned or validated on our dataset. As a result, their predictions may introduce a degree of uncertainty and noise typical of automated annotation pipelines, particularly for nuanced phenomena such as emotion, persuasion, and speech acts. Second, as our dataset consists mostly of public cryptocurrency-related Telegram channels, the communicative dynamics we observe may reflect domain-specific norms, which limits the generalizability of our findings and highlights the need to examine other domains in future work. Doing so would also clarify our interpretation of the positivity bias: although we view it as consistent with social-approval, engagement, or banter signaling, other explanations remain possible, including community norms specific to this domain that may encourage more positive reactions. Nonetheless, evidence that emoji reactions do not reliably reflect emotional resonance, at least in this setting, remains informative. At the same time, our analytical approach is general and can be applied to other domains and platforms. Beyond these interpretative factors, the Telegram reaction system itself introduces structural constraints. The set of standard emoji reactions is finite, and our study does not consider custom emoji sets. In addition, not all channels enable reactions, and when they do, administrators may restrict the available options to a small and often predominantly positive set. These features can introduce selection bias and promote approval-oriented responses by design. Future work could extend these insights by examining how reaction emojis operate within a broader range of digital communication practices.

Acknowledgements

This work was partly supported by SoBigData.it which receives funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: “SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics” – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.; and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU.

References

- Lorenzo Alvisi, Victoria Popa, Guglielmo Cola, Serena Tardelli, and Maurizio Tesconi. 2025a. From Toxicity to Conformity: Adaptive user behavior to social norms in Telegram communities. *arXiv preprint arXiv:2511.17333*.
- Lorenzo Alvisi, Serena Tardelli, and Maurizio Tesconi. 2024. Unraveling the Italian and English Telegram Conspiracy Spheres through Message Forwarding. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 204–213. Springer.
- Lorenzo Alvisi, Serena Tardelli, and Maurizio Tesconi. 2025b. Mapping the Italian Telegram Ecosystem: Communities, Toxicity, and Hate Speech. *arXiv preprint arXiv:2504.19594*.
- William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.
- Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.
- Andrea Calamusa, Serena Tardelli, Marco Avvenuti, Stefano Cresci, Iliaria Federigi, Maurizio Tesconi, Marco Verani, and Annalaura Carducci. 2020. Twitter monitoring evidence of Covid-19 infodemic in Italy. *European Journal of Public Health*, 30(Supplement_5):ckaa165–066.
- CardiffNLP. 2021. twitter-roberta-base-sentiment. <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>. Accessed: July 22, 2025.
- V. D. Chaithra and 1 others. 2019. Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments. *International Journal of Electrical and Computer Engineering*, 9(5):4034–4041.
- Emil Chifu, Roxana Hatos, Monica Borda, and Alexandra Terec. 2015. Web harvesting and sentiment analysis of consumer feedback. *Acta Technica Napocensis: Electronics and Telecommunications*, 56(3):27–32.
- Lorenzo Cima, Benedetta Tessa, Amaury Trujillo, Stefano Cresci, and Marco Avvenuti. 2025. Investigating the heterogeneous effects of a massive content moderation intervention via Difference-in-Differences. *Online Social Networks and Media*, 48:100320.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.
- Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. 2024. The anatomy of conspiracy theorists: Unveiling traits using a comprehensive Twitter dataset. *Computer Communications*, 217:25–40.
- Lisa Graziani, Stefano Melacci, and Marco Gori. 2019. Jointly Learning to Detect Emotions and Predict Facebook Reactions. In *International Conference on Artificial Neural Networks*, pages 185–197. Springer.
- Md Abrar Jahin, Md Sakib Hossain Shovon, MF Mridha, Md Rashedul Islam, and Yutaka Watanobe. 2024. A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. *Scientific Reports*, 14(1):24882.
- Olusegun Oladele Jegede. 2024. Speech Acts in the Digital Sphere: A Corpus-Based Exploration of Interactional Dynamics on Social Media Platforms. *Corpus-based Studies across Humanities*, 2(2):359–383.
- Amit Khan, Dipankar Majumdar, and Bikromaditya Mondal. 2025. Sentiment analysis of emoji fused reviews using machine learning and Bert. *Scientific Reports*, 15(1):7538.
- Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS one*, 10(12):e0144296.
- Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.
- Yinxia Lou, Junxiang Zhou, Jun Zhou, Donghong Ji, and Qing Zhang. 2024. Emoji multimodal microblog sentiment analysis based on mutual attention mechanism. *Scientific Reports*, 14(1):29314.
- Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems*, 30.
- Ahmed Mahrous, Jens Schneider, and Roberto Di Pietro. 2023. The Role of Emojis in Sentiment Analysis of Financial Microblogs. In *2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 76–84. IEEE.
- Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. “Blissfully Happy” or “Ready to Fight”: Varying Interpretations of Emoji. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):259–268.
- Esteban Morales, Jaigris Hodson, Anatoliy Gruzd, and Philip Mai. 2025. I [thumbs-up] your Hate: Emojis as Infrastructural Platform Violence on Telegram. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2316–2325. University of Hawai’i Press.

- Lynnette Hui Xian Ng and Jia Yuan Loke. 2020. Analyzing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat. *IEEE Internet Computing*, 25(2):84–91.
- John C Paolillo. 2023. The awkward semantics of Facebook reactions. *First Monday*.
- F Plaza Del Arco, S Jiménez-Zafra, María-Teresa Martín-Valdivia, and L Alfonso Urena-López. 2018. Using Facebook Reactions to Recognize Emotion in Political Domain. In *Proceedings of CAEPIA*.
- Robert Plutchik. 1994. *The psychology and biology of emotion*. HarperCollins College Publishers.
- Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39.
- Fatima Zahra Qachfar and Rakesh Verma. 2023. ReDASPersuasion at SemEval-2023 Task 3: Persuasion Detection using Multilingual Transformers and Language Agnostic Features. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2124–2132.
- Bin Tareaf Raad, Berger Philipp, Hennig Patrick, and Meinel Christoph. 2018. ASEDS: Towards Automatic Social Emotion Detection System Using Facebook Reactions. In *2018 IEEE 20th international conference on high performance computing and communications; IEEE 16th international conference on Smart City; IEEE 4th international conference on data science and systems (HPCC/SmartCity/DSS)*, pages 860–866. IEEE.
- Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. A Transformer based Approach for Identification of Tweet Acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Bhadresh Savani. 2021. bert-base-uncased-emotion. <https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion>. Accessed: July 22, 2025.
- Jonas P Schöne, David Garcia, Brian Parkinson, and Amit Goldenberg. 2023. Negative expressions are shared more on Twitter for public figures than for ordinary users. *PNAS nexus*, 2(7):pgad219.
- John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.
- Mohammed Shiha and Serkan Ayvaz. 2017. The Effects of Emoji in Sentiment Analysis. *Int. J. Comput. Electr. Eng.(IJCEE.)*, 9(1):360–369.
- Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2025. Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention. *Engineering Applications of Artificial Intelligence*, 162:112375.
- Arsenii Tretiakov, Sergio D’Antonio-Maceiras, Áurea Anguera de Sojo Hernández, and Alejandro Martín. 2025. Toxic Discourse in the Digital Battlefield: Analysing Telegram Channels During the Russia-Ukraine ‘Conflict’. *Expert Systems*, 42(7):e70081.
- Aleksandra Urman and Stefan Katz. 2022. What they do in the shadows: Examining the far-right networks on Telegram. *Information, communication & society*, 25(7):904–923.
- Valhalla. 2020. distilbart-mnli-12-1. <https://huggingface.co/valhalla/distilbart-mnli-12-1>. Accessed: July 22, 2025.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- Dong Wang, Tao Xiao, Teyon Son, Raula Gaikovina Kula, Takashi Ishio, Yasutaka Kamei, and Kenichi Matsumoto. 2023. More Than React: Investigating The Role of Emoji Reaction in GitHub Pull Requests. *Empirical Software Engineering*, 28(5):123.
- Joe Watson, Sander van der Linden, Michael Watson, and David Stillwell. 2024. Negative online news articles are shared more to social media. *Scientific Reports*, 14(1):21592.
- Yuhang Zhou, Yimin Xiao, Wei Ai, and Ge Gao. 2025. The Hidden Language of Harm: Examining the Role of Emojis in Harmful Online Communication and Content Moderation. *arXiv preprint arXiv:2506.00583*.
- Xiaorui Zuo, Yao-Tsung Chen, and Wolfgang Karl Härdle. 2024. Emoji driven crypto assets market reactions. *Management & Marketing*, 19(2).

Quantifying Social Sentiment in Hostels Using A Domain-Specific Transformer Pipeline

Ian W. McMurry

Georgia Institute of Technology

Atlanta, GA, United States

imcmurry3@gatech.edu

Abstract

This paper presents a domain-specific transformer pipeline for quantifying social atmosphere in hostel reviews, an experiential dimension that travelers consistently prioritize but that existing NLP methods and booking platforms fail to capture. We train a cross-encoder on 4,994 manually annotated reviews and use it to pseudo-label 162,840 additional reviews; these labels are then distilled into a sentence-transformer bi-encoder, producing embeddings where proximity reflects social interaction level rather than generic sentiment. On held-out human-labeled data, the domain-adapted embeddings achieve $F1 = 0.826$, outperforming generic sentence embeddings (0.671) and zero-shot GPT-4o (0.774), with a 40-fold improvement in intra-class versus inter-class similarity. Aggregating predictions to the property level reveals that hostel socialness follows an approximate exponential distribution, confirming that highly social hostels are rare. This work formalizes socialness as a measurable semantic construct and provides a general template for extracting implicit experiential attributes from text at scale.

1 Introduction

For travelers staying in hostels, the social atmosphere, that is, how lively, friendly, and communal a property feels, is often more important than traditional quality indicators such as cleanliness, location, or amenities. Prior research in tourism and backpacking culture has consistently shown that younger and solo travelers select hostels specifically to meet other travelers, form temporary communities, and engage in shared social experiences (Oliveira-Brochado and Gameiro, 2013). Industry data similarly underscores this priority: over half of solo travelers report that “meeting interesting people along the way” is the single most memorable aspect of a trip, ranking above destination or attractions (Hostelworld, 2024, 2025). Despite this

clear behavioral pattern, the social dimension of hostels remains poorly captured in both academic hospitality research and commercial booking platforms.

Traveler reviews present a promising yet underutilized source of information about social atmosphere. These reviews often contain rich descriptions of friendliness, group dynamics, shared meals, pub crawls, communal areas, and the ease or difficulty of meeting other guests. However, the signals are embedded in unstructured, subjective language that traditional NLP tools fail to extract meaningfully. General-purpose sentiment analysis models primarily detect positive versus negative valence, overlooking experiential or social nuance. Topic-modeling studies in hospitality typically emphasize concrete attributes such as cleanliness, facilities, location, or price (Li et al., 2015; Krishnan et al., 2024), and existing aspect-based sentiment models for hotels largely exclude social interaction as a defined category. Prior NLP work in the tourism domain repeatedly notes the challenge of modeling experiential, intangible, and socially constructed attributes, which often require domain-specific cues and contextual understanding (Xiang et al., 2015).

Even platforms explicitly designed for socially oriented travelers do not provide a direct measurement of social atmosphere. Hostelworld, for example, includes an “atmosphere” score, but this metric is likely degraded through expectation alignment. According to expectation–disconfirmation theory (EDT) (Oliver, 1980; Pizam and Milman, 1993), satisfaction ratings primarily measure whether a guest’s expectations were met, not what the true underlying attribute is. A tranquil hostel can receive a high atmosphere rating if guests expected calmness and it delivered; a party hostel can receive the same high score when it meets expectations of social intensity. Because both extremes are rewarded equally, the rating cannot distinguish high-social from low-social hostels.

Due to the complexity of extracting social cues from unstructured reviews and the expectation-alignment bias embedded in atmosphere ratings under EDT, neither the research community nor the tourism industry currently possesses an operational or data-driven metric for measuring the true social interaction level of a hostel. Consequently, travelers cannot reliably compare hostels by social atmosphere, and operators lack tools to monitor or improve the communal experience that many guests value most.

To address this gap, we develop the first domain-specific transformer model designed to identify social behavior and communal atmosphere in hostel reviews. Our approach begins by training a high-accuracy cross-encoder to detect socialness cues in a curated seed dataset, enabling large-scale pseudo-labeling of more than 160k hostel reviews. Pseudo-labeling has a long history as an effective semi-supervised learning technique for low-density separation and improving classifier confidence (Chapelle et al., 2009; Lee et al., 2013). We use the term *pseudo-labeling* to emphasize that these automatically assigned labels are treated as supervision and used only for downstream representation learning rather than as ground truth annotations.

We then fine-tune a sentence-transformer bi-encoder on the pseudo-labeled corpus to create a vector space where semantic proximity reflects social interaction levels rather than generic sentiment or topic similarity. This mirrors successful strategies in other specialized NLP domains, such as BioBERT for biomedical literature (Lee et al., 2020) and FinBERT for financial communication (Huang et al., 2023), which demonstrate that pre-trained models often need to be adapted to domain-specific corpora to capture specialized meaning.

Together, the cross-encoder supervision and domain-adapted bi-encoder directly address the two core challenges outlined above. First, the cross-encoder enables precise identification of social cues that generic sentiment and topic models fail to capture, allowing the model to learn from examples where social interaction is subtly or implicitly expressed. Second, the fine-tuned sentence-embedding space overcomes the inadequacy of expectation-aligned ratings by providing a representation that reflects the actual semantic content of reviews rather than guests’ prior expectations. In combination, this pipeline produces embeddings that encode the underlying social vi-

brancy of hostels, something neither raw ratings nor out-of-the-box NLP models can isolate.

This work contributes four advances:

1. A new annotated and pseudo-labeled dataset capturing social interaction themes in hostel reviews, the first collection focused explicitly on socialness rather than sentiment, amenities, or destination attributes.
2. A formal introduction of “socialness” as a measurable semantic construct within NLP for tourism, distinct from sentiment polarity, topic categories, or expectation-aligned ratings.
3. A domain-specific transformer embedding that captures hostel social atmosphere more accurately than general-purpose models, enabling fine-grained discrimination of communal versus non-communal environments.
4. Applied empirical evaluation, showing that the resulting embeddings support downstream tasks such as hostel ranking, clustering, retrieval, and social-atmosphere mapping at global scale.

By quantifying an attribute that travelers consistently value but that platforms cannot meaningfully measure, this work fills a methodological and practical gap at the intersection of natural language processing, tourism analytics, and experiential hospitality research. Beyond academic contributions, the resulting domain-specific embeddings enable a wide range of real-world applications: social-based hostel ranking, similarity search (for example, “find hostels with a similar vibe”), clustering of hostels into social typologies, temporal monitoring of a property’s social atmosphere, and operator-facing analytics that help hostel managers understand, benchmark, and improve the communal guest experience. These downstream uses demonstrate both the feasibility and the immediate utility of modeling socialness as a quantifiable semantic dimension.

2 Related Work

2.1 Tourism NLP and the Absence of Social Atmosphere Modeling

Tourism and hospitality research has extensively mined online reviews, but this work consistently focuses on amenity-based and operational hotel

attributes rather than experiential or social atmosphere. For example, Li et al. (Li et al., 2015) analyze 118,000 TripAdvisor hotel reviews using emerging pattern mining and find that the dominant extracted features are location, price, room, service, staff, breakfast, and cleanliness, all of which are concrete physical characteristics of the property. Their automatically generated feature list such as room, staff, breakfast, cleanliness, lounge, and pool contains no representations of guest-to-guest interaction or communal social dynamics. Similar patterns appear in alternative accommodations. Krishnan et al. (Krishnan et al., 2024) apply sentiment analysis and topic modeling to homestay reviews and identify topics related to rooms, facilities, bathrooms, cleanliness, location, and host hospitality, but again no indicators of the vibe, communal energy, or interaction among guests that characterize social atmosphere. Across both hotels and homestays, existing tourism NLP reliably extracts tangible service attributes while entirely omitting interpersonal or community-oriented experiences, leaving the social dimension of accommodations computationally unmodeled.

2.2 Subjectivity and Experiential Meaning in NLP

A substantial body of work shows that many subjective or experiential attributes in text are conveyed implicitly rather than through explicit lexical markers. Kim and Klinger (Kim and Klinger, 2018) demonstrate that emotions are often expressed by describing situations rather than affective words, noting that implicit emotion requires reasoning over events and commonsense knowledge rather than surface sentiment cues. This finding highlights a general limitation of polarity-based or keyword-driven approaches when meaning is embedded in context.

Recent work on implicit semantics extends this challenge to modern embedding models. Sun et al. (Sun et al., 2025) show that widely used text embeddings prioritize surface similarity and topical overlap while performing poorly on tasks involving implicature, speaker stance, or social meaning. Their experiments indicate that even state-of-the-art models capture only a small fraction of pragmatic or attitudinal content.

These observations indicate that experiential constructs often require specialized modeling to infer meaning that is not explicitly stated. Social atmosphere in hostel reviews exhibits similar character-

istics. Descriptions of social interaction, friendliness, or communal energy are frequently implied through situational cues rather than explicit terminology. This motivates the need for a domain-specific modeling approach capable of capturing these implicit social signals

2.3 Domain-Specific Transformer Models

Large pretrained language models often require domain-specific adaptation when applied to specialized corpora, since general-purpose models do not fully capture the terminology, semantic patterns, or contextual cues present in domain-specific text. BioBERT (Lee et al., 2020) demonstrates this clearly in the biomedical domain. By continuing BERT pretraining on PubMed abstracts and PMC full-text articles, BioBERT achieves substantial performance gains across biomedical named entity recognition, relation extraction, and question answering tasks, as demonstrated by substantial gains across biomedical NER, relation extraction, and question answering. FinBERT (Huang et al., 2023) applies the same principle to financial communication, where nuanced sentiment and domain-specific vocabulary lead general pretrained models to underperform. The authors show that pretrained transformers adapted to financial filings and earnings reports provide significantly more accurate sentiment extraction, outperforming general BERT on financial-specific evaluation tasks. Together, these models show that domain-adapted transformers consistently outperform general-purpose variants when the target domain contains specialized terminology or subtle contextual cues. This motivates our use of a domain-specific sentence-transformer trained on hostel reviews, where social interaction cues are similarly domain dependent and not well captured by general pretrained models.

3 Methodology

Our goal is to learn a representation of hostel reviews in which distances reflect the underlying social interaction level rather than generic sentiment or topic similarity. The pipeline consists of (1) defining socialness and constructing a labeled seed set, (2) training a cross-encoder classifier, (3) pseudo-labeling the full corpus (4) fine-tuning a domain-specific bi-encoder on these labels, (5) calibrating a linear classifier on human-annotated data, and (6) aggregating review-level predictions into hostel-level socialness scores.

3.1 Defining Socialness and Seed Labels

3.1.1 Construct Definition

Socialness refers to the extent to which a review describes guest-to-guest interaction, communal energy, or situational cues indicating that travelers are likely to meet, engage, or form temporary communities. Unlike amenities (e.g., cleanliness, facilities) or evaluative constructs such as sentiment, socialness captures the interpersonal dynamics of the property as experienced by guests. Socialness and overall sentiment are independent dimensions: a review may be strongly positive or negative in tone without being social at all, and a highly social environment may be described unfavorably.

Linguistically, socialness is expressed primarily through implicit situational descriptions rather than explicit labels such as “this is a social hostel.” Reviews rarely state socialness directly. Instead, social cues appear through descriptions of activities, shared spaces, and interaction patterns among guests. For example, a review describing “parties every night” or “something happening every day like drinking games and going to the clubs” conveys a highly social environment even when accompanied by complaints about noise or cleanliness. These cases illustrate that high socialness can co-occur with negative sentiment.

Conversely, not all mentions of party-related vocabulary or positive affect indicate genuine social interaction. A reviewer may note that “while it is a party hostel, I found it impersonal and hard to meet people,” indicating the presence of a party-oriented setting without meaningful guest-to-guest engagement. Similarly, reviews emphasizing “very friendly staff” or a welcoming host environment provide no evidence of interactions among guests and are therefore considered non-social. These distinctions motivate grounding the construct in observed interpersonal dynamics rather than topical keywords, facilities, or affective polarity.

Formally, we view socialness as a continuous experiential dimension. However, for supervision at the review level, we operationalize it as a binary label: $\text{social} = 1$ if the review provides clear textual evidence of guest interaction or communal participation, and $\text{social} = 0$ otherwise. This binary operationalization prioritizes annotation consistency and high-precision supervision. The underlying continuity of socialness emerges downstream through aggregation across reviews and through the geometry of the learned embedding space, rather

than from individual review labels.

The full annotation rubric, including decision rules and a labeling flowchart, is provided in Appendix A.

3.1.2 Seed Dataset and Annotation

The cross-encoder (Section 3.3.1) is trained on a manually annotated seed dataset of 4,994 reviews sampled from a relational database of hostel reviews. Because explicitly social reviews are relatively rare in purely random samples, we use keyword-guided enrichment as a sampling strategy only to increase the prevalence of candidate socialness edge cases during annotation. Concretely, we draw a mixed pool consisting of (i) reviews containing social-event related strings (e.g., “pub crawl”, “party hostel”), (ii) reviews containing calm or low-sociality strings (e.g., “quiet”, “not a party hostel”), and (iii) randomly sampled reviews. Full list of strings that were used to draw out the enrichment can be seen in Appendix B. To avoid collapsing onto narrow lexical triggers, these keyword pools are used solely to select reviews for annotation, and final labels are assigned based on full-text review content using the rubric in Appendix A rather than keyword presence. In practice, enriched candidates constituted a minority of the seed dataset, with the remainder drawn from random sampling to preserve coverage of the natural review distribution.

All reviews in the seed dataset were labeled manually by a single annotator with domain familiarity, prioritizing internal consistency in applying the socialness rubric. Ambiguous cases were resolved conservatively as non-social to maintain high precision in the positive class. This annotated seed set provides the sole human supervision used to train the cross-encoder model.

3.2 Dataset and Preprocessing

The full corpus consists of 162,840 hostel reviews from two major booking platforms, covering 2,230 distinct hostels worldwide. Reviews were programmatically collected from publicly accessible online review pages and stored in a relational database prior to preprocessing. The texts are short but information-dense, with most reviews well within standard transformer context windows, making them suitable for sentence-level modeling.

Because hostel travel is highly international, the corpus is multilingual. We first apply automatic language identification and then translate all non-English reviews into English using the Google

Cloud Translation API. The detected language code is stored alongside the original text. Reviews identified as English are kept verbatim; others are translated in batched requests. If translation fails, the original text is retained. Manual inspection of a random subset confirms that translations remain fluent and preserve the situational cues relevant for socialness.

Preprocessing is intentionally minimal. Aggressive normalization could remove precisely the cues needed to detect guest interaction. We therefore preserve punctuation, capitalization, emojis, and nonstandard spellings. The only substantive cleaning step concerns reviews that are scraped into multiple fields (e.g., separate positive and negative comments); these are concatenated into a single free-text field. Empty or null reviews are discarded. Otherwise, text is kept exactly as written by travelers.

3.3 Modeling Pipeline

3.3.1 Cross-Encoder Socialness Classifier

The first stage of the pipeline is a high-precision cross-encoder that maps each review to the probability that it describes a socially active environment. We fine-tune the publicly available cross-encoder/ms-marco-MiniLM-L-12-v2 model from Sentence-Transformers, originally trained for passage–query relevance. The model receives as input a pair consisting of (1) the review text and (2) a fixed natural-language query that specifies the target construct, such as:

“Does this review describe a lively social environment with friendly guests?”

The cross-encoder jointly encodes the concatenated review and query and outputs a scalar probability via a sigmoid layer. We train the model with binary cross-entropy loss on the annotated seed set, using standard hyperparameters for learning rate, batch size, and number of epochs. Because social reviews are less frequent than non-social ones, we modestly upsample the positive class during training to expose the model to a wider variety of social cues. To assess generalization across properties, we use both a stratified row-level split and a hostel-level split in which all reviews from held-out hostels appear only in validation or test sets.

3.3.2 Pseudo-Labeling and High-Confidence Subset

After training, we apply the cross-encoder to the entire translated corpus to obtain, for each review x , a probability $p_{\text{CE}}(y = 1 | x)$ that the review is social. We then derive a binary pseudo-label $\hat{y}_{\text{CE}} \in \{0, 1\}$ for every review using a fixed decision threshold selected on the validation set.

To reduce noise in downstream representation learning, pseudo-labeling pipelines often apply confidence-based filtering to discard ambiguous cases near the decision boundary. In our setting, however, the cross-encoder produces a sharply bimodal probability distribution, with relatively few reviews assigned intermediate probabilities. We therefore retain the full pseudo-labeled corpus for bi-encoder training, as explicit high-confidence thresholding would remove only a small number of borderline instances and does not materially affect downstream performance. This approach yields a large, automatically labeled dataset while preserving the diversity of social and non-social signals present in the corpus.

3.3.3 Domain-Specific Bi-Encoder Training

Although the cross-encoder yields high-quality predictions, it is computationally costly because each inference requires full joint attention over the review and query. To support large-scale analysis (e.g., unsupervised clustering, embedding-space exploration) and to enable low-latency inference for retrieval or recommendation settings, we distill the cross-encoder into a lightweight sentence-transformer bi-encoder that produces socialness-aware embeddings.

We start from the sentence-transformers/all-MiniLM-L6-v2 model, which maps each review x independently to a 384-dimensional embedding $f(x)$. Training is performed on the pseudo-labeled reviews. We first create train/validation/test splits with respect to \hat{y}_{CE} , optionally subsampling to maintain a balanced and tractable training set.

Rather than conditioning on a query, the bi-encoder is trained with a pairwise similarity objective over review–review pairs. Positive pairs are constructed from reviews that share the same pseudo-label (both social or both non-social), and negative pairs from reviews with different pseudo-labels. The model is optimized with a cosine-similarity loss that encourages embeddings of same-label reviews to be close and embeddings

of different-label reviews to be far apart. We use standard settings for batch size, learning rate, and number of epochs. The best checkpoint is selected based on validation performance on the pseudo-label classification task.

3.3.4 Linear Classifier Head and Continuous Scores

To align the distilled representation directly with human annotation and to obtain a lightweight classifier for downstream use, we freeze the fine-tuned bi-encoder and train a logistic regression head on top of its embeddings. Each of the 4,994 human-labeled reviews is embedded with the bi-encoder, and the resulting vectors are split into train/validation/test sets using the same proportions as for the cross-encoder. An ℓ_2 -regularized logistic regression model with class-balanced weights is then trained to predict the manual socialness label from the embedding.

In addition to binary decisions, both the logistic regression output and the underlying embedding space provide continuous socialness scores. For example, hostel-level socialness can be computed as the fraction of a hostel’s reviews predicted as social or by aggregating the continuous probabilities across its reviews. These aggregation procedures are described in the following subsection.

3.3.5 Hostel-Level Aggregation

Finally, we derive property-level socialness scores by aggregating review-level predictions. For each hostel, we collect all associated reviews and compute the fraction of reviews classified as social by the bi-encoder + logistic regression model. This fraction serves as a continuous socialness score in $[0, 1]$, where higher values indicate a higher prevalence of social interaction in the textual record of the property. Alternative aggregation schemes (e.g., averaging probabilities, weighting by recency) are straightforward but beyond the scope of this work.

4 Results

4.1 Dataset Descriptive Statistics

The full corpus consists of 162,840 hostel reviews collected from two major booking platforms: Booking.com (88,656 reviews) and Hostelworld (74,184 reviews). Reviews span 72 cities worldwide and correspond to 2,230 unique hostels, reflecting substantial geographic and property-level diversity.

The dataset is multilingual, reflecting the international nature of hostel travel. Automatic language

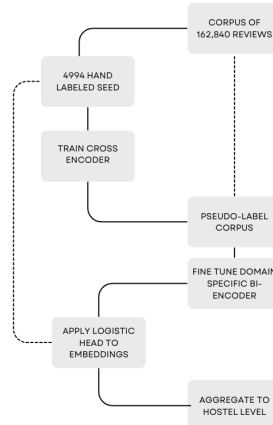


Figure 1: Methodology Workflow

identification indicates that 68.08% of reviews are written in English, with the remaining reviews distributed across more than twenty languages. Table 1 summarizes the most common languages. Because cues related to social interaction may appear across diverse linguistic contexts, all subsequent analyses operate on a translated English corpus.

Language	Percent of Corpus
English	68.08%
Spanish	9.25%
French	4.59%
Portuguese	4.06%
German	3.25%
Japanese	1.47%
Italian	1.31%
Russian	1.24%
Korean	0.91%
Dutch	0.89%

Table 1: Most common languages in the 162,840-review corpus prior to translation.

Reviews are generally short to moderate in length. The mean review length is 55.8 tokens, with a median of 43 tokens and a standard deviation of 48.6 tokens. The distribution is right-skewed, with the 90th percentile at 110 tokens, the 95th percentile at 128 tokens, and the 99th percentile at 235 tokens. Extremely long reviews are rare: only 0.05% of reviews exceed 512 tokens, and fewer than one hundred reviews exceed 1,024 tokens. This length profile is well suited to transformer-based encoders without requiring aggressive truncation.

4.2 Human-Labeled Seed Dataset

The cross-encoder is trained on a manually annotated seed set of 4,994 reviews. This dataset provides the only human supervision used in the

pipeline and therefore anchors the definition of socialness. Table 2 summarizes its key properties. Social reviews constitute 25.15% of the dataset, reflecting the relative rarity of explicit guest–guest interaction cues in hostel reviews. Review lengths are short (median 50 tokens), and the dataset spans both major platforms, ensuring coverage of diverse writing styles and traveler perspectives.

Statistic	Value
Total reviews	4,994
Positive class (social = 1)	25.15%
Negative class (social = 0)	74.85%
Median length	50 tokens
95th percentile length	146 tokens
Platforms represented	Hostelworld & Booking.com

Table 2: Summary statistics of the human-labeled seed dataset used to train the cross-encoder.

4.3 Evaluation Metrics

We report standard metrics for binary classification, chosen to reflect both thresholded classification performance and ranking quality:

- **F1 score** is the harmonic mean of precision and recall and provides a single summary measure of classification performance at a fixed decision threshold. Because we care equally about avoiding false positives and false negatives, F1 serves as the primary metric for comparing models in this setting.
- **AUC** (area under the ROC curve) measures ranking quality independent of any particular threshold.
- **Average Precision (AP)** summarizes the precision–recall curve by weighting precision at different recall levels.

4.4 Cross-Encoder Performance

We first evaluate the cross-encoder on the 4,994 human-labeled reviews. Table 3 reports performance under two evaluation regimes: (a) a standard row-level split and (b) a stricter hostel-level split that tests generalization to unseen properties. The model achieves high discrimination ($AUC > 0.97$) and ranking quality ($AP = 0.954$) in both settings. The slightly higher F1 on the hostel-level split indicates that the classifier captures generalizable cues of social interaction beyond property-specific phrasing.

Split	AP	AUC	F1	Thresh.
Row-level	0.9540	0.9819	0.8829	0.0733
Hostel-level	0.9540	0.9797	0.8916	0.0972

Table 3: Cross-encoder performance on human-labeled data.

4.4.1 Pseudo-Labeling the Full Corpus

Applying the cross-encoder to all 162,840 reviews yields continuous socialness probabilities and pseudo-labels. The model predicts that 20.34% of reviews describe socially active environments, representing a downward shift from the hand-labeled seed set. This difference is expected because the seed dataset was constructed using a sampling enrichment procedure that increased the prevalence of candidate socialness edge cases (in addition to random reviews) during annotation, whereas the full corpus reflects the natural distribution of review content. Figure 2 shows the resulting probability distribution, which is sharply bimodal: most reviews lie near 0 or near 1.

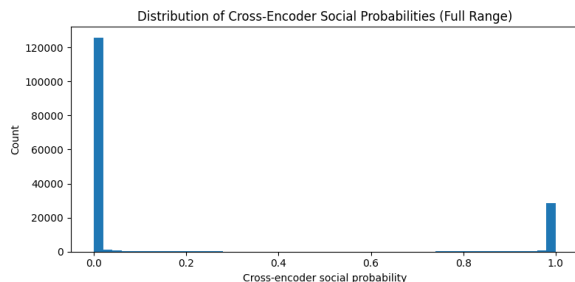


Figure 2: Cross-encoder probability distribution for all reviews.

The cross-encoder probability distribution is sharply bimodal, with most reviews assigned probabilities near 0 or 1. As a result, we do not apply explicit high-confidence filtering during pseudo-labeling, since removing borderline cases would discard only a small fraction of the data and does not materially affect downstream performance.

Platform differences. Hostelworld reviews exhibit a substantially higher proportion of social-labeled content (27.87%) compared to Booking.com reviews (14.04%). This pattern aligns with differences in platform focus: Hostelworld targets backpackers and hostel-goers, for whom social interaction is often a primary motivation for accommodation choice, while Booking.com serves a broader traveler population. The result provides correlational evidence that social atmosphere is a

meaningful and behaviorally relevant dimension in hostel travel.

4.5 Bi-Encoder Distillation and Classification

4.5.1 Classification Performance

We embed each review with the fine-tuned bi-encoder and train a logistic regression classifier on the human-labeled dataset. Table 4 compares performance against two baselines: (1) generic sentence-transformer embeddings produced by `sentence-transformers/all-mpnet-base-v2` without domain adaptation, and (2) a zero-shot classification prompt to GPT-4o-mini. Domain-adapted embeddings yield the strongest overall performance, improving F1 by 15.5 points and AUC by 8 points over the generic embedding baseline.

Model	F1	Prec.	Rec.	AUC
Generic embeddings + LogReg	0.671	0.594	0.772	0.872
GPT-4o-mini (zero-shot)	0.774	0.983	0.638	0.817
Fine tuned Bi-encoder + LogReg	0.826	0.790	0.865	0.952

Table 4: Classification performance on the human-labeled test set.

4.5.2 Embedding Structure

To assess whether socialness forms a coherent dimension in embedding space, we project embeddings of the 4,994 labeled reviews using UMAP. As shown in Figure 3, generic embeddings exhibit substantial overlap between classes, whereas the domain-specific bi-encoder produces a clear separation between social and non-social reviews.

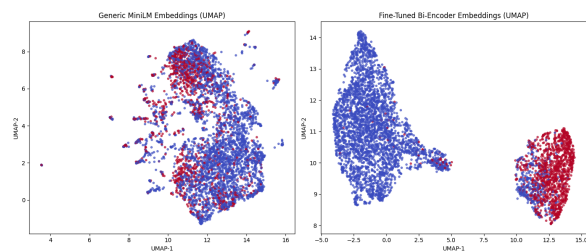


Figure 3: UMAP projection of generic embeddings (left) and domain-specific embeddings (right).

Pairwise cosine similarities confirm this structure: the intra–inter class similarity gap increases from 0.018 (generic) to 0.704 (fine-tuned), a 40-fold improvement. Unsupervised k -means clustering (with $k = 2$) achieves purity of 0.909 on the bi-encoder embeddings, compared to 0.748 for the generic baseline, indicating strong alignment between the learned representation and the socialness construct.

4.6 Hostel-Level Socialness Distribution

For each hostel, we aggregate review-level predictions to compute the fraction of its reviews classified as social, yielding a continuous property-level socialness score in $[0, 1]$. Prior to distribution fitting, we remove hostels with socialness = 0 (i.e., no reviews predicted as social) and drop any invalid or missing values, resulting in $n = 2,072$ hostels retained out of 2,230.

Figure 4 shows the empirical distribution of these nonzero hostel-level scores. Fitting a one-parameter exponential model via maximum likelihood yields $\hat{\lambda} = 5.39$ (mean = 0.186). The exponential provides a strong descriptive approximation in the upper tail: the empirical survival curve is close to linear on a log scale ($R^2 = 0.9748$). However, formal goodness-of-fit tests reject the exponential assumption (KS: $p = 3.8 \times 10^{-10}$; χ^2 : $p = 4.2 \times 10^{-4}$), indicating statistically detectable deviations from an exact exponential model.

Despite this rejection, the near-linear log-survival behavior suggests a clear rarity structure: highly social hostels are uncommon, while most properties exhibit limited or intermittent social cues. We therefore interpret the exponential fit as a compact summary of the decay pattern rather than a strictly correct generative model.

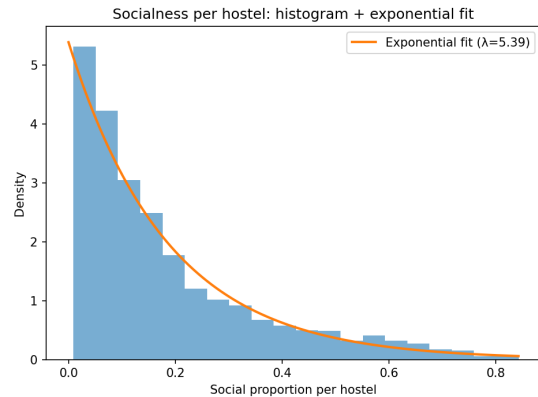


Figure 4: Distribution of hostel-level socialness scores with exponential fit.

5 Conclusion

This work introduces a data-driven approach for quantifying the social atmosphere of hostels from unstructured traveler reviews. By formalizing *socialness* as a semantic construct distinct from sentiment, amenities, and expectation-aligned ratings, we address a long-standing gap in both tourism

research and commercial travel platforms: the absence of an operational metric for guest-to-guest interaction and communal experience.

Empirically, we show that socialness can be reliably inferred from text using a domain-specific modeling pipeline. A cross-encoder trained on a small, carefully annotated seed set generalizes well to unseen hostels and produces sharply bimodal probability distributions when applied at scale. Distilling these judgments into a sentence-transformer bi-encoder yields an embedding space in which socialness emerges as a coherent latent dimension, enabling both accurate classification and unsupervised structure discovery. The resulting representations substantially outperform generic embeddings and zero-shot prompted language models, demonstrating the value of domain adaptation for capturing implicit experiential meaning.

At the property level, aggregating review-level predictions reveals a striking and well-behaved distribution: hostel socialness follows an exponential decay, with highly social properties being rare and most hostels exhibiting limited or intermittent social cues. This structure provides independent validation that the learned signal is not arbitrary, and it offers a compact statistical characterization of social atmosphere at global scale.

Beyond academic contributions, the learned embeddings enable practical applications that are not supported by existing rating systems. These include social-based hostel ranking, similarity search (e.g., discovering hostels with comparable social profiles), unsupervised clustering into social typologies, and temporal monitoring of changes in communal atmosphere. By shifting the measurement of social experience from expectation-aligned ratings to text-derived semantic structure, this work demonstrates how implicit social behavior can be operationalized at scale.

More broadly, the approach illustrates a general strategy for modeling experiential constructs that are difficult to define explicitly but are consistently described in natural language. The combination of targeted supervision, pseudo-labeling, and domain-specific representation learning provides a template for extracting other forms of social and experiential meaning from reviews, opening new directions for NLP in tourism analytics and beyond.

Limitations

While the proposed pipeline provides a coherent and scalable representation of hostel social atmosphere, several limitations remain.

Construct ambiguity and absence of ground truth. Socialness is an inherently fuzzy and context-dependent construct, with no external ground-truth benchmark against which model predictions can be definitively validated. Many reviews lie near conceptual boundaries. Examples include, party-oriented hostels with limited mingling, quiet hostels with occasional communal encounters, or reviews emphasizing staff friendliness rather than guest interaction. These ambiguities introduce unavoidable subjectivity into the labeling process and make fine-grained distinctions difficult for both humans and models. As a result, evaluation at the hostel level necessarily relies on internal consistency and distributional structure (e.g., rarity patterns) rather than comparison to a trusted external reference. This limitation reflects the broader challenge of operationalizing socially constructed experiences from text rather than a deficiency of the proposed model.

Single-annotator supervision. The seed dataset used to train the cross-encoder was annotated by a single domain-familiar annotator, prioritizing internal consistency in the initial definition of socialness. While this choice reduces inter-annotator disagreement during early development, it limits the ability to quantify annotation reliability or capture alternative interpretations of borderline cases. Future work could substantially strengthen the empirical foundation of this task by incorporating multiple annotators, measuring inter-annotator agreement, and introducing verification or adjudication layers for ambiguous reviews.

Data coverage and translation effects. The corpus is drawn exclusively from Hostelworld and Booking.com, which constrains representativeness across the broader hostel ecosystem and reflects the preferences of travelers who choose to leave reviews. In addition, approximately one-third of the corpus relies on machine translation. Although manual inspection suggests that translations preserve the situational cues most relevant to socialness, subtle shifts in tone, pragmatics, or culturally specific expressions of social interaction may affect classification. These factors introduce systematic biases that future work could address through expanded platform coverage and multilingual model-

ing.

References

- Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.
- Hostelworld. 2024. [State of solo travel report 2024](#). Accessed: 2025-12-13.
- Hostelworld. 2025. [State of solo travel report 2025](#). Accessed: 2025-12-13.
- Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.
- Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jay Krishnan, Biplab Bhattacharjee, Maheshwar Pratap, Janardan Krishna Yadav, and Moinak Maiti. 2024. [Survival strategies for family-run homestays: Analyzing user reviews through text mining](#). *Data Science and Management*, 7(3):228–237.
- Dong-Hyun Lee and 1 others. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Gang Li, Rob Law, Huy Quan Vu, Jia Rong, and Xinyuan Zhao. 2015. [Identifying emerging hotel preferences using emerging pattern mining technique](#). *Tourism Management*, 46:311–321.
- Ana Oliveira-Brochado and Carla Gameiro. 2013. [Toward a better understanding of backpackers' motivations](#). *Tékhne*, 11(2):92–99.
- Richard L. Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4):460–469.
- Abraham Pizam and Ady Milman. 1993. [Predicting satisfaction among first-time visitors to a destination by using the expectancy disconfirmation theory](#). *International Journal of Hospitality Management*, 12(2):197–209.
- Yiqun Sun, Qiang Huang, Anthony K. H. Tung, and Jun Yu. 2025. [Text embeddings should capture implicit semantics, not just surface meaning](#). *Preprint*, arXiv:2506.08354.
- Zheng Xiang, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. 2015. [What can big data and text analytics tell us about hotel guest experience and satisfaction?](#) *International Journal of Hospitality Management*, 44:120–130.

A Annotation Rubric for Socialness

This appendix documents the review-level annotation rubric used to construct the human-labeled seed dataset. The goal of the rubric is to operationalize *socialness* as evidence of guest-to-guest interaction and communal participation, while explicitly distinguishing it from sentiment polarity, staff friendliness, facilities, and topic keywords.

A.1 Task Framing and Unit of Annotation

Unit of annotation. Each *review* is annotated independently. Labels are assigned based only on the textual content of the review, without using metadata (e.g., rating, reviewer nationality, hostel name).

Label set. Socialness is annotated as a binary label:

- $\text{social} = 1$: the review provides clear evidence of guest-to-guest interaction or communal participation.
- $\text{social} = 0$: the review does not provide such evidence, or explicitly indicates low sociality.

Core principle. A review is labeled $\text{social} = 1$ only when it contains *direct or strongly implied* evidence that guests interact with other guests or participate in shared social activities. Mentions of facilities (e.g., bar, common room), staff friendliness, or general positivity are not sufficient on their own.

A.2 Decision Rules

A.2.1 Conditions for $\text{social} = 1$

A review receives $\text{social} = 1$ if it satisfies at least one of the following:

- **Explicit sociality.** The review directly states that the hostel is social or that it is easy to meet other travelers (e.g., “very social,” “easy to meet people,” “met lots of people”).
- **Guest interaction described.** The review describes guests meeting, talking, hanging out, cooking/eating together, playing games, or spending time together in shared spaces.
- **Hostel-organized activities with participation.** The review mentions naturally social group events (e.g., pub crawls, group dinners, drinking games, dance classes, or parties) (e.g., “pub crawl was super fun”).

- **Lively atmosphere with interpersonal content.** The review describes a vibrant communal environment where social interaction is clearly occurring, even if overall sentiment is negative (e.g., complaints about noise while describing nightly social activity).

A.2.2 Conditions for $\text{social} = 0$

A review receives $\text{social} = 0$ if any of the following apply:

- **Explicit statements of low sociality.** The review states that it is hard to meet people, that the hostel is not social, boring, impersonal, or lacks atmosphere.
- **Facilities mentioned without interaction.** The review mentions a bar, rooftop, common area, or events *without* describing guests interacting (e.g., “nice bar” with no interaction evidence).
- **Staff-only friendliness.** The review focuses on friendly staff/hosts or a welcoming reception but provides no evidence of guest-to-guest interaction.
- **Keyword-only or vague cues.** The review contains isolated or vague atmosphere descriptions (e.g., “good vibe”) without evidence of interaction. These cases are labeled $\text{social} = 0$ unless interpersonal participation is described.

A.3 Representative Examples

The examples below illustrate how socialness can be expressed implicitly and independently from sentiment polarity, staff friendliness, or topical keywords. Excerpts are lightly edited for brevity.

Example Type	Review Excerpt
High social, implicit (social=1)	“There are parties every night, and the music can be heard in the rooms.”
High social, negative sentiment (social=1)	“Don’t go if you want a clean place, it’s only good for partying. There is something happening every day like drinking games and going to the clubs.”
Party keywords but low interaction (social=0)	“While it is a party hostel I found it impersonal and hard to meet people.”
Low social, positive sentiment (social=0)	“Amazing hostel! Very friendly staff and they gave great advice on things to do.”

A.4 Labeling Decision Flow

Figure 5 summarizes the decision process used during annotation.

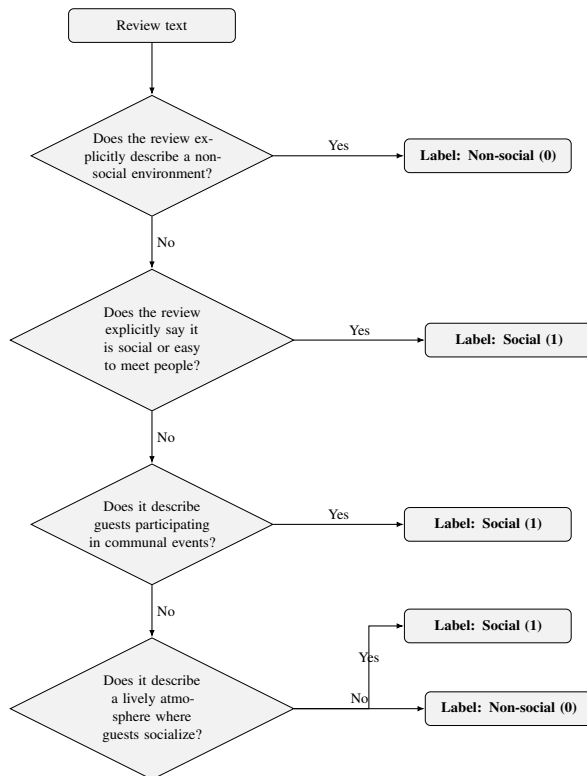


Figure 5: Decision flow for assigning binary socialness labels at the review level.

B Keyword-Guided Sampling Enrichment

This appendix documents the keyword lists used for sampling enrichment when selecting candidate reviews for annotation. Keywords were used solely to increase the probability of sampling reviews likely to contain high- or low-socialness cues and were not used as labeling rules, model features, or inference heuristics.

B.1 Party and Calm Token Lists

Party-related strings.

party hostel; pub crawl; bar crawl; club crawl;
beer pong; drinking games; DJ; nightlife; shots;
happy hour; karaoke; music until late; wild party;
parties every night;

Calm or low-sociality strings.

quiet; peaceful; calm; good sleep; sleep well; not
a party hostel; relaxing; chill; no noise; quiet
hours; silent

Predicting Convincingness in Political Speech: How Emotional Tone Shapes Persuasive Strength

Bhuvanesh Verma^{1*} Mounika Marreddy^{1*} Alexander Mehler¹

¹Goethe University, Frankfurt am Main, Germany

verma@em.uni-frankfurt.de, mmarredd@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de

Abstract

Emotional tone plays a central role in persuasion, yet its impact on computational assessments of political argument quality in real world election campaign speeches remains understudied. In this work, we investigate whether positive emotional framing correlates with higher perceived convincingness in political arguments. We fine-tune language models on argument quality datasets and test their ability to transfer convincingness predictions to real-world campaign speeches. Using a corpus of U.S. presidential campaign speeches, we analyze emotional polarity in relation to predicted persuasive strength to test whether positively framed arguments are judged more convincing than neutral or negative ones. Our empirical analysis shows that political parties rely heavily on argumentation during their election campaigns. Also, we found the evidence that politicians strategically employ emotional cues within their arguments during these campaign speeches, with positive emotions being more strongly associated with persuasive strength, for example in topics such as *USMCA's Effect on American Jobs and Agriculture*, *Border Control Policies*, *Progressive Tax Reforms*. At the same time, we find that negative emotions have a weaker yet still non-negligible influence on voter persuasion in topics such as *City Crime and Civil Unrest* and *White Supremacist Violence (Charlottesville Incident)*.

1 Introduction

Persuasion modeling estimates how convincing language is based on its linguistic and rhetorical features. Prior political psychology research shows that emotional framing significantly shapes information processing, attitude formation, and belief revision (Petty and Cacioppo, 2012; Early, 2015; Lerner et al., 2015; Rodriguez-Hernandez et al., 2024). Positive affect is linked to greater mes-

sage acceptance, trust, and credibility in persuasion (Shah, 2022; Hassell et al., 2022; Grüning and Schubert, 2022). Despite the known link between emotion and persuasion, it remains unclear whether emotional tone consistently affects the perceived convincingness of political arguments and whether modern Natural Language Processing (NLP) models capture this in large-scale political discourse. To address this gap, we hypothesize that positively framed arguments are predicted to be more convincing than neutral or negative ones.

Computational persuasion research has largely focused on modeling argument quality, stance, and rhetoric across debates, online discussions, and social media (Marreddy et al., 2025; Lippi and Torroni, 2016; Haddadan et al., 2019; Tan et al., 2016; Qiu et al., 2015). In contrast, campaign speeches, which are key to electoral messaging and identity building, remain understudied computationally. Unlike debates, which are adversarial and episodic, campaign speeches are narrative and affective, aimed at sustaining momentum and alignment across elections (Ross, 2006; Chong and Druckman, 2007). Large-scale analyses reveal systematic variation in emotional rhetoric across political actors and time (Gennaro and Ash, 2022; Aroyehun et al., 2025), yet most focus on sentiment, lexical patterns, or themes rather than persuasion outcomes (Finity et al., 2021; Mohapatra and Mohapatra, 2022). Thus, the link between emotional tone and persuasive strength in campaign speech remains underexplored.

Advances in argument mining and persuasion modeling make it possible to study such questions at scale (Lawrence and Reed, 2019; Wachsmuth et al., 2016; Druckman, 2022). Neural language models in particular have demonstrated strong performance in predicting argument convincingness and agreement across domains (Lawrence and Reed, 2019; Wachsmuth et al., 2016; Druckman, 2022). However, whether these models exhibit

*Equal contribution.

stable emotion-sensitive persuasion patterns in political speech remains unknown.

To investigate this question, we examine persuasive framing in 2020 U.S. presidential campaign speeches using [Chalkiadakis et al. \(2025\)](#)’s corpus of 1 056 Democratic and Republican speeches (Jan 2019–Jan 2021). We extract topic-aligned arguments and assess their predicted convincingness using models fine-tuned on persuasion and argument-quality datasets. We also examine how emotional polarity, especially positivity, relates to predicted persuasiveness across candidate, topic, and model scale.

We make three contributions: (1) We provide the first large-scale empirical analysis examining how emotional tone interacts with predicted persuasive strength in real-world campaign speech. (2) We evaluate whether persuasion modeling trained on benchmark datasets transfers to political speech through significance testing on downstream analysis. (3) We present a comprehensive empirical analysis showing that both parties rely heavily on argumentation in campaign speeches, with no significant differences in overall argument usage or aggregate argument quality, but with clear topic-specific asymmetries in persuasive effectiveness driven by emotional cues. In particular, we show that positive emotions are more strongly associated with persuasive strength, while negative emotions also play a non-negligible role in persuasion for certain topics.

2 Related Work

NLP has explored persuasion in political communication through argument mining, emotional rhetoric, and convincingness prediction. Early argument mining emphasized structure over persuasive impact in political communication. For example, [Lippi and Torroni \(2016\)](#) found that prosodic and lexical features improve claim detection in debates, while [Menini et al. \(2018\)](#) showed that argument mining reveals ideological contrasts in presidential speeches. Large-scale resources like the 29K-component corpus from [Haddadan et al. \(2019\)](#) enabled systematic modeling of argument structure across 50+ years of U.S. presidential campaigns.

Related literature examines emotion in political speech. It shows systematic variation in sentiment and affect across parties, candidates, audiences, and campaign contexts ([Mohapatra and Mohapatra, 2022](#); [Gennaro and Ash, 2022](#); [Aroyehun et al.,](#)

[2025](#)). However, these studies mostly focus on descriptive emotional patterns rather than how emotional tone shapes persuasive outcomes. Communication theory and psychology suggest that emotional framing affects message acceptance, credibility judgments, and belief updating ([Petty and Cacioppo, 2012](#); [Lerner et al., 2015](#); [Shah, 2022](#)). Similarly, computational studies link emotion to stance and engagement ([Benlamine et al., 2015, 2017](#)), with recent work revealing modeling biases, such as overpredicting negative emotion in argumentation ([Greschner and Klinger, 2024](#)).

A third line of work focuses on computational persuasion modeling and convincingness prediction. Early work introduced pairwise ranking frameworks to estimate persuasive strength ([Habernal and Gurevych, 2016](#)), later improved through neural architectures capable of capturing stylistic and discourse level signals ([Chowanda et al., 2017](#)). [Cano-Basave and He \(2016\)](#) investigated the argumentation styles of speakers in political debates and modeled persuasion using audience reactions such as applause and booing as proxies for emotional response. Similarly, [Quensel et al. \(2025\)](#) investigated the influence of subjective factors, including emotions, on argument strength, with a particular focus on persuasion.

Persuasion has also been modeled in online debates and stance settings, including forecasting agreement shifts and outcome prediction ([Tan et al., 2016](#); [Qiu et al., 2015](#)). Yet little work tests whether persuasion-trained models capture emotional framing, especially in campaign speech, or whether effects generalize across topics, speakers, or model scale.

Unlike this prior work, we examine how emotional tone relates to predicted persuasiveness in real-world U.S. campaign speeches and whether persuasion-trained models reflect emotion-sensitive persuasion patterns. In addition, we evaluate these effects across multiple model families and speaker contexts, allowing us to test the robustness and generalizability of emotion–persuasion interactions beyond the domains typically considered in earlier computational persuasion studies.

3 Datasets Description

We use three datasets in this study: (1) an argument detection corpus used to identify argumentative text spans, (2) an argument quality dataset used to train convincingness models, and (3) a campaign speech

corpus used for downstream analysis of persuasive framing in political communication.

3.1 Argument Detection Dataset

In this work, we adopt a broad definition of argumentation. A statement is considered argumentative if it expresses reasoning, justification, or support within a text, regardless of whether it functions as a claim or as evidence. Accordingly, both claims and evidence are treated as instances of argumentative language. To operationalize this definition, we combine two existing argument detection datasets, Ein-Dor et al. (2020) and Levy et al. (2018), to construct the ArgDetect dataset. One dataset provides claim–topic pairs, while the other consists of evidence–topic pairs; we merge these resources to create a unified argument detection corpus. Dataset statistics are reported in Appendix A and Table 2.

3.2 Argument Quality Dataset

For training a model to detect argument quality, we used EviConv by Gleize et al. (2019) which is a relative evidence ranking dataset. This dataset contains a topic claim and two evidence statements and a label. Label assigns which evidence is better among the two. Since our objective is to conduct a comparative analysis of argumentative strength between political parties, this relative formulation is a natural fit for our use case. In addition, adopting a ranking-based framework reduces sensitivity to dataset and modeling calibration issues that commonly arise when transferring absolute convincingness scores across domains (Ivanova et al., 2024). The dataset statistics are presented in the Appendix A Table 3.

3.3 Campaign Speeches Text Data (CSTD)

For our experiments, we use the Campaign Speeches Text Data (CSTD) introduced by Chalkiadakis et al. (2025) and released in Scientific Data. The corpus comprises 1,056 campaign speeches delivered by the Democratic and Republican presidential and vice-presidential candidates during the 2020 U.S. election cycle, spanning January 2019 to January 2021.

4 Methodology

Rhetorical analysis of political speeches involves identifying rhetorical components, such as arguments, and quantifying them. To this end, we first identify the latent topics present in campaign

speeches, followed by the extraction of argumentative statements and the evaluation of their persuasive quality and associated emotional strength. Therefore, we train models for argument detection and argument quality assessment and along with that topic extraction is performed using unsupervised topic modeling technique, while emotion classification is conducted using *SamLowe/roberta-base-go_emotions*¹ which was trained on the GoEmotions dataset that includes 27 distinct emotion labels. In the following subsections, we describe each methodology in detail.

4.1 Argument Detection

For argument detection, we train traditional machine learning (ML) models such as *Logistic Regression*, *Linear SVM*, and *Naive Bayes*, transformer-based models including *bert-base-uncased*, *roberta-base*, and *deberta-v3-large*, and large language models (LLMs) like *Qwen2.5-Coder-14B*, *Mistral-8B-Instruct-2410*, and *Llama-3.1-8B-Instruct*. The dataset is split into training and test sets in an 80:20 ratio for all experiments. For ML and transformer-based models, we concatenate the topic title with the sentence, separated by a space, to form each training instance. With ML-based models, we use TF-IDF and sentence embeddings (Reimers and Gurevych, 2019) (SE) features based on unigrams and bigrams (maximum 10,000 features) with stopword removal. For transformer-based modeling, we added a classification head on top of each pretrained model for sequence classification. To address the natural class imbalance of argumentative versus non-argumentative sentences, we apply a class-weighted cross-entropy loss (Phan and Yamamoto, 2020). Model training uses a learning rate of 2e-5 and a weight decay of 0.01.

For efficient utilization of LLMs, we evaluate their two-shot prompting capabilities, as zero-shot settings often struggled to produce correctly formatted labels. The prompt used for two-shot evaluation is provided in the Appendix B Section B.1. To optimize inference efficiency, we employ constrained generation (Beurer-Kellner et al., 2023), restricting the LLMs to output a single token corresponding to the label (0 or 1). During generation, we consider the probabilities of these two tokens to determine the predicted label. For evaluation, we randomly sample 200 instances from the test set, calculate

¹https://huggingface.co/SamLowe/roberta-base-go_emotions

metrics, and repeat the experiment five times to ensure stability.

4.1.1 Argument Quality Assessment

Similar to the argument detection task, we train a set of traditional machine learning (ML) models, transformer-based models, and evaluate two-shot capabilities of LLMs. We divide the training set of EviConv to create a validation split using a 9:1 ratio. For ML models, the topic and its two associated evidence sentences are concatenated into a single input text. Similar features and preprocessing steps are applied as those used in the argument detection modeling. For transformer-based modeling, we implement a Topic-Evidence Ranker that scores evidence sentences relative to a given topic. Each claim-evidence pair is encoded using a shared pretrained transformer (e.g., *bert-base-uncased*), and the resulting [CLS] token embedding is passed through a two-layer MLP with a hidden size of 256 to produce a scalar relevance score. During inference, evidence sentences are ranked according to these scores, with higher scores indicating more persuasive argument. The model is trained using pairwise ranking loss for three epochs to distinguish the more convincing evidence. For LLM evaluation, we adopt the same two-shot prompting approach, providing two-shot examples in the prompt (see prompt in the [Appendix B Section B.2](#)).

4.1.2 CSTD Analysis

With the argument detection and argument quality ranking models in place, we first extract topics from the CSTD using *BERTopic* (Grootendorst, 2022). To prepare the data, we aggregate clean text from all speeches across all speakers into a single corpus. The text is then chunked into segments of 512 tokens to prevent truncation, and each chunk is embedded using sentence embeddings (*all-MiniLM-L6-v2*) and passed to *BERTopic*. This setting allow us to capture the whole argument within a text chunk as an argument can span to multiple sentences. For the generated topics, we assign labels using LLMs. Specifically, multiple LLMs including *Gemma3* (Team et al., 2025) and *Qwen2.5* (Qwen et al., 2024) are prompted to suggest labels, and a third LLM (*Llama3.2*) is used to align and finalize the topic labels. Once topics are established, we obtain the corresponding sentences for each topic, which serve as input for the argument detection model. Each topic-sentence pair is processed to identify argumentative state-

ments within the CSTD. Subsequently, arguments are grouped by political affiliation Republican or Democratic. To compare arguments between parties, we form pairs of opposing arguments for each topic. These pairs, along with the associated topics, are then evaluated using the argument quality ranking model to determine which argument is more persuasive. This process yields a ranked list of the most compelling arguments for each topic, along with their associated political party.

Finally, to examine the emotional content of the identified arguments, all argumentative statements are passed through an emotion classification model. This enables analysis of the nuanced emotional tone present in persuasive political statements.

5 Results and Discussion

In this section, we present, results from training argument detection and argument quality model. We also present the analysis on Campaign Speech dataset including topic modeling and argument prediction. Finally, we analyse the results with respect to political affiliation, argument convincingness and emotions involved.

5.1 Argument Modeling Results

Transformer-based models achieved the strongest performance across both the argument detection and argument quality assessment tasks. In argument detection, our primary objective was to maximize precision in order to identify true argumentative statements with higher reliability. Among ML models, Linear SVM with pretrained sentence embeddings performed notably well, achieving a precision of 0.77 and an F1-score of 0.74. In the two-shot LLM setting, *Llama-3.1-8B-Instruct* demonstrated competitive performance, suggesting that parameter-efficient fine-tuning (e.g., PEFT) could further enhance results. However, transformer-based models particularly *DeBERTa-v3-large* achieved the best performance, with a precision of 0.87 and an F1-score of 0.86 (see full results in [Table 1](#)). This model was therefore selected for extracting arguments from CSTD.

For argument quality assessment, we prioritized accuracy, as the goal was to reliably determine the more persuasive argument in each pair. ML models again benefited substantially from sentence-embedding features, which outperformed traditional TF-IDF representations. Among LLMs, *Ministral-8B-Instruct-2410* achieved the highest ac-

Model	ARGDETECT		EVI CONV	
	Precision	F1	Accuracy	F1
Logistic Regression (tf-idf)	0.78	0.72	0.50	0.48
Linear SVM	0.75	0.74	0.47	0.46
Naive Bayes	0.79	0.59	0.52	0.48
LR with SE	0.77	0.74	0.54	0.52
BERT-base-uncased	0.84	0.84	0.75	0.76
RoBERTa-base	0.84	0.84	0.78	0.79
DeBERTa-v3-large	0.87	0.86	0.78	0.78
Qwen2.5-Coder-14B	0.6321 ± 0.0211	0.4581 ± 0.0169	0.616 ± 0.0305	0.6166 ± 0.0305
Minstral-8B-Instruct-2410	0.6473 ± 0.0077	0.5098 ± 0.0154	0.658 ± 0.0223	0.6573 ± 0.0223
Llama-3.1-8B-Instruct	0.6353 ± 0.0196	0.6162 ± 0.0219	0.64 ± 0.045	0.6350 ± 0.0466

Table 1: Results of ML, Transformer, and LLM models on ARGDETECT and EVI CONV, with macro-averaged metrics. For LLMs, mean and standard deviation are from five two-shot samples of 200 instances. Italics mark the best ML model; bold marks the best overall per dataset.

curacy (0.65). Transformer-based models such as *RoBERTa-base* and *DeBERTa-v3-large* performed comparably, and our Topic–Evidence Ranker improved upon the baseline accuracy of 0.73 (Gleize et al., 2019) across all pretrained model. For downstream analysis of CSTD, we selected the *RoBERTa-base* ranker (accuracy 0.78) for argument quality assessment.

5.2 Topic Modeling Results

After applying topic modeling to the CSTD, we identified 88 interpretable topics (see full list in the Appendix C). Although several topics exhibited semantic proximity, we retained them as distinct categories to preserve thematic granularity in downstream analysis. Across these topics, we extracted 11,944 topic-relevant sentences, of which 73% were produced by Republican candidates and 27% by Democratic candidates. Notably, 72 of the 88 topics were discussed by speakers from both parties.

The most frequently discussed topics across all campaign speeches were *Border Control Policies*, *Small Business Policy*, and *Political Alliances* and the *USMCA’s Impact on American Jobs and Agriculture*. Party-specific rankings reveal both convergence and divergence. For Republicans, the top topics include *Border Control Policies* and *USMCA’s Effect on American Jobs and Agriculture*, both of which appear in the overall top three, as well as *U.S. Election: Allegations of Government Corruption and Family Scandals*. In contrast, the Democratic Party’s top three topics are markedly different: *Universal Education Equity*, *Universal Access to Comprehensive Healthcare (Including Pre-existing Conditions)*, and *American Values Advocacy: Constitu-*

tion, Law Enforcement, Faith, Family, and History.

Overall, Democratic speakers tend to emphasize the expansion of social programs and civil rights, including universal healthcare and education, gun control, voting rights, and climate policy through clean-energy initiatives. They also highlight pandemic-related economic relief for individuals and small businesses. Republican speakers, by contrast, prioritize national security, traditional values, and limiting federal intervention. Their discourse frequently centers on border control, “American values” (e.g., constitutionalism, faith, and family), conservative judicial appointments (such as pro-life positions regarding the Supreme Court), investigations into alleged government corruption, and election integrity. Additionally, Republicans devote attention to trade-related economic effects (e.g., USMCA) and veterans’ affairs.

5.3 Political Argument Detection

Using trained argument detection model (see subsection 4.1), we extract arguments used by the politicians in the campaign speech data irrespective of the fact if its a claim or evidence. Overall 36.84% of the total text chunks were arguments. Republican vice presidential candidate Mike Pence emerged as one with highest argument percentage while Donald Trump with least argument percentage as 34.81%.

Within the Democratic party, Joe Biden produced a higher proportion of arguments than Kamala Harris. Certain topics demonstrated particularly high argument rates, including *Tax Relief for Family Farms and Small Businesses*, *COVID-19 Testing Capacity Growth (US, Feb ’20)*, and *Job Dignity & Workplace Respect*, which ranked

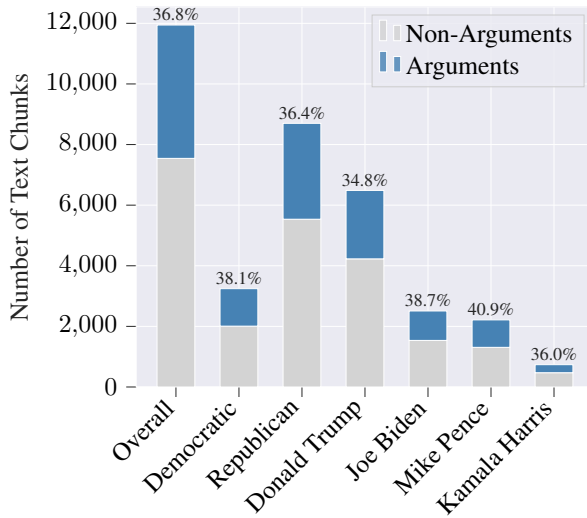


Figure 1: Distribution of predicted arguments in the campaign speech dataset. The stacked bars show the number of argument (steelblue) and non-argument (light gray) text chunks for the overall dataset, broken down by party, and by individual candidate.

as the top three in argument density. Conversely, topics such as *U.S. Election: Allegations of Government Corruption and Family Scandals* and *America Reclamation Movement* exhibited lower argument rates. As illustrated in Figure 1, Democratic speeches contained a higher percentage of arguments overall compared to Republican speeches. It should be noted, however, that the total number of text chunks was greater for Republican speakers, resulting in a higher absolute number of arguments despite the lower proportion.

5.4 Political Argument Quality Analysis

For analyzing argument quality, we automatically paired arguments for each topic between politicians from different parties. For each topic, we extract arguments are extracted independently for each party, after which all possible cross-party argument pairs are generated exhaustively. This setup allows a direct comparison of argumentative strength between parties and aligns naturally with our argument quality model, which is designed to perform relative comparisons between arguments. As shown in Figure 2, the overall distribution of convincing argument pairs is approximately balanced between the two parties. Examining individual performance, Kamala Harris achieves the highest overall convincingness rate, outperforming both Donald Trump and Mike Pence. When examining convincingness at the topic level, Democratic ar-

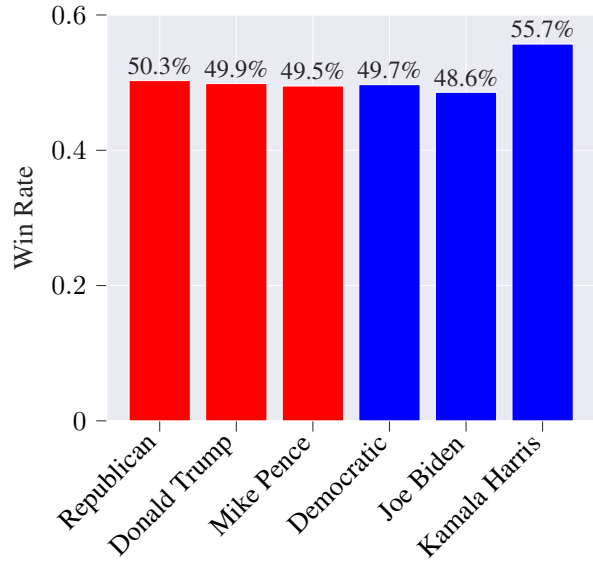


Figure 2: Pairwise convincingness by party and speaker. Bars show win rates, with colors indicating party affiliation (red = Republican, blue = Democratic).

guments exhibit a substantially higher overall rate: 65.5% of their arguments are judged more convincing, compared to 34.4% for Republican arguments. This suggests that, despite variation across individual topics, Democratic speakers generally produce more persuasive arguments. An overall semantic map illustrating topic-level argument convincingness is presented in Figure 3.

5.5 Emotions in Persuasive Political Discourse

In the previous section, we presented pairwise results of argument convincingness. To examine the role of emotions in persuasive political discourse, it is necessary to convert these pairwise results into a single, argument-level convincingness score. To achieve this, we calculate the *argument convincingness rate* by counting the number of times a given argument is selected as the winning argument. These arguments are then processed through an emotion detection model to obtain both the emotion category and corresponding intensity.

Our analysis shows that more than half of the arguments (56.5%) are classified as **neutral**. Beyond neutrality, positive emotions such as admiration, approval, gratitude, and optimism are most frequently associated with arguments. Among negative emotions, disapproval occurs most frequently, followed by sadness, disappointment, and annoyance. Interestingly, although only two arguments are associated with anger, they exhibit high convincingness rate (97.3%). Other negative emotions,

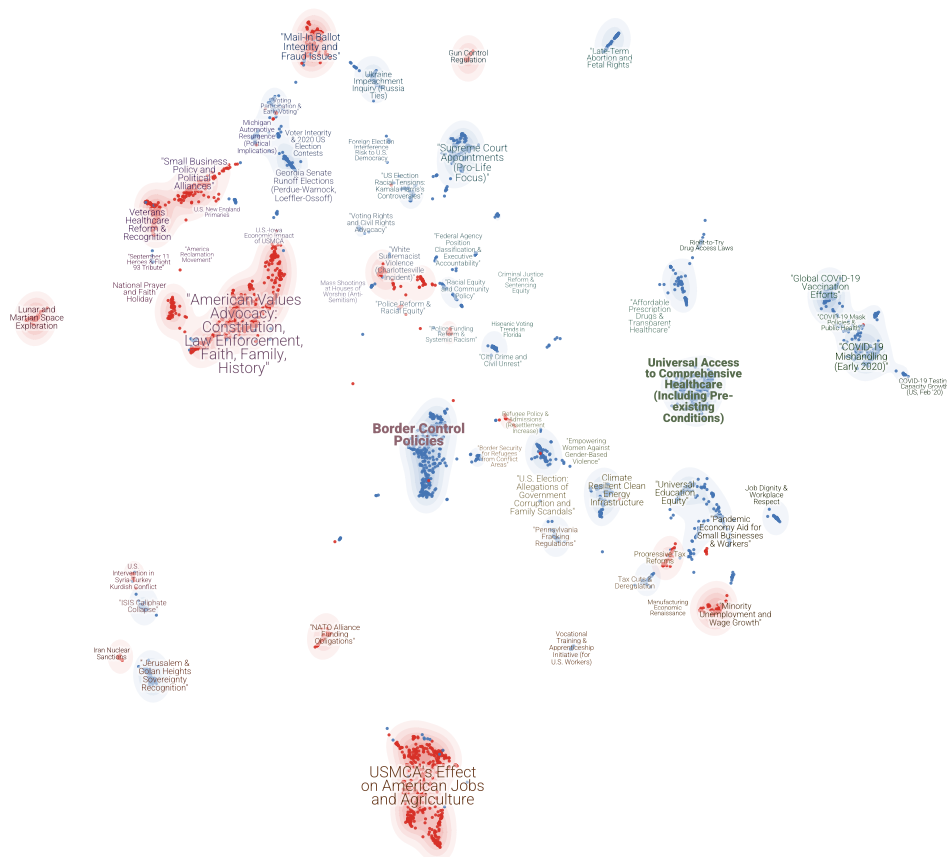


Figure 3: Topic-Convincingness Map for CSTD with blue representing Democratic and red representing Republican

including fear, disapproval, and annoyance, also correspond to arguments with high convincingness. For positive emotions, curiosity, love, amusement, desire, and surprise are among those linked to particularly persuasive arguments. Figure 4 shows distribution of the most dominant emotions across topics in CSTD.

In Republican arguments, the topics with the highest convincing rates were *City Crime and Civil Unrest* (100%) and *Pennsylvania Fracking Regulations* (92%). For *City Crime and Civil Unrest*, emotions observed included admiration, sadness, love, annoyance, and disappointment, with intensity analysis showing that **sadness and annoyance** were more pronounced than admiration and love. For *Pennsylvania Fracking Regulations*, disapproval and admiration appeared with comparable intensity. Examining the most-discussed Republican topics, such as *USMCA's Effect on American Jobs and Agriculture* and *Border Control Policies*, we found that **admiration, approval, and gratitude** were prevalent, with admiration and gratitude exhibiting the highest intensity.

In Democratic arguments, the most persuasive

topics were *White Supremacist Violence (Charlottesville Incident)* (98%) and *Progressive Tax Reforms* (77%). For the Charlottesville Incident related topic, key emotions included fear, approval, and sadness, with **fear and sadness showing high intensity**, though neutral statements constituted 83% of the total. For *Progressive Tax Reforms*, the dominant emotions were approval and curiosity. Looking at the most-discussed Democratic topics, such as *American Values Advocacy*, *Universal Access to Comprehensive Healthcare*, and *COVID-19 Mishandling (Early 2020)*, emotions varied by topic. In *COVID-19 Mishandling*, sadness and gratitude were prominent, with gratitude exhibiting high intensity. In healthcare-related discussions, approval predominated, accompanied by other emotions such as sadness and curiosity.

Overall, the analysis indicates that politicians employ emotions strategically in their argumentation. For example, Democrats leverage negative emotions such as fear and sadness when discussing topics related to white supremacist violence, aligning their messaging with public concern on these issues. In contrast, when addressing *Progressive Tax*

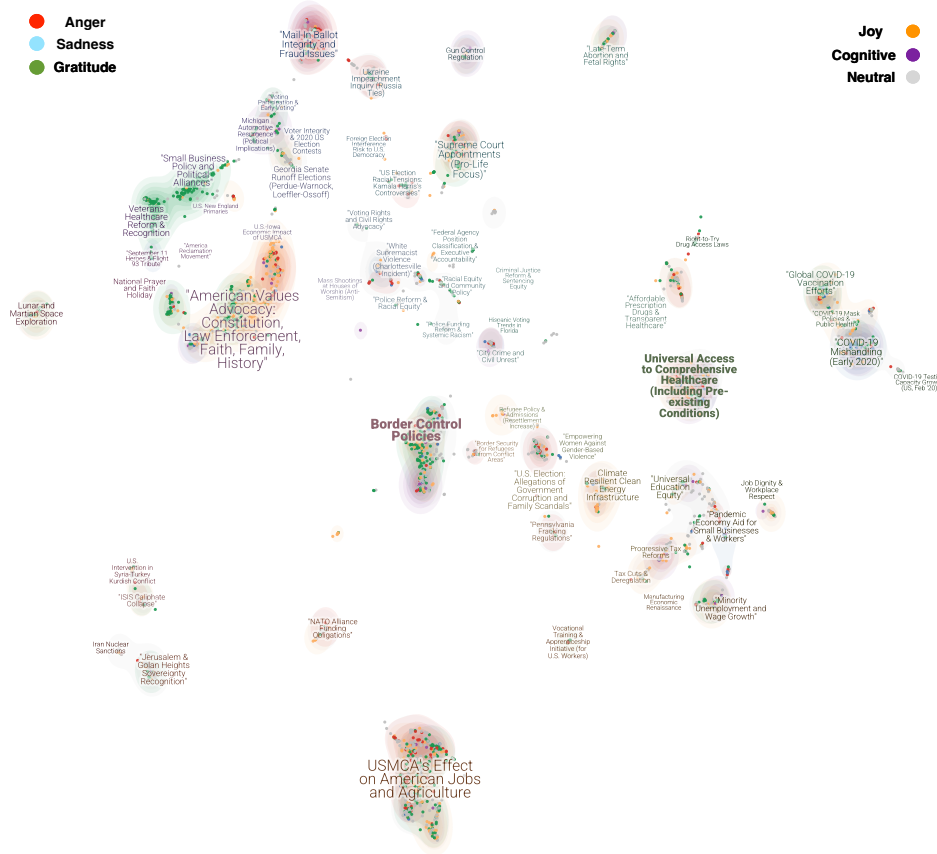


Figure 4: Topic-Emotion Map for CSTD. Emotional tone used by both political parties across various topics identified in the CSTD.

Reform, they predominantly use positive emotions like approval to appeal to voters who support taxing the wealthy. Similarly, Republican candidates utilize positive emotions such as admiration and approval when discussing the USMCA trade agreement, whereas for topics like city crime and civil unrest, particularly in states with active protests, they employ negative emotions to resonate with their voter base.

5.6 Randomization Based Significance Tests

To assess the statistical significance of our findings, we conducted a series of randomization tests. First, we examined argument detection results. By maintaining the original proportion of text excerpts per party, we randomly reassigned excerpts to each party. A direct comparison between Democrats and Republicans showed no significant difference relative to the random baseline, indicating that the observed difference in argumentative text proportions could occur by chance.

Next, we investigated argument quality, measured via convincingness scores. All arguments were randomly reassigned across parties while

keeping the number of arguments per party constant, and average convincingness was computed in each randomization. The observed results indicated comparable convincingness between parties, with Republicans exhibiting a slight advantage (0.5%). Randomization testing, however, revealed no evidence of a significant overall pairwise advantage for either party.

We assessed topic-level argument comparisons. For each topic, multiple pairwise comparisons between arguments were available, with the winner of each comparison recorded. Randomization was applied within topics by shuffling winners across comparisons, and the resulting distributions were compared to the observed win ratios. This analysis identified 10 topics in which Democratic arguments exhibited a significant edge, indicating a topic-specific advantage in convincingness. Similarly, 15 topics were found in which Republican arguments showed significant superiority.

Finally, we aggregated results across topics to examine overall topic-level wins. In the observed data, Democrats were the majority winner in 64%

of topics based on argument convincingness (see subsection 5.4). Randomization testing showed that, despite near equivalence in global pairwise wins, Democrats won significantly more topics than expected by chance. This indicates a topic-level advantage for Democrats that was not evident from global pairwise comparisons alone.

6 Conclusion

We presented a computational analysis of campaign speeches, focusing on emotion’s role in persuasive discourse. To achieve this, we trained models for argument detection and argument quality assessment. Our argument detection model achieved **0.87** precision, accurately identifying argumentative statements in political speech. The argument quality model achieved **0.78** accuracy, offering reliable persuasiveness evaluation. Using these models, we found that both parties rely heavily on argumentation in their U.S. political speeches. To assess argument effectiveness, we applied a rhetorical lens and found that high-quality arguments boost persuasion, with Democrats winning 65% of the time. Our analysis also shows that politicians strategically use emotional cues to enhance message persuasiveness. While our study is purely computational, we view human supervision as an important direction for future work to validate and complement these findings. Future work will analyze arguments logically to identify fallacies in political speech. Combining logical, rhetorical, and emotional analyses could reveal whether fallacious arguments are especially persuasive and which emotions they evoke.

Limitations

There are few limitations to our study that should be noted. First, while training the argument detection model, we did not split the data based on topics. Although we subsequently tested the model using topic-based splits and observed minimal differences in performance, this approach may still overlook subtle topic-specific patterns in argumentation. Second, our emotion detection model assigns a single dominant emotion to each argument. In reality, multiple emotions with varying intensities can be associated with a single utterance. While we partially addressed this by aggregating emotion distributions at the speaker and topic levels, a more fine-grained analysis would require a system capable of capturing the full emotion distribution for each individual argument. Third, to identify argu-

mentative content in political speeches, we segment the text into non-overlapping chunks of up to 512 tokens. While this chunking strategy allows us to capture arguments that span multiple sentences, it may also dilute argumentative signals, potentially causing some argumentative content to be misclassified as non-argumentative. Future work could explore sequence-level or hierarchical modeling approaches to more precisely capture argumentative structure within longer text spans. Finally, we acknowledge that pretrained language models may encode political or ideological biases, which could influence convincingness predictions. Although our randomization-based analyses mitigate some aggregate effects, we cannot fully rule out such biases, and we view this as an important limitation of the current study.

References

- Segun T Aroyehun, Almog Simchon, Fabio Carrella, Jana Lasser, Stephan Lewandowsky, and David Garcia. 2025. Computational analysis of us congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, pages 1–12.
- Mohamed S Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *International Conference on Human-Computer Interaction*, pages 659–671. Springer.
- Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. [Emotions in argumentation: an empirical evaluation](#). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 156–163. AAAI Press.
- Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969.
- Amparo Elizabeth Cano-Basave and Yulan He. 2016. [A study of the impact of persuasive argumentation in political debates](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California. Association for Computational Linguistics.
- Ioannis Chalkiadakis, Louise Anglès d’Auriac, Gareth W Peters, and Divina Frau-Meigs. 2025. A text dataset of campaign speeches of the main tickets in the 2020 us presidential election. *Scientific Data*, 12(1):662.

- Dennis Chong and James N Druckman. 2007. A theory of framing and opinion formation in competitive elite environments. *Journal of communication*, 57(1):99–118.
- Alan Darmasaputra Chowanda, Albert Richard Sanyoto, Derwin Suhartono, and Criscentia Jessica Setiadi. 2017. Automatic debate text summarization in online debate forum. *Procedia computer science*, 116:11–19.
- James N Druckman. 2022. A framework for the study of persuasion. *Annual Review of Political Science*, 25:65–88.
- Barbara Early. 2015. The righteous mind: Why good people are divided by politics and religion, by jonathan haidt: (2012). new york, ny: Pantheon books, illustrated, 419 pp., \$28.95 (hard cover).
- Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. **Corpus wide argument mining - A working solution**. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691. AAAI Press.
- Kevin Finitiy, Ramit Garg, and Max McGaw. 2021. A text analysis of the 2020 us presidential election campaign speeches. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE.
- Gloria Gennaro and Elliott Ash. 2022. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.
- Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkovich, Ranit Aharonov, and Noam Slonim. 2019. **Are you convinced? choosing the more convincing evidence with a Siamese network**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.
- Lynn Greschner and Roman Klinger. 2024. **Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms**. *ArXiv preprint*, abs/2412.15993.
- Maarten Grootendorst. 2022. **Bertopic: Neural topic modeling with a class-based tf-idf procedure**. *ArXiv preprint*, abs/2203.05794.
- David J Grüning and Thomas W Schubert. 2022. Emotional campaigning in politics: being moved and anger in political ads motivate to support candidate and party. *Frontiers in Psychology*, 12:781851.
- Ivan Habernal and Iryna Gurevych. 2016. **What makes a convincing argument? empirical analysis and detecting attributes of convincingsness in web argumentation**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.
- Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. **Yes, we can! mining arguments in 50 years of US presidential campaign debates**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.
- Hans JG Hassell, Christopher D Johnston, Jessica Khan, and Edgar Cook. 2022. The nature and impact of emotional content in congressional candidate emails to supporters. *Electoral Studies*, 79:102501.
- Rositsa Ivanova, Thomas Huber, and Christina Niklaus. 2024. **Let’s discuss! quality dimensions and annotated datasets for computational argument quality assessment**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779.
- John Lawrence and Chris Reed. 2019. **Argument mining: A survey**. *Computational Linguistics*, 45(4):765–818.
- Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. Emotion and decision making. *Annual review of psychology*, 66(1):799–823.
- Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. **Towards an argumentative content search engine using weak supervision**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Marco Lippi and Paolo Torrioni. 2016. **Argument mining from speech: Detecting claims in political debates**. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2979–2985. AAAI Press.
- Mounika Marreddy, Subba Reddy Oota, Venkata Charan Chinni, Manish Gupta, and Lucie Flek. 2025. **Usdc: A dataset of user stance and dogmatism in long conversations**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23715–23759.
- Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. **Never retreat, never retract: Argumentation analysis for political speeches**. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4889–4896. AAAI Press.

- Sovesh Mohapatra and Somesh Mohapatra. 2022. [Sentiment is all you need to win US presidential elections](#). In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 15–20, Taipei, Taiwan. Association for Computational Linguistics.
- Richard E Petty and John T Cacioppo. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.
- Trong Huy Phan and Kazuma Yamamoto. 2020. [Resolving class imbalance in object detection with weighted cross entropy losses](#). *ArXiv preprint*, abs/2006.01413.
- Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. [Modeling user arguments, interactions, and attributes for stance prediction in online debate forums](#). In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 855–863. SIAM.
- Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2024. [Qwen2.5 technical report](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Valerie Rodriguez-Hernandez, Vanesa Hidalgo, and Alicia Salvador. 2024. Emotional and cognitive processes underlying persuasion, moderating factors, and physiological reactions: A systematic review. *Psychological Reports*, page 00332941241291497.
- Sandy Ross. 2006. Don’t think of an elephant: Know your values and frame the debate. *Melbourne Journal of Politics*, 31:145–149.
- Tamanna M Shah. 2022. Emotions in politics: A review of contemporary perspectives and trends. *International Political Science Abstracts*, 74(1):1–14.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. [Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. [Gemma 3 technical report](#). *ArXiv preprint*, abs/2503.19786.
- Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. [Using argument mining to assess the argumentation quality of essays](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Overview of Appendix Sections

- Section A: Dataset Statistics
- Section B: LLM Prompts
- Section C: CSTD Topics

A Dataset Statistics

Statistic	Train	Validation
Total instances	25,542	6,386
Unique topics	240	236
Avg. sentences per topic	106.4	27.1
Avg. sentence length (tokens)	31.0	31.0

Table 2: Dataset statistics for the ARGDETECT corpus split into train and validation.

Statistic	Train	Validation
Total instances	4,319	1,378
Unique topics	48	21
Avg. instances per topic	90	65
Avg. sentence length (tokens)	28	28

Table 3: Dataset statistics for the EVICONV corpus across train and validation splits.

In this section, we present the full statistics of the dataset used for argument detection (Table 2) and argument quality assessment (Table 3) training.

B LLM Prompts

B.1 Argument Detection

Figure 5 shows the prompt that we used as annotation guidelines to obtain argument detection annotations from LLMs.

B.2 Argument Quality Prompt

Figure 6 shows the prompt that we used as annotation guidelines to obtain argument quality assessment annotations from LLMs.

C Topics in CSTD

88 topics are shown in Table 4 extracted using BERTopic and labelled using LLMs like Gemma3, Qwen2.5 and Llama3.2.

C.1 Top 10 topics

Table 5 shows top 10 topics for Democratic party where they exhibit higher convincingness. Similarly, Table 6 shows the same for the Republican party.

Argument Detection

You are an expert in argument mining and discourse analysis. Your task is to determine if a given Sentence/Statement functions as an argument (a claim, reason, or piece of evidence) for or against the provided Topic.

The output should be a single integer:

- Label: 1 if the Sentence/Statement acts as an argument (pro or con) for the Topic.
- Label: 0 if the Sentence/Statement is purely descriptive, definitional, or factually related to the Topic but does not attempt to persuade, support, or challenge a viewpoint about it.

In other words, check if the Sentence/Statement offers a reason why someone should think a certain way about the Topic.

Example 1:

Topic: Amnesty International

Sentence: In a report of November 1988, Amnesty International said that the number of prisons had increased to 644 and their capacity had been raised from 55,000 to more than 80,000 [REF].

Label: 0

Example 2:

Topic: Environmental technology

Sentence: Designed to transform the site into a 21st-century global laboratory which brings together business, education and green technologies, it aims to create over 1,000 jobs for the local area and generate £29m private sector investment.

Label: 1

Now answer the following:

Topic: {topic}

Sentence: {sentence}

Label:

Figure 5: Prompt for Argument Detection

Argument Quality Assessment

You are an expert at evaluating the quality of arguments based on evidence provided. Given a claim and two pieces of evidence, your task is to determine which piece of evidence better supports the claim.

Even if both pieces of evidence are weak or imperfect, you must choose the one that is relatively better at supporting the claim.

Example 1:

Claim: We should legalize prostitution

Evidence 1: The appellants' argument then, more precisely stated, is that in criminalizing so many activities surrounding the act itself, Parliament has made prostitution de facto illegal if not de jure illegal.

Evidence 2: Feminists who hold such views on prostitution include Kathleen Barry, Melissa Farley, Julie Bindel, Sheila Jeffreys, Catharine MacKinnon and Laura Lederer; the European Women's Lobby has also condemned prostitution as "an intolerable form of male violence".

Answer: 1

Example 2:

Claim: We should subsidize condoms

Evidence 1: In 2009, Lewis strongly criticized Pope Benedict XVI's assertion that condom use only makes the AIDS crisis worse.

Evidence 2: Green said that according to the "best studies," condoms makes people take wilder sexual risks, thus worsening the spread of the disease.

Answer: 2

Now answer the following:

Claim: {claim}

Evidence 1: {evidence_1}

Evidence 2: {evidence_2}

Answer:

Figure 6: Prompt for Argument Quality Assessment

ID	Topic	ID	Topic
1	Veterans Healthcare Reform & Recognition	45	Venezuelan Democratic Crisis (Regime Transition)
2	U.S. Military Modernization Budget	46	Pennsylvania Fracking Regulations
3	Border Control Policies	47	Job Dignity & Workplace Respect
4	Georgia Senate Runoff Elections (Perdue–Warnock, Loeffler–Ossoff)	48	U.S. New England Primaries
5	Jerusalem & Golan Heights Sovereignty Recognition	49	US Election Racial Tensions: Kamala Harris’s Controversies
6	ISIS Caliphate Collapse	50	Voting Participation & Early Voting
7	Iran Nuclear Sanctions	51	Voter Integrity & 2020 US Election Contests
8	Mass Shootings at Houses of Worship (Anti-Semitism)	52	NATO Alliance Funding Obligations
9	U.S. Space Governance and NASA Growth	53	Middle Class Empowerment & Labor Strength
10	Lunar and Martian Space Exploration	54	Progressive Tax Reforms
11	Small Business Policy and Political Alliances	55	Federal Agency Position Classification & Executive Accountability
12	Tax Cuts & Deregulation	56	Universal Access to Comprehensive Healthcare (Including Pre-existing Conditions)
13	Minority Unemployment and Wage Growth	57	White Supremacist Violence (Charlottesville Incident)
14	Vocational Training & Apprenticeship Initiative (for U.S. Workers)	58	Climate Resilient Clean Energy Infrastructure
15	Criminal Justice Reform & Sentencing Equity	59	Racial Equity and Community Policy
16	Gun Control Regulation	60	Foreign Election Interference Risk to U.S. Democracy
17	National Prayer and Faith Holiday	61	Refugee Policy & Admissions (Resettlement Increase)
18	American Values Advocacy: Constitution, Law Enforcement, Faith, Family, History	62	Empowering Women Against Gender-Based Violence
19	U.S. Agricultural Policy & Farmer Support	63	Policy on National Security, Gun Control, & Immigration Economics
20	USMCA’s Effect on American Jobs and Agriculture	64	North Korea’s Nuclear Diplomacy (U.S., Russia focus)
21	Energy Independence & Resource Dominance	65	U.S. Intervention in Syria–Turkey Kurdish Conflict
22	U.S.-Iowa Economic Impact of USMCA	66	Suburban Women’s & Housing Safety Regulations
23	National Defense Expansion	67	Historic, Diverse, Inclusive Cabinet – Environmental Priorities
24	Universal Education Equity	68	US COVID-19 Health Policy & Treatments
25	Supreme Court Appointments (Pro-Life Focus)	69	Affordable Prescription Drugs & Transparent Healthcare
26	Voting Rights and Civil Rights Advocacy	70	City Crime and Civil Unrest
27	Manufacturing Economic Renaissance	71	Dreamer Citizenship Policy
28	Impact of Soleimani’s Death on Terrorism in Iraq	72	Hispanic Voting Trends in Florida
29	Ukraine Impeachment Inquiry (Russia Ties)	73	Right-to-Try Drug Access Laws
30	Michigan Automotive Resurgence (Political Implications)	74	COVID-19 Mask Policies & Public Health
31	Late-Term Abortion and Fetal Rights	75	Mail-In Ballot Integrity and Fraud Issues
32	Economic Freedom and Anti-Socialism in America	76	America Reclamation Movement
33	Police Reform & Racial Equity	77	Public Gathering Management Policies & Controversies
34	Law Enforcement Reform & Support	78	Stock Market Performance & Investing
35	COVID-19 Manufacturing Response (Ventilators)	79	Tax Relief for Family Farms and Small Businesses
36	COVID-19 Testing Capacity Growth (US, Feb ’20)	80	Opioid Crisis Treatment Across Regions (Montana & Kentucky)
37	Pandemic Economy Aid for Small Businesses & Workers	81	Politics of Anti-Immigration in Minnesota (re: Omar)
38	COVID-19 Mishandling (Early 2020)	82	Household Appliance Water Efficiency Regulations
39	Veterans Healthcare Accountability Act	83	Media Ownership Impact on News Reporting & Bias
40	Police Funding Reform & Systemic Racism	84	Boeing Contract Cancellation Fees
41	Global COVID-19 Vaccination Efforts	85	Monument Preservation vs Removal Conflict
42	U.S. Election: Allegations of Government Corruption and Family Scandals	86	Politics and Law Enforcement
43	Firearms Regulation and Second Amendment	87	Rape Kit Backlog & Victim Support Services
44	Border Security for Refugees from Conflict Areas	88	September 11 Heroes & Flight 93 Tribute

Table 4: Topics extracted using BERTopic and labelled using LLMs like Gemma3, Qwen2.5 and Llama3.2

Topic	Dem Wins	Rep Wins	Dem Win Rate
COVID-19 Mask Policies & Public Health	54	0	1.000
Border Security for Refugees from Conflict Areas	30	0	1.000
City Crime and Civil Unrest	34	0	1.000
Job Dignity & Workplace Respect	74	0	1.000
Michigan Automotive Resurgence	27	0	1.000
Pennsylvania Fracking Regulations	72	6	0.923
Hispanic Voting Trends in Florida	11	1	0.917
USMCA (Iowa Economic Impact)	56	7	0.889
September 11 Heroes & Flight 93 Tribute	14	2	0.875
Voting Participation & Early Voting	47	7	0.870

Table 5: Top 10 topics where Democratic speakers were more convincing.

Topic	Dem Wins	Rep Wins	Dem Win Rate
U.S. New England Primaries	0	4	0.000
America Reclamation Movement	0	7	0.000
White Supremacist Violence (Charlottesville)	1	42	0.023
Minority Unemployment & Wage Growth	16	124	0.114
Lunar and Martian Space Exploration	11	65	0.145
Iran Nuclear Sanctions	15	54	0.217
Empowering Women Against Gender-Based Violence	18	48	0.273
Veterans Healthcare Reform & Recognition	55	141	0.281
Mail-In Ballot Integrity & Fraud Issues	309	559	0.356
NATO Funding Obligations	36	58	0.383

Table 6: Top 10 topics where Republican speakers were more convincing.

Measuring LLMs’ Sensitivity to Paraphrased Opinion Prompt

Bushra Alhetelah

Department of Computer Engineering
King Fahd University of Petroleum
& Minerals (KFUPM)
Dhahran 31261, Saudi Arabia
g202401980@kfupm.edu.sa

Irfan Ahmad

Department of Information and Computer
Science, King Fahd University of Petroleum
& Minerals (KFUPM),
SDAIA–KFUPM Joint Research Center
for Artificial Intelligence
Dhahran 31261, Saudi Arabia
irfan.ahmad@kfupm.edu.sa

Abstract

Large language models (LLMs) are now widely used in applications that depend on closed-ended decisions, including automated surveys, policy screening, and decision-support tools. In such contexts, these models are typically expected to produce consistent binary or ternary responses (for example, Yes, No, or Neither) when presented with questions that are semantically equivalent. However, recent studies show that LLM outputs can be influenced by relatively minor changes in prompt wording, raising concerns about the reliability of their decisions under paraphrasing. In this paper, we conduct a systematic analysis of paraphrase robustness across five widely used LLMs. To support this evaluation, we develop a controlled dataset consisting of 200 opinion-based questions drawn from multiple domains, each accompanied by five human-validated paraphrases. All models are evaluated under deterministic inference settings and constrained to a fixed Yes/No/Neither response format. We assess model behavior using a set of complementary metrics that capture the stability of each evaluated model. DeepSeek Reasoner and Gemini 2.0 Flash show the highest stability when responding to paraphrased inputs, whereas Claude 3.7 Sonnet exhibits strong internal consistency but produces judgments that differ more frequently from those of other models. By contrast, GPT-3.5 Turbo and LLaMA 3 70B display greater sensitivity to surface-level variations in prompt phrasing. Overall, these findings suggest that robustness to paraphrasing is driven more by alignment strategies and reasoning design choices than by model size alone.

1 Introduction

LLMs have become an enabling technology across many aspects of life. They can generate natural language and assist across a wide range of domains, including healthcare, education, public policy, and

e-commerce (Naveed et al., 2025; Siino et al., 2025; Raiaan et al., 2024). However, in many real-world scenarios LLMs are not mainly used for long-form text generation. Instead, they are also deployed to produce direct decisions, typically in the form of binary or ternary responses such as Yes, No, or Neither. These decisions can influence significant processes, including clinical triage, policy assessment, and public opinion analysis, making the reliability and stability of model outputs especially important (Buhnila et al., 2024; Karanjai et al., 2025). Despite their strong performance, LLMs are known to be sensitive to variations in prompt formulation (Wang et al., 2021; Errica et al., 2025). Multiple studies have shown that semantically equivalent inputs may cause models to generate different outputs which slightly differ in wording, syntax, or structure. This has been examined extensively in open-ended generation tasks, while it has received less attention in closed-ended settings. In practice, even minor rephrasing—such as substituting synonyms or reordering clauses—can lead to changes in binary judgments, potentially undermining trust in downstream systems.

This study addresses this gap by systematically evaluating the robustness of five state-of-the-art LLMs—OpenAI GPT-3.5 Turbo, Claude 3.7 Sonnet, DeepSeek Reasoner, Google Gemini 2.0 Flash, and LLaMA 3 70B—when responding to paraphrased opinion prompts. We construct a dataset of 200 opinion questions spanning multiple domains, each paired with five LLM-generated and human-validated paraphrases. Model behavior is analyzed under controlled conditions using a fixed Yes/No/Neither response format. By measuring consistency, agreement, and variability across paraphrased inputs, this work provides insight into how different model design and alignment choices influence decision stability under semantic equivalence.

Our findings provide new insights into how LLMs respond to paraphrased opinion prompts,

highlighting both recent progress in improving decision stability and the continued need for careful evaluation of robustness in closed-ended, decision-oriented applications.

The remainder of this paper is organized as follows. Section 2 reviews prior work on LLM robustness, paraphrasing, and opinion-focused question answering. Section 3 describes the methodology and experimental setup. Section 4 presents the results along with a comparative analysis, followed by discussion, ethical considerations and limitations.

Prior studies have shown that LLM outputs can change under paraphrased inputs. However, most existing work primarily examines response stability within specific application domains, such as psychometric questionnaires or political opinion surveys (Haller et al., 2024). For example, recent large-scale analyses generate multiple paraphrases per item and evaluate response validity and stability across models. While these studies provide important insights into paraphrase robustness, they typically focus on domain-specific settings or evaluate robustness using a single perspective, such as response variance or validity.

In contrast, this work systematically measures sensitivity to paraphrased opinion prompts using controlled prompt groups and consistency-based evaluation metrics across multiple modern LLM families. Rather than evaluating stability only at the individual prompt level, we analyze consistency patterns across structured paraphrase groups and across models. This enables a comparative and model-agnostic analysis of paraphrase sensitivity in closed-ended opinion tasks. Furthermore, the dataset and evaluation framework introduced in this work are designed to support reproducible cross-model analysis under semantically equivalent prompt variations, providing a complementary perspective to prior stability-focused studies.

Fine-tuning and instruction modification are as important as pretraining when adapting LLMs for practical deployment (Wu et al., 2025). Instruction tuning ensures that models behave in accordance with human intent by training them on carefully curated prompt–response pairs, often combined with reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). Modern LLMs such as ChatGPT and Gemini demonstrate that instruction tuning improves instruction-following, coherence, and safety alignment (Kalla et al., 2023; Deng et al., 2025; Team et al., 2024). These alignment

and tuning choices strongly influence how models respond to variations in prompt formulation.

Paraphrasing preserves the semantic meaning of a question while modifying its surface form through lexical, syntactic, or grammatical changes (Bhagat and Hovy, 2013). An increasing body of research demonstrates that surface-level variations in prompt wording can significantly influence LLM behavior. These findings highlight the strong sensitivity of LLMs to prompt phrasing. Paraphrasing is commonly used to improve model performance and robustness by rephrasing prompts while preserving semantic intent. However, this also exposes an important limitation: semantically equivalent prompts do not always produce consistent model outputs, particularly in decision-making and classification settings. The CheckList behavioral testing framework (Ribeiro et al., 2020) evaluates model robustness using capability-based test suites designed to expose systematic model weaknesses. Recent studies further confirm that LLM outputs can vary substantially across paraphrased prompts, even when response formats are constrained and tasks are closed-ended (Haller et al., 2024).

In the context of opinion-focused question answering, paraphrasing introduces additional challenges beyond sentiment classification, including the identification of opinion holders. Kim and Hovy 2005 established opinion-holder recognition as a ranking problem among candidate entities, demonstrating that grammatical structure and contextual signals are critical for accurate attribution. Prior work on human respondents has also shown that paraphrased question formulations can alter how individuals interpret opinion expressions, affecting categorization and extraction performance (Uemlianin, 2000). Similarly, studies on LLMs show that semantically equivalent paraphrases can lead to different model outputs and reduced decision stability (Haller et al., 2024).

Paraphrasing also presents challenges for robustness and detection. Prior work shows that adversarial paraphrasing can significantly degrade the performance of AI-generated text detectors, while training on paraphrased examples can improve detection robustness (Kurt Pehlivanoglu et al., 2024; Lau and Zubiaga, 2025). Beyond automated systems, prompt formulation has also been shown to shape human–AI interaction, influencing user trust, reliance, and decision diversity in mixed human–AI decision-making settings (Lai et al., 2023; Chen et al., 2023). Together, these findings demonstrate

that prompt phrasing affects both model behavior and downstream human interpretation.

Although previous studies have investigated the effects of paraphrasing in open-ended generation, retrieval, detection, and relevance judgment tasks, comprehensive evaluation of paraphrase robustness in closed-ended, decision-oriented opinion prompts remains limited. In particular, the consistency of fixed Yes/No/Neither responses under controlled paraphrase variation has not been systematically evaluated across multiple modern LLM families. This gap directly motivates the empirical evaluation presented in this work, which measures response consistency under controlled paraphrase variation across diverse LLM architectures.

2 Methodology

This section describes the experimental design used to evaluate the robustness of large language models (LLMs) to paraphrased opinion prompts. We outline the construction of the paraphrase dataset, the selection of evaluated models, the prompt execution protocol, and the quantitative metrics used to measure consistency and agreement.

The experimental setup used in this work was designed specifically to address the research objective of measuring paraphrase sensitivity in closed-ended opinion prompts, rather than directly adopting a single existing evaluation methodology. However, the design is conceptually aligned with prior work on robustness and behavioral evaluation of NLP systems, which emphasizes evaluating model behavior under semantically equivalent input transformations and paraphrased prompts (Ribeiro et al., 2020; Haller et al., 2024).

Building on these evaluation principles, this work introduces a task-specific framework based on controlled paraphrase groups and consistency-based metrics across multiple modern LLM families. This enables systematic cross-model comparison while maintaining consistency with established robustness evaluation practices.

All experiments were conducted under controlled and deterministic conditions to ensure that observed variations in model outputs were attributable to paraphrase differences rather than sampling randomness or decoding variability.

2.1 Data Collection

A set of 200 distinct opinion prompts spanning technology, public policy, work culture, and ed-

ucation was generated using OpenAI’s ChatGPT-4o-mini model. Prompt generation was guided by domain-balanced topic selection to ensure coverage across major societal and decision-oriented themes. For each original prompt, five paraphrased variants were generated using the same model, yielding a total of 1,200 items (200 original prompts and 1,000 paraphrases). All automatically generated paraphrases were manually reviewed and verified by the authors to ensure semantic equivalence with the original question. To illustrate the nature of the paraphrasing process, Table 1 presents a representative example consisting of one original opinion question and its corresponding paraphrased variants.

The paraphrase validation process was conducted through manual review by one of the authors to ensure semantic equivalence between each paraphrased prompt and its corresponding original question. During this process, paraphrases were evaluated for preservation of intent, opinion framing, and decision context. Paraphrases that introduced semantic drift, ambiguity, or unintended bias were either corrected or regenerated. Overall, approximately 7% of the automatically generated paraphrases required minor manual adjustments to ensure semantic consistency with the source prompt.

Because validation was performed by a single annotator, no inter-annotator agreement metric was computed. However, a consistent validation protocol was applied across all prompts, focusing on semantic equivalence, grammatical correctness, and preservation of response intent.

To support reproducibility and facilitate future research, the dataset is publicly released via Hugging Face under the name Paraphrased Opinion Prompt Sensitivity (POPS) Dataset¹

2.2 Model Selection and Specifications

We deliberately selected a diverse set of language models to capture variation across model size, reasoning capability, and deployment properties, as summarized in Table 2. The selected models represent diversity across deployment type, reasoning specialization, and alignment design rather than strict parameter scale comparisons. The evaluation includes both open-weight and closed, provider-managed systems, as well as models optimized

¹https://huggingface.co/datasets/mazew2000/Paraphrased_Opinion_Prompt_Sensitivity_Dataset_POPS

1	Original Opinion Question	Do you believe higher education should be provided at no cost to students?
2	Paraphrased Questions	Should governments eliminate tuition fees for university-level education?
3		In your opinion, should college education be free for everyone?
4		Would it be fair to make university education accessible without financial burden?
5		Do you support the idea of publicly funded college education for all citizens?
6		Should students have the right to attend university without paying tuition fees?

Table 1: Sample Original Opinion Prompt and Corresponding Paraphrased Questions

for general-purpose reasoning and models specifically designed for structured reasoning tasks. This diversity enables comparative analysis across performance, operational cost, latency, and real-world deployment constraints.

As shown in Table 2, Claude 3.7 Sonnet and GPT-3.5 Turbo represent closed, provider-managed models designed for strong safety alignment, reliability, and production integration. LLaMA 3 70B represents a large open-weight model supporting full customization, local deployment, and research flexibility (Ersoy and Erşahin, 2024; Gue et al., 2024). DeepSeek Reasoner was included to capture reasoning-specialized behavior under structured decision tasks, while Gemini 2.0 Flash represents a low-latency model optimized for real-time interaction scenarios (Comanici et al., 2025). Together, these models enable comparative evaluation across differences in deployment architecture, reasoning behavior, alignment strategy, and system-level trade-offs.

2.3 Prompt Design and Experimental Execution

To ensure consistent and deterministic model behavior, all prompts were issued under a fixed system instruction: “You are a strict yes-or-no answerer. For each question, answer with exactly ‘Yes’, ‘No’, or ‘Neither’. No numbering, no extra text.” This instruction was applied verbatim to

every request. In addition, the temperature parameter was set to 0.0 for all model calls, minimizing stochastic variation and ensuring that observed response differences arose from paraphrase variation rather than sampling randomness. To comply with rate-limit constraints and maintain execution stability, prompts were dispatched sequentially in fixed-size batches of ten, with a 200 ms delay enforced between API calls. No parallelization was employed, ensuring strict request ordering.

2.4 Evaluation Metrics

Four primary metrics were computed for each model to quantify robustness and comparison under paraphrase variation.

Consistency Rate (CR). Consistency Rate measures the proportion of paraphrased responses that match the majority answer within each question group. For a question group q with P paraphrases, it is defined as:

$$CR_q = \frac{1}{P} \sum_{i=1}^P I(a_{q,i} = a_q^{\text{maj}}), \quad (1)$$

where $I(\cdot)$ is the indicator function. The overall model consistency is obtained by averaging CR_q across all Q questions.

Pairwise Agreement (PA). Pairwise Agreement measures the proportion of identical responses produced by two models M_1 and M_2 on the same prompts:

$$PA(M_1, M_2) = \frac{1}{N} \sum_{i=1}^N I(a_i^{(M_1)} = a_i^{(M_2)}), \quad (2)$$

where N denotes the total number of prompts.

Cohen’s Kappa (κ). Cohen’s Kappa evaluates agreement between a model’s responses to original prompts and their paraphrases while correcting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (3)$$

where p_o is the observed agreement and p_e is the expected agreement under random labeling.

Aggregate Statistics. For each model, the mean consistency rate and 95% confidence intervals across question groups were computed to summarize overall robustness and uncertainty in model responses under paraphrased prompts.

Attribute	Claude 3.7 Sonnet	LLaMA 3 70B	DeepSeek Reasoner	Gemini 2.0 Flash	GPT-3.5 Turbo
Parameters	Not publicly disclosed	70B (open weights)	Not publicly disclosed	Not publicly disclosed	Not publicly disclosed
Reasoning Orientation	General-use	General-use	Reasoning-oriented	Paraphrase-robust	General-use
Context Window	Large (varies by plan/API)	8k tokens (standard LLaMA 3)	Varies by offering/API	Varies by offering/API	Up to 16k (variant-dependent)
Inference Cost	Moderate–High (relative)	Moderate (deployment-dependent)	Moderate (relative)	Low (relative)	Low (relative)
Latency	Low–Moderate (deployment-dependent)	Moderate (hardware-dependent)	Moderate (relative; deployment-dependent)	Low (deployment-dependent)	Low (deployment-dependent)
Fine-tuning / Plugins	No public fine-tuning	Fully customizable (open-weight)	No public fine-tuning	Limited (prompt-based)	Supports fine-tuning and function calling
Safety & Robustness	Provider-aligned safety tuning	User/deployment controlled	Provider-dependent safety	Provider-aligned safety tuning	Provider-aligned safety tuning
Primary Strengths	Strong reasoning and writing, long-context use cases	Open-weight flexibility; on-prem deployment	Reasoning-centric behavior	Fast interactive responses	Mature tooling ecosystem
Ideal Use Cases	Document-heavy workflows, assistants	Research and custom deployments	Reasoning-centric evaluation pipelines	Real-time chat systems	Prototyping and assistants

Table 2: Model Selection and Comparative Specifications

3 Results and Comparison

This section presents a quantitative analysis of how consistently each evaluated language model responds to paraphrased versions of the same opinion questions. Using multiple complementary metrics, we examine both within-model stability and cross-model alignment under controlled rewording conditions.

Figure 1 presents the mean consistency rates across question groups for the five evaluated language models together with 95% confidence intervals. Overall, all models exhibit high consistency under paraphrased versions of the same opinion questions, indicating substantial robustness to superficial prompt rewording. However, measurable differences in consistency are observed across models. DeepSeek Reasoner achieves the highest mean consistency rate (approximately 91%), followed closely by Gemini 2.0 Flash (approximately 90–91%). Both models also show relatively narrow confidence intervals, indicating stable behavior across paraphrase groups. Claude 3.7 Sonnet demonstrates strong performance with a mean consistency rate around 90%, though with slightly higher variability. GPT-3.5 Turbo and LLaMA 3 70B show lower mean consistency and wider uncertainty ranges, suggesting greater sensitivity to paraphrase variation. In practical terms, while all models demonstrate generally robust behavior,

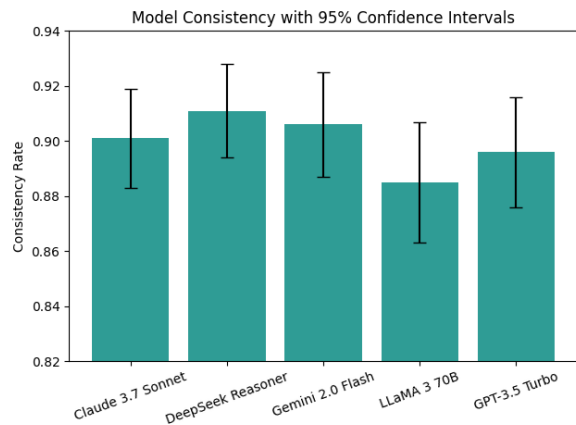


Figure 1: Mean consistency rate across question groups for each evaluated model with 95% confidence intervals. Error bars indicate uncertainty in consistency estimates across paraphrase groups.

DeepSeek Reasoner and Gemini 2.0 Flash exhibit the most stable response patterns under paraphrasing, making them particularly suitable for applications requiring consistent closed-ended decision outputs.

Figure 2 illustrates the pairwise agreement matrix across all models, revealing how frequently two models produce identical responses to the same prompts. As expected, perfect self-agreement appears along the diagonal. Among off-diagonal entries, the highest agreement is observed between GPT-3.5 Turbo and LLaMA 3 70B at 92.9%, in-

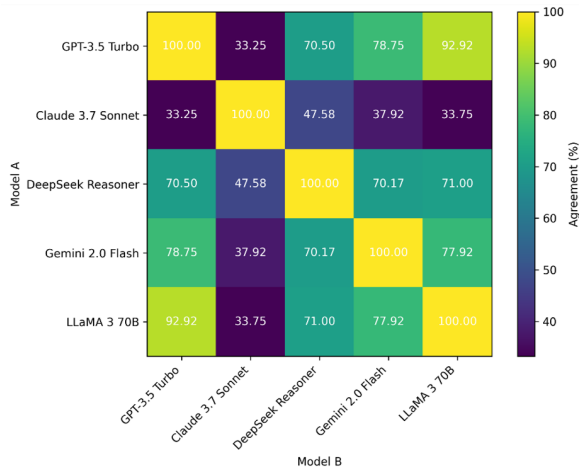


Figure 2: Pairwise Agreement Matrix Between Evaluated Language Models

indicating that these two models frequently reach the same binary conclusions despite originating from different providers. High agreement is also observed between Gemini 2.0 Flash and LLaMA 3 70B, as well as between Gemini 2.0 Flash and GPT-3.5 Turbo, with values in the 77–79% range. These results suggest that Gemini’s decision behavior aligns closely with that of GPT-3.5 Turbo and LLaMA 3 70B. DeepSeek Reasoner demonstrates moderate agreement (approximately 70%) with most other models, indicating broadly similar but less tightly coupled decision boundaries. In contrast, Claude 3.7 Sonnet consistently exhibits low agreement with all other models, ranging from 33% to 48%. This divergence suggests that Claude applies a substantially different internal decision strategy for binary opinion questions. In ensemble settings, Claude’s outputs may therefore require careful weighting or fallback mechanisms to avoid instability arising from inter-model disagreement. Figure 3 reports each model’s self-consistency using Cohen’s Kappa (κ), which measures agreement between responses to original prompts and their paraphrased variants while correcting for chance agreement. Claude 3.7 Sonnet achieves the highest Kappa score ($\kappa = 0.665$), indicating substantial agreement beyond chance and strong internal stability under paraphrasing. DeepSeek Reasoner ($\kappa = 0.638$) and Gemini 2.0 Flash ($\kappa = 0.642$) follow closely, confirming that these models generalize their binary decisions reliably across lexical and syntactic variations. By comparison, LLaMA 3 70B ($\kappa = 0.312$) and GPT-3.5 Turbo ($\kappa = 0.307$) fall within the fair agreement range. While these models still achieve high raw consis-

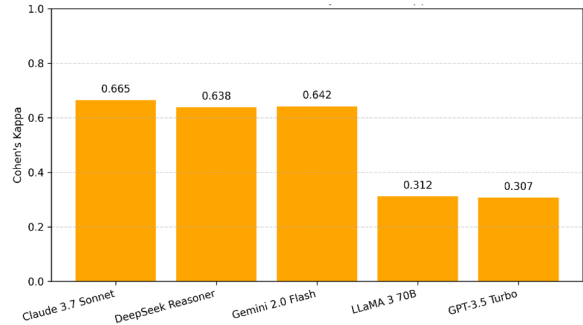


Figure 3: Self-Consistency of Language Model Outputs (Cohen’s Kappa)

tency rates, their lower Kappa values indicate that a larger portion of the observed agreement can be attributed to chance. This suggests greater sensitivity to paraphrase-induced variation when evaluated under a chance-corrected reliability framework, rather than implying direct flip probabilities.

Collectively, these results reveal three distinct robustness profiles:

- **High and stable:** DeepSeek Reasoner and Gemini 2.0 Flash.
- **High but moderately variable:** Claude 3.7 Sonnet and GPT-3.5 Turbo.
- **Lower and highly variable:** LLaMA 3 70B.

For applications requiring predictable binary outputs under paraphrase variation, DeepSeek Reasoner and Gemini 2.0 Flash emerge as the strongest candidates. Claude 3.7 Sonnet and GPT-3.5 Turbo remain viable but may benefit from ensemble voting or post-processing safeguards, while LLaMA 3 70B should be deployed with additional validation mechanisms when phrasing stability is critical.

4 Discussion and Conclusion

This paper presented a systematic evaluation of large language model sensitivity to paraphrased opinion prompts using a controlled dataset and consistency-based evaluation framework. The results demonstrate that while modern LLMs exhibit high overall consistency across paraphrased inputs, measurable variability remains even under semantically equivalent prompt transformations. These findings reinforce the importance of considering prompt formulation as a key factor influencing model behavior in structured decision tasks.

Beyond confirming previously observed prompt sensitivity behavior, this work contributes three primary advances. First, it introduces a new evaluation setting focused specifically on closed-ended, decision-oriented opinion prompts, which remain underexplored compared to open-ended generation and domain-specific robustness studies. Second, it proposes a reproducible evaluation methodology based on controlled paraphrase groups and consistency metrics, enabling systematic cross-model comparison under deterministic inference conditions. Third, the empirical results provide new insight into robustness mechanisms, suggesting that alignment strategies, reasoning specialization, and deployment design contribute more strongly to paraphrase robustness than parameter scale alone.

The publicly released Paraphrased Opinion Prompt Sensitivity (POPS) dataset further supports reproducibility and future research by enabling consistent benchmarking of paraphrase robustness across models and architectures. This resource allows researchers to extend evaluation beyond single-prompt testing and toward structured robustness assessment under semantically controlled variation.

Future work may expand this framework to multilingual settings, open-ended generation tasks, and interactive multi-turn decision scenarios. Additionally, integrating human evaluation and uncertainty-aware benchmarking methods could provide deeper insight into how paraphrase sensitivity influences real-world human–AI decision workflows.

Overall, this work highlights the importance of evaluating LLM robustness not only across tasks and domains, but also across semantically equivalent prompt formulations. Understanding and improving decision stability under paraphrase variation will be critical for reliable deployment of LLMs in real-world decision-support applications.

Limitations

Our evaluation was conducted under a single deterministic inference setting with the temperature fixed to zero, and therefore does not capture how paraphrase robustness may vary under stochastic decoding conditions. In addition, paraphrased prompts were generated by a language model and manually validated, which may not fully reflect the linguistic diversity of paraphrases produced by human experts. Exploring robustness across different inference settings and human-authored paraphrases

is an important direction for future work.

References

- Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational linguistics*, 39(3):463–472.
- Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.
- Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Zehang Deng, Wanlun Ma, Qing-Long Han, Wei Zhou, Xiaogang Zhu, Sheng Wen, and Yang Xiang. 2025. Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica*, 12(5):872–893.
- Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did i do wrong? quantifying llms’ sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558.
- Pınar Ersoy and Mustafa Erşahin. 2024. Benchmarking llama 3 70b for code generation: A comprehensive evaluation. *Orclever Proceedings of Research and Development*, 4(1):52–58.
- Celeste Ci Ying Gue, Noorul Dharajath Abdul Rahim, William Rojas-Carabali, Rupesh Agrawal, Palvannan Rk, John Abisheganaden, and Wan Fen Yip. 2024. Evaluating the openai’s gpt-3.5 turbo’s performance in extracting information from scientific articles on diabetic retinopathy. *Systematic reviews*, 13(1):135.
- Patrick Haller, Jannis Vamvas, and Lena Ann Jäger. 2024. Yes, no, maybe? revisiting language models’ response stability under paraphrasing for the assessment of political leaning. In *First Conference on Language Modeling*.
- Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

- Rabimba Karanjai, Boris Shor, Amanda Austin, Ryan Kennedy, Yang Lu, Lei Xu, and Weidong Shi. 2025. Synthesizing public opinions with llms: Role creation, impacts, and the future to edemocracy. *arXiv preprint arXiv:2504.00241*.
- Soo-Min Kim and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 1367–1373.
- Meltem Kurt Pehlivanoglu, Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la Fuente-Valentín. 2024. Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and t5 language models using a new chatgpt generated dataset: Paragpt. *Expert Systems*, 41(11):e13699.
- Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.
- Hui Ting Lau and Arkaitz Zubiaga. 2025. Understanding the effects of human-written paraphrases in llm-generated text detection. *Natural Language Processing Journal*, page 100151.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunos Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.
- Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Ivan A Uemlianin. 2000. Engaging text: Assessing paraphrase and understanding. *Studies in Higher Education*, 25(3):347–358.
- Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, and 1 others. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural 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: System Demonstrations*, pages 347–355.
- Xiao-Kun Wu, Min Chen, Wanyi Li, Rui Wang, Limeng Lu, Jia Liu, Kai Hwang, Yixue Hao, Yanru Pan, Qingguo Meng, and 1 others. 2025. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, 9(4):87.

The Impact of Highlighting Subjective Language on Perceived News Trustworthiness

Mohammad Shokri¹ Vivek Sharma¹ Emily Klapper²
Elena Filatova³ Shweta Jain⁴ Sarah Ita Levitan²

¹Graduate Center, CUNY ²Hunter College, CUNY
³City Tech, CUNY ⁴John Jay College, CUNY

Abstract

The rise of misinformation and opinionated articles has made understanding how misleading or biased content influences readers an increasingly important problem. While most prior work focuses on detecting misinformation or deceptive language in real time, far less attention has been paid to how such content is perceived by readers, which is an essential component of misinformation’s effectiveness. In this study, we examine whether highlighting subjective sentences in news articles affects perceived trustworthiness. Using a controlled user experiment and 1,334 article–participant ratings, we find that highlighting subjective content produces a modest yet statistically significant decrease in trust, with substantial variation across articles and participants. To explain this variation, we model trust change after highlighting subjective language as a function of article-level linguistic features and reader-level attitudes. Our findings suggest that readers’ reactions to highlighted subjective language are driven primarily by characteristics of the text itself, and that making subjective language explicit plays a role in shaping perceived trustworthiness.

1 Introduction

Trust in media has reached an unprecedented low,¹ even as the internet enables news to circulate with unprecedented speed and reach. This combination has created fertile ground for misinformation and disinformation, which continue to produce serious societal and public health consequences (Vosoughi et al., 2018; Vasist et al., 2024; Ruohonen, 2024). The impact of misleading content is especially pronounced when individuals find it believable.

Believability is closely tied to perceptions of trust, including trust in the author, the outlet, or the

¹<https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx>

narrative being presented. Prior work shows that readers often rely on textual cues to assess whether an article appears balanced, fair, or biased, and these trust judgments strongly shape whether they accept or reject the information (Rodrigo-Ginés et al., 2024; Gabriel et al., 2022).

Understanding how people respond to news as they encounter it is therefore a critical (Gabriel et al., 2022) and understudied challenge. Readers do not process all information uniformly: the same article may elicit skepticism, acceptance, or indifference depending on how it is framed and perceived. Capturing these responses has practical implications for how media platforms allocate fact-checking resources, prioritize interventions, and curate content for different audiences. In particular, expending effort on content that readers easily dismiss as untrustworthy may be less effective than addressing cases where subtle linguistic cues shape perceptions of credibility and persuasion (Babaei et al., 2021).

While much prior work has focused on computational approaches for identifying fake news or manipulative content, including propaganda (Sprenkamp et al., 2023; Sharma et al., 2025), deception (Velutharambath et al., 2024; Rani et al., 2025), and misleading claims in text (Pérez-Rosas et al., 2018; Volkova and Jang, 2018), these approaches largely operate independently of how content is perceived by readers. Yet, as trust judgments are formed during reading an article, understanding which linguistic signals influence these perceptions is critical for explaining why certain content appears credible or persuasive in the first place.

A central factor in these judgments is subjective language, which refers to the presence of opinions, attitudes, or emotionally charged language in the text (Wiebe et al., 2004). Subjective writing may signal author bias, but readers do not always notice such cues on their own. This creates an impor-

tant intersection between misinformation, trust, and subjectivity, as subjective framing can increase the persuasive power of content even when the underlying claim is unsupported, provided that readers interpret it as trustworthy.

In this study, we focus not on determining whether news is factually accurate, but on understanding how readers perceive the trustworthiness of the text they encounter. While detecting subjectivity has been extensively studied in the literature (Wiebe et al., 2004; Antici et al., 2024; Shokri et al., 2024; Elbouanani et al., 2025), far less is known about how making subjectivity explicit affects readers’ trust judgments. We address this gap by investigating whether explicitly highlighting subjective language alters perceived trustworthiness. Specifically, we conduct a controlled within-subjects human study in which subjective sentences in news articles are highlighted and compare participants’ trust ratings before and after this intervention. Through quantitative analysis of the responses, we examine the extent to which subjectivity influences trust perceptions and identify the conditions under which a simple transparency cue can meaningfully affect readers’ evaluations of news content.

2 Related Work

Trust in media has primarily been studied through qualitative approaches, such as interviews and surveys examining how audiences form trust and mistrust in news outlets (Tong, 2024; Garusi and Splendore, 2023; Wenzel et al., 2018). In the psychology literature, trust is commonly defined as a belief in the competence, benevolence, honesty, or predictability of another party, coupled with a willingness to rely on them despite uncertainty or risk (Mayer et al., 1995; McKnight et al., 1998). In the context of online settings and digital media, trust is often defined with respect to the perceived reliability and credibility of information sources and the technology platforms that mediate them. In particular, trust in online news involves not only evaluations of content quality, but also judgments about the integrity of the platform, the authenticity of content producers, and the transparency of information flows (Heuer and Breiter, 2018; Kolo et al., 2022; Lee and Lee, 2023; Cha, 2024).

Relatedly, misinformation and fake news detection have received substantial attention in the NLP community, with recent work focused on

large language models as part of the solution (Ma et al., 2024; Han et al., 2025; Tong et al., 2025; Modzelewski et al., 2025). Earlier approaches examined linguistic differences between legitimate and deceptive news using supervised learning and crowd-sourced datasets (Pérez-Rosas et al., 2018; Rubin et al., 2016). Other methods incorporate fact-checking signals or metadata, such as speaker identity or political affiliation, to improve detection of deceptive or hyper-partisan content (Potthast et al., 2018; Wang, 2017; Long et al., 2017). While these efforts have significantly advanced computational techniques for identifying misinformation, they largely frame the problem as a classification task, offering limited insight into how linguistic cues are perceived by readers or how they shape trust judgments during consumption.

Only a limited number of computational studies have explicitly examined how readers form trust judgments about news, rather than focusing primarily on factual accuracy. In a recent study, Bohacek et al. (2023) introduces a dataset of Czech news articles annotated by humans into four trustworthiness categories: trustworthy, partially trustworthy, misleading, and manipulative. The authors fine-tune a range of transformer-based models on this dataset and report that RoBERTa (Liu et al., 2019) achieves the strongest performance, while overall results indicate substantial room for improvement. In a more reader-focused study, Gabriel et al. (2022) model how readers interpret news headlines, including reactions such as distrust and perceived reliability. A key component of their framework is the perceived label, which captures whether readers interpret a headline as real news or misinformation. Through a controlled user study, the authors show that machine-generated explanations such as inferred writer intent can influence readers’ trust judgments. In particular, these explanations were found to increase trust in real news and decrease trust in misinformation, with significant correlations observed between gold labels and shifts in perceived trustworthiness for certain models.

In contrast to these approaches, we focus on how a specific and interpretable linguistic signal, namely subjectivity, influences perceived trustworthiness during reading. Rather than predicting trust labels or generating explanations, we study the causal effect of highlighting subjective sentences on readers’ trust judgments through a controlled human-subjects experiment.

3 Dataset and Survey

This section describes the datasets used in our study, along with the data collection process, preprocessing steps, and the design of the human-subjects survey.

3.1 Article Selection and Preparation

We draw news articles from the CoAID dataset (Cui and Lee, 2020), a widely used benchmark for misinformation research that aggregates COVID-19-related content from both reliable news outlets and fact-checking sources. All articles used in this study are written in English. The dataset includes true articles collected from cross-verified authoritative outlets, as well as false or misleading articles identified through professional fact-checking organizations.

From the set of true articles, we randomly sampled 100 news articles related to COVID-19 vaccination in the United States. To ensure that articles contained a sufficient amount of subjective language for our intervention, we applied a finetuned RoBERTa-based (Liu et al., 2019) subjectivity classification model to each article and computed the proportion of sentences classified as subjective. We retained articles with a subjectivity ratio greater than 10%, resulting in a final set of 26 articles. Details of the finetuned model and threshold selection are provided in Appendix A.1.

To isolate the effect of textual content on trustworthiness perception, we presented articles in plain text format, removing potentially confounding factors such as author information, news outlet name, images, and advertisements. For each article, we created two versions: an original version and a highlighted version in which sentences annotated as subjective were visually marked using a yellow highlight.

3.2 Annotating Subjective Sentences

To reduce noise from automatic classification and ensure high-quality sentence-level labels, we manually annotated the filtered articles for subjective language. Three members of the research team independently annotated each sentence in the selected articles as either subjective or objective. During this process, annotators did not consistently agree that some articles contained any subjective sentences, and these articles were excluded. As a result, the final set was reduced from 26 to 17 articles. Inter-annotator agreement was assessed us-

ing Krippendorff’s α (Krippendorff, 2018), which accounts for chance agreement among multiple annotators. The resulting Krippendorff’s α agreement score was 0.73, indicating acceptable agreement. Final sentence labels were determined via majority vote among the three annotators.

3.3 Participant Recruitment

We recruited 150 participants through Prolific,² an online crowdsourcing platform. Eligibility criteria required participants to be: (1) native English speakers, (2) born in the United States, and (3) between 18 and 65 years of age. These criteria ensured participants had relevant cultural context for evaluating articles about U.S. COVID-19 policy and healthcare issues.

3.4 Survey Procedure

The survey consisted of three phases. First, participants answered a screening question about their stance on vaccines as a public health measure, with response options ranging from “Strongly Supportive” to “Strongly Opposed.” This question captured participants’ prior beliefs, which could influence their perceptions of article trustworthiness. Second, participants were provided with a definition of trust for the study:

Trust Definition

For this task, trust refers to how fair and unbiased you believe the author is in presenting the information.

Third, participants evaluated 10 randomly selected articles from the pool of 17. We employed a within-subjects design with counterbalanced presentation order: for each participant, 5 randomly selected articles were presented in original-first order (original version rated, then highlighted version), while the remaining 5 articles were presented in highlighted-first order (highlighted version rated, then original version). All ratings used a 4-point scale (1 = no trust, 2 = somewhat untrustworthy, 3 = somewhat trustworthy, 4 = fully trustworthy).

3.5 Data Quality Control

To ensure data quality, we implemented attention checks and excluded responses from participants who completed the survey in less than 6 minutes, which represented the bottom 5% of completion

²<https://www.prolific.com>

times. This threshold was established to filter out participants who may not have carefully read the articles. After applying this criterion, our final dataset consisted of 136 participants and 1334 paired article ratings.

4 Survey Results

This section examines the relationship between highlighting subjective sentences and perceived trustworthiness. We analyze data from 136 participants who evaluated 10 randomly selected news articles, yielding 1,334 paired observations where each participant rated both the original and highlighted versions of the articles.

4.1 Effect of Highlighting Subjectivity

To assess whether highlighting subjective sentences affects perceived trustworthiness, we compared participants’ trust ratings for each article before and after highlighting. We conducted a paired-samples *t*-test and, as a robustness check, a Wilcoxon signed-rank test, which does not assume normally distributed paired differences (Wilcoxon, 1945). Table 1 presents the results. The paired *t*-test revealed a statistically significant negative effect of highlighting subjectivity ($t = -8.42, p < 0.001$), with participants rating highlighted articles as less trustworthy ($\mu = 2.47, \sigma = 1.05$) compared to non-highlighted versions ($\mu = 2.68, \sigma = 1.02$), a mean decrease of 0.21 points. The Wilcoxon signed-rank test corroborated this finding ($p < 0.001$). However, the effect size was small (Cohen’s $d = -0.23$), and two-thirds of ratings remained unchanged, suggesting modest practical impact.

4.2 Order Effect Analysis

To verify that the observed reduction in trust was not attributable to presentation order, we conducted a 2×2 repeated measures ANOVA with version (original vs. highlighted) as a within-subjects factor and presentation order (original-first vs. highlighted-first) as a between-subjects factor. Results confirmed the main effect of highlighting ($F(1, 134) = 21.16, p < 0.001, \eta_p^2 = 0.136$), replicating our paired *t*-test findings. Critically, we found no significant main effect of presentation order ($F(1, 134) = 0.42, p = 0.52, \eta_p^2 = 0.003$) and no version \times order interaction ($F(1, 134) = 0.18, p = 0.67, \eta_p^2 = 0.001$), confirming that the effect of highlighting was consistent regardless of whether participants saw the original or highlighted version first. Separate paired *t*-tests within each

Measure	Original	Highlighted
Mean (SD)	2.68 (1.02)	2.47 (1.05)
Median	3.00	3.00
Statistical Tests		
Paired <i>t</i> -test	$t = -8.42, p < 0.001$	
Wilcoxon test	$p < 0.001$	
Cohen’s d_z	-0.23	
Direction of Change		
Decreased	274 (20.5%)	
No change	879 (65.9%)	
Increased	181 (13.6%)	

Table 1: Effect of highlighting subjective sentences on perceived trustworthiness. Values reflect paired comparisons across $N = 1,334$ participant–article pairs. Cohen’s d_z denotes the paired-samples effect size.

order condition further validated this pattern, with both original-first and highlighted-first conditions showing statistically significant decreases in trust. These results rule out order effects as an alternative explanation and support a causal interpretation of the highlighting intervention.

4.3 Variation Across Articles

While the overall effect of highlighting subjective sentences was negative, Figure 1 reveals substantial heterogeneity across individual articles. The effect ranged from -0.35 (Article 1) to +0.18 (Article 6), a span of 0.53 points on the 4-point scale. Exploratory paired *t*-tests for each article individually showed that 9 of 17 articles exhibited statistically significant effects at $p < 0.05$ (uncorrected), with 4 articles surviving Bonferroni correction for multiple comparisons ($p < 0.003$). Notably, 14 articles showed decreased trustworthiness, while 3 showed increased trustworthiness after highlighting. This variability suggests that the impact of highlighting subjective content is moderated by article-specific characteristics, such as topic, content, writing style, or the nature of the subjective statements themselves. The consistent direction of effects (predominantly negative) combined with the varying magnitude indicates that while highlighting generally reduces perceived trustworthiness, the strength of this effect depends on contextual factors that warrant further investigation.

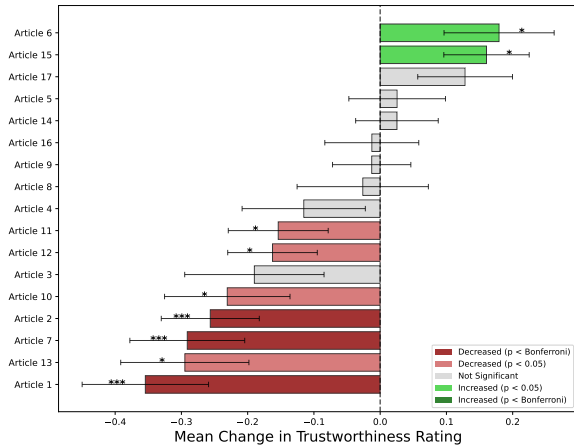


Figure 1: Mean change in trustworthiness ratings by article after highlighting subjective content. Error bars represent standard error of the mean. Asterisks indicate statistical significance from individual paired t-tests: * $p < 0.05$, *** $p < 0.003$ (Bonferroni-corrected). Negative values indicate decreased trustworthiness; positive values indicate increased trustworthiness.

4.4 Variation Across Raters

Beyond article-level effects of highlighting subjective content, we assess variability across participants. Some individuals may be more sensitive to the presence of highlighted subjectivity than others, leading to heterogeneous responses even when the overall trend is negative. Examining this participant-level variation helps clarify whether the observed effect reflects a broadly shared response or is driven by a subset of readers.

Analysis. To examine how consistently participants responded to highlighting, we computed the mean change in trustworthiness for each participant across all articles they rated ($\Delta trust = \text{highlighted} - \text{original}$). Figure 2 shows the distribution of these mean changes, where negative values indicate a reduction in perceived trustworthiness after highlighting.

Results. Across 136 participants, the average change in perceived trustworthiness was small but consistently negative ($M = -0.21$, $SD = 0.34$), indicating that most participants rated highlighted articles slightly less trustworthy than their non-highlighted counterparts. However, there was substantial individual variability (range = $[-0.9, +1.0]$), as illustrated in Figure 2, suggesting that some participants were more sensitive to highlighting than others.

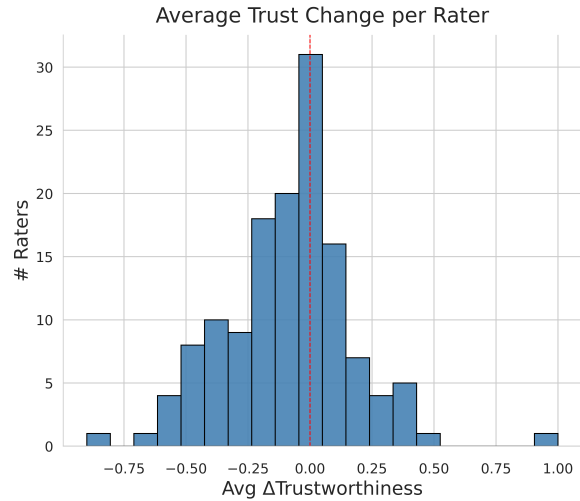


Figure 2: Distribution of mean change in trustworthiness ratings per participant (highlighted – original)

Moderation by Initial Vaccine Stance. To explore whether participants’ prior attitudes toward vaccines moderated their response to highlighted subjectivity, we compared the mean change in trustworthiness ratings ($\Delta trust$) across stance groups. Participants reported their general stance toward vaccines as a public health measure at the beginning of the survey. As shown in Figure 3, the direction of the effect was consistent across groups, with highlighting generally reducing perceived trustworthiness, though the magnitude of the decrease varied by stance. Participants who were somewhat supportive of vaccines exhibited the largest average decline ($M = -0.30$, $SD = 0.24$), followed by somewhat opposed ($M = -0.18$) and strongly opposed ($M = -0.05$) participants, while strongly supportive participants showed no change ($M = 0.00$).

A one-way ANOVA and Kruskal-Wallis test, which assess differences in central tendency across groups under parametric and non-parametric assumptions respectively, did not reveal a statistically significant difference across groups ($p > 0.05$). Nevertheless, the observed pattern is consistent with the possibility that highlighting subjective language has a stronger effect among participants with more moderate positions, while having little influence on those with firmly held views.

5 Features

To analyze factors associated with changes in perceived trustworthiness, we extracted interpretable features from both the articles and the participants

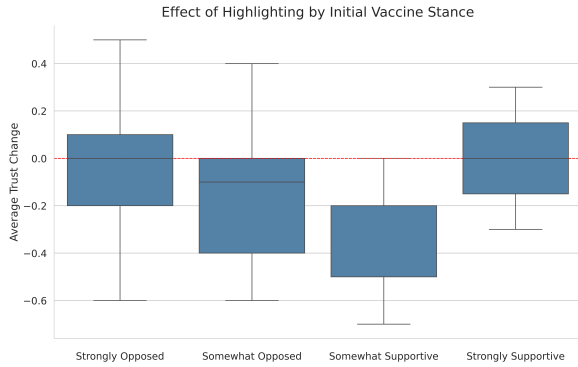


Figure 3: Mean change in perceived trustworthiness (highlighted – original) by participants’ initial stance toward vaccines (Q59). Negative values indicate that highlighting subjective language reduced trust. Although the overall effect of highlighting was modestly negative across all groups, the magnitude of this reduction varied by stance, with the largest decline among participants who were somewhat supportive of vaccines. Error bars represent standard deviation within each stance group.

who rated them. Article-related features capture the linguistic, stylistic, and rhetorical properties of each article, while reader-related features describe participants’ prior attitudes and behavioral tendencies that may form their sensitivity to highlighted subjective sentences.

5.1 Article Related Features

Subjectivity Density. This feature quantifies the proportion of sentences in an article classified as subjective, with values ranging from 0 to 1. It directly captures the extent of subjective language within an article. We hypothesize that articles with higher subjectivity density should elicit stronger trust reductions after highlighting, as the highlighted sentences occupy a greater share of the text.

Sentiment Polarity and Intensity. To capture the affective tone of each article, we computed two sentiment-based metrics: (1) the mean sentiment polarity, ranging from negative to positive, and (2) the average sentiment intensity, defined as the absolute magnitude of emotional valence regardless of sign. We derived these measures using the `cardiffnlp/twitter-roberta-base-sentiment` model (Loureiro et al., 2022), a transformer-based RoBERTa encoder that represents a widely adopted state-of-the-art approach for sentiment analysis in political and news discourse. For each sentence, the model outputs class probabilities for negative, neutral, and positive sentiment. We convert these

probabilities to a continuous polarity score in $[-1, 1]$ by taking the expectation over class labels, and compute intensity as its absolute value. We then average these sentence-level scores to obtain article-level polarity and intensity and scale both features to the 0–1 range via min–max normalization for comparability across articles. Polarity reflects the direction of evaluative language, whereas intensity measures how emotionally charged the text is. We posit that articles with either highly positive or highly negative tone may be perceived as less objective, and thus may undergo greater reductions in perceived trustworthiness when subjective sentences are highlighted.

Persuasiveness. We quantify persuasiveness using a transformer-based regression model that predicts argument quality as a proxy for persuasive strength. Specifically, we apply the `webis/argument-quality-ibm-reproduced` model on Hugging Face, a recent reproduction of IBM’s Project Debater argument-quality predictor (Zelch et al., 2025). The model estimates the perceived quality of written arguments on a continuous scale based on linguistic, syntactic, and discourse-level features learned from the large-scale Argument Quality dataset introduced by Gretz et al. (2020). For each article, we compute the model’s predicted argument-quality score and normalize it to the 0–1 range to produce the persuasiveness score. Higher values indicate greater persuasive strength. This feature captures variation in an article’s argumentative style that may be related to changes in perceived trustworthiness when subjective content is highlighted.

Readability. We measure textual readability using the Flesch–Kincaid Grade Level (Kincaid et al., 1975), a widely used readability index that estimates the years of education required to comprehend a text. The score is computed from two core components: the average number of words per sentence and the average number of syllables per word, following the formulation of Flesch (1948). For each article, we compute the Flesch–Kincaid score and then invert it so that higher values correspond to greater ease of reading. The resulting values are subsequently scaled to the 0–1 range via min–max normalization, where 0 indicates highly complex language and 1 corresponds to simple, accessible writing. Readability functions as a proxy for cognitive load: articles that are easier to read may render

subjective statements more salient, whereas dense or syntactically intricate texts may obscure such cues and thus diminish the effect of highlighting.

Propaganda Techniques. To capture rhetorical strategies beyond sentiment and subjectivity, we annotated each article for the presence of specific propaganda techniques. We follow the taxonomy introduced in SemEval-2020 Task 11 on the Detection of Propaganda Techniques in News Articles (Da San Martino et al., 2020a), which defines a closed set of fine-grained rhetorical devices designed for span-level analysis in news text. This taxonomy includes commonly studied techniques such as Loaded Language, Name Calling or Labeling, Appeal to Fear or Prejudice, and Flag Waving. The definitions of all techniques are provided in Appendix A.2.

Because existing pretrained propaganda detectors are typically optimized for sentence-level span annotation rather than article-level presence, we employed a controlled, instruction-following large language model to perform document-level classification. Prior work has shown that large language models achieve performance comparable to state-of-the-art systems on propaganda detection tasks (Sprenkamp et al., 2023), motivating their use in this setting. Specifically, we prompted the OPENAI GPT-4.1-MINI model (OpenAI, 2024) with a standardized, closed-set annotation prompt requiring JSON-formatted outputs. For each article, the model identified which (if any) techniques appeared at least once, and we encoded these predictions as binary indicators. This approach provides an interpretable representation of rhetorical bias that complements the linguistic and affective features described above. The exact prompt used to obtain these annotations is shown below.

Propaganda Annotation Prompt

You are an expert annotator with extensive experience in analyzing political discourse and identifying rhetorical persuasion strategies. Your task is to determine whether an article contains any of a predefined set of propaganda techniques.

Follow these constraints carefully:

1. Use only the predefined propaganda technique labels.
2. A technique is present if it appears in at least one sentence.

3. Do not invent new labels and do not provide explanations.
4. Output a JSON dictionary of the form:


```
{"techniques_present": ["Loaded Language", "Red Herring", ...]}
```

Examples:

<FEW SHOT EXAMPLES>

Article:

<ARTICLE TEXT>

5.2 Reader Related Features

Stance Toward Vaccines. Participants’ initial attitudes toward vaccines were encoded as an ordinal variable ranging from strongly opposed (1) to strongly supportive (4). Prior stance plays an important role in how individuals interpret and spread misinformation, and may shape how readers respond to subjective language depending on whether it aligns with their beliefs (Derczynski et al., 2017; Hardalov et al., 2022). Stance values were normalized to the 0–1 range for use in the model.

6 Feature Analysis

Our analysis in Section 4 revealed that highlighting subjective language generally reduced perceived trustworthiness, though the magnitude of this effect varied across both articles and raters. To better understand the underlying drivers of this variability, we aim to identify the linguistic and attitudinal features that best explain the observed change in trustworthiness (Δ_{trust}).

We hypothesize that both textual properties and individual differences (Verma et al., 2018) plausibly contribute to Δ_{trust} , so we model the outcome as a weighted linear combination of article-related and reader-related features with separate scaling parameters for each group. Let $\mathbf{x}_j \in \mathbb{R}^p$ denote the article features for article j (e.g., subjectivity density, sentiment, readability, propaganda indicators) and $\mathbf{z}_i \in \mathbb{R}^q$ denote the rater features for participant i (e.g., stance, engagement). The model is expressed as

$$\Delta_{trust_{ij}} = \alpha (\mathbf{w}^\top \mathbf{x}_j) + \beta (\mathbf{v}^\top \mathbf{z}_i) + b, \quad (1)$$

where α and β scale the relative influence of article- and reader-related predictors, \mathbf{w} and \mathbf{v} are the learned feature coefficients within each group, b is the intercept. Continuous predictors are normalized prior to estimation so that α and β are directly

comparable in magnitude. This allows us to assess whether variation in trust change is driven more by article characteristics or by rater attributes.

We fit a linear regression model to estimate $\Delta trust$ using the feature set described above. Article-level predictors included subjectivity density, sentiment polarity and intensity, readability, and a set of one-hot propaganda indicators, while rater-level predictors consisted of each participant’s vaccine stance. To ensure comparability across heterogeneous features, article and rater predictors were standardized separately before fitting the model. The regression jointly estimates the feature weights \mathbf{w} and \mathbf{v} as well as the group-level scaling parameters α and β , allowing us to directly compare the aggregate contribution of textual characteristics to that of rater attributes. The model was trained on all 1,334 participant-article pairs for which both original and highlighted trust ratings were available.

The model reveals a strong asymmetry in predictive influence between article- and participant-related features. The norm of the article-feature coefficients was substantially larger than that of the rater-feature coefficients ($\alpha = 0.34$ vs. $\beta = 0.039$), yielding an α/β ratio of **8.78**. This indicates that article properties account for nearly nine times more variance in $\Delta trust$ than individual differences in vaccine stance. Among article-level predictors, higher subjectivity density was associated with larger decreases in trust, while certain rhetorical devices such as Appeal to Authority and Black-and-White Fallacy showed notable associations with trust change. Participant stance exhibited a smaller but directionally consistent effect, with more pro-vaccine readers showing slightly larger trust reductions. Table 2 summarizes the standardized coefficients for all predictors. To assess the robustness of these estimates, we additionally report 95% bootstrap confidence intervals for the largest-magnitude coefficients in Appendix A.3.

7 Conclusion

In this work, we examined how making subjective language explicit influences readers’ perceptions of news trustworthiness. Through a controlled within-subjects survey, we showed that highlighting subjective sentences generally reduces perceived trust, though the magnitude of this effect varies across articles and individuals. Our feature-based analysis reveals that this variation is driven predomi-

Feature	Coefficient	Type
subj_ratio	-0.1968	Article
Black-and-White Fallacy	+0.1775	Article
polarity_raw	+0.1602	Article
Appeal to Authority	-0.0994	Article
Appeal to Fear or Prejudice	+0.0658	Article
Bandwagon	+0.0430	Article
Glittering Generalities	-0.0358	Article
vaccine_stance	-0.0388	Rater
intensity_raw	+0.0237	Article
fk_raw	+0.0236	Article
Loaded Language	+0.0207	Article
Name Calling or Labeling	+0.0147	Article
Exaggeration or Minimization	-0.0101	Article
readability	-0.0236	Article
Whataboutism	+0.0017	Article
Doubt	+0.0017	Article

Table 2: Standardized regression coefficients for the $\Delta trust$ prediction model, excluding features with zero-valued coefficients. Positive coefficients indicate that the feature is associated with increased trust in the highlighted version relative to the original, while negative coefficients indicate greater trust reduction. Article-level predictors dominate the model, with the aggregate coefficient magnitude of article features ($\alpha = 0.34$) substantially exceeding that of the rater feature ($\beta = 0.039$).

nantly by article-level properties rather than readers’ stance toward the topic. Among the article-related features, subjectivity density and specific rhetorical devices explain substantially more variance in trust change. These findings indicate that trust judgments are sensitive to how linguistic cues are surfaced during reading, even when factual content remains unchanged. This suggests that perception-focused cues may play a role in shaping reader evaluations alongside existing detection and verification efforts.

In light of these results, relying solely on real-time fact checking may be insufficient, since trust judgments often form before such interventions occur. Instead, our results point to the value of perception-focused interventions at the point of reading, complementing existing detection and verification efforts.

Limitations

Our study has several limitations that should be considered when interpreting the results. First, the number of news articles included in the experiment is relatively small. This choice reflects a deliberate trade-off driven by budgetary and statistical considerations. Because our primary objective is to measure how readers’ trust judgments change in response to highlighting subjective language, reli-

able estimation of this effect requires a sufficiently large number of pre- and post-highlighting ratings for each article. Given a fixed budget, increasing the number of articles would necessarily reduce the number of evaluations per article. We prioritized depth of evaluation, resulting in approximately 75 ratings per article, at the cost of broader article coverage. Future work with larger budgets could extend this design to a wider range of articles while maintaining adequate per-article sampling.

Second, our feature analysis assumes a linear relationships between predictors and changes in perceived trustworthiness. While this choice supports interpretability and enables direct comparison between article-level and reader-level influences, it may bias the analysis toward simple relationships and overlook nonlinear effects or interactions between features. More flexible modeling approaches could capture richer patterns in how linguistic cues and reader characteristics jointly shape trust judgments.

Third, the set of features considered in this study is necessarily limited. Article-level predictors focus on a subset of interpretable linguistic and rhetorical properties, while reader-level features are restricted primarily to prior stance toward vaccines. Although this design facilitates transparent analysis, it does not exhaust the range of factors that may influence trust perceptions. Incorporating broader representations of discourse structure, narrative framing, or additional reader attributes may provide a more comprehensive account of trust formation.

References

- Francesco Antici, Federico Ruggeri, Andrea Galassi, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2024. A corpus for sentence-level subjectivity detection on english news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 273–285.
- Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. 2021. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, 9(3):839–850.
- Matyas Bohacek, Michal Bravansky, Filip Trhlík, and Václav Moravec. 2023. Czech-ing the news: Article trustworthiness dataset for czech. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 96–109.
- Jiyoung Cha. 2024. Predictors of the credibility of social media as a news outlet: An examination of the influences of social media contacts, source perceptions, and media use. *International Journal on Media Management*, 26(1-2):68–93.
- Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *CoRR*.
- Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1377–1414.
- Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Yu Seunghak, Roberto Di Pietro, Preslav Nakov, and 1 others. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI-20}*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization.
- Lingjia Deng and Janyce Wiebe. 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.
- Akram Elbouanani, Evan Dufraisie, Aboubacar Tuo, and Adrian Popescu. 2025. Cea-list at checkthat! 2025: evaluating llms as detectors of bias and opinion in text. *arXiv preprint arXiv:2507.07539*.
- Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.
- Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers’ reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.
- Diego Garusi and Sergio Splendore. 2023. Advancing a qualitative turn in news media trust research. *Sociology compass*, 17(4):e13075.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Asaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument

- quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.
- Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. [Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15125–15140, Suzhou, China. Association for Computational Linguistics.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.
- Hendrik Heuer and Andreas Breiter. 2018. Trust in news on social media. In *Proceedings of the 10th Nordic conference on human-computer interaction*, pages 137–147.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
- Castulus Kolo, Joschka Mütterlein, and Sarah Anna Schmid. 2022. Believing journalists, ai, or fake news: The role of trust in media. In *HICSS*, pages 1–10.
- Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. Sage Publications.
- Seung Yeop Lee and Sang Woo Lee. 2023. Normative or effective? the role of news diversity and trust in news recommendation services. *International Journal of Human-Computer Interaction*, 39(6):1216–1229.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)*, pages 252–256.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.
- Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. [On fake news detection with LLM enhanced semantics mining](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.
- Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.
- D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review*, 23(3):473–490.
- Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. [PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.
- OpenAI. 2024. [GPT-4o mini: advancing cost-efficient intelligence](#). Technical report.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3391–3401.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 231–240.
- Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2025. Sepsis: I can catch your lies—a new paradigm for deception detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 97–128.
- Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.
- Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.
- Jukka Ruohonen. 2024. A comparative study of online disinformation and offline protests. *SN Social Sciences*, 4(12):232.

- Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314.
- Vivek Sharma, Mohammad Mahdi Shokri, Sarah Ita Levitan, Elena Filatova, and Shweta Jain. 2025. Analysis of propaganda in tweets from politically biased sources. *arXiv preprint arXiv:2507.08169*.
- Mohammad Shokri, Vivek Sharma, Elena Filatova, Shweta Jain, and Sarah Levitan. 2024. **Subjectivity detection in English news using large language models**. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 215–226, Bangkok, Thailand. Association for Computational Linguistics.
- Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.
- Jingrong Tong. 2024. From content to context: A qualitative case study of factors influencing audience perception of the trustworthiness of covid-19 data visualisations in uk newspaper coverage. *Journalism*, 25(7):1481–1499.
- Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290.
- Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2024. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, 26(2):663–688.
- Aswathy Velutharambath, Amelie Wüthrl, and Roman Klinger. 2024. How entangled is factuality and deception in german? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9538–9554.
- Nitin Verma, Kenneth R Fleischmann, and Kolina S Koltai. 2018. Demographic factors and trust in different news sources. *Proceedings of the Association for Information Science and Technology*, 55(1):524–533.
- Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Andrea Wenzel, Anthony Nadler, Melissa Valle, and Marc Lamont Hill. 2018. Listening is not enough: Mistrust and local news in urban and suburban philly. *Columbia Journalism Review*.
- Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.
- Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Ines Zelch, Matthias Hagen, Benno Stein, and Johannes Kiesel. 2025. Reproducing the argument quality prediction of project debater. In *Proceedings of the 12th Argument mining Workshop*, pages 181–188.

A Appendix

A.1 Training a Subjectivity Classifier

To filter articles with a subjective sentence ratio greater than 10%, we trained a sentence-level subjectivity classifier based on RoBERTa (Liu et al., 2019). The model was trained on two subjectivity-annotated datasets: the MPQA opinion corpus (Deng and Wiebe, 2015) and a recently released dataset containing sentence-level subjectivity annotations for COVID-19 and crime-related news (Savinova and Del Prado, 2023).

MPQA. The MPQA (Multi-Perspective Question Answering) dataset is a widely used resource in sentiment analysis and opinion mining. It is designed to capture the multifaceted nature of subjective language and contains text segments annotated for sentiment polarity and subjectivity. The dataset includes content from multiple sources, such as news articles, product reviews, discussion forums, and social media, reflecting the varied contexts in which subjective expressions occur. To control for genre-related variation, we restrict our experiments to MPQA sentences drawn from news articles. We use version three of the MPQA opinion corpus.

News Dataset. The second dataset consists of sentences from news articles and Facebook posts related to *crime* and *COVID-19*, published by four major UK news outlets, for a total of 7,751 sentences (Savinova and Del Prado, 2023). We exclude Facebook posts, as they are shorter and may not be authored by journalists. Consequently, all experiments in this paper use only news sentences, yielding a dataset of 2,973 sentences, including 1,013 subjective and 1,960 objective sentences.

Training Details. We fine-tuned a RoBERTa-base model for binary sentence-level subjectivity classification using the Adam optimizer with a learning rate of 1×10^{-5} . Training was performed for 10 epochs with standard cross-entropy loss. Across the two training datasets, the classifier achieved an average F1 score of 0.88. The resulting model was used to estimate the proportion of subjective sentences in each article for the purpose of article filtering.

A.2 Propaganda Techniques

Below is a list of the definitions for propaganda techniques we used for propaganda detection in news articles (Da San Martino et al., 2020b):

- Loaded language: Using words/phrases with strong emotional implications (positive or negative) to influence an audience.
- Name calling or labeling: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises.
- Repetition: Repeating the same message over and over again, so that the audience will eventually accept it.
- Appeal to fear: support an idea by instilling fear against other alternatives
- Exaggeration or minimization: Either representing something in an excessive manner: making things larger, better, worse.
- Doubt: Questioning the credibility of someone or something. - appeal to fear/prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on pre-conceived judgments.
- Flag-waving: Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea.
- Causal oversimplification: Assuming one cause when there are multiple causes behind an issue. We include scapegoating as well which is defined as the transfer of the blame to one person or group of people without investigating the complexities of an issue.
- Slogans: A brief and striking phrase that may include labeling and stereotyping.
- Appeal to authority: Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence. Include the special case where the reference is not an authority/expert, although it is referred to as testimonial in the literature.
- Black-and-white fallacy: Presenting two alternative options as the only possibilities, when in fact more possibilities exist, eliminating any other possible choice. and as an extreme telling the audience exactly what actions to take, which is also called as dictatorship.

- **Thought-terminating cliché:** Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought.
- **Whataboutism:** Discredit an opponent’s position by charging them with hypocrisy without directly disproving their argument.
- **Reductio ad hitlerum:** Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.
- **Red herring:** Introducing irrelevant material to the issue being discussed, so that everyone’s attention is diverted away from the points made.
- **Bandwagon:** Attempting to persuade the target audience to join in and take the course of action because “everyone else is taking the same action”.
- **Obfuscation/intentional vagueness/confusion:** Using deliberately unclear words, so that the audience may have its own interpretation."
- **Straw men:** Refuting arguments that were not presented.

A.3 Confidence Intervals for Feature Coefficients

To quantify uncertainty in the feature-level effects reported in Section 6, we estimate 95% confidence intervals for the regression coefficients using non-parametric bootstrap resampling. Specifically, we repeatedly resample the participant-article pairs with replacement and refit the linear model described in Equation (1) on each resampled dataset. For each feature, we compute percentile-based confidence intervals from the resulting empirical distribution of coefficients.

Figure 4 shows the bootstrap means and 95% confidence intervals for the features with the largest absolute coefficients. Subjectivity density exhibits a consistently negative association with changes in perceived trustworthiness, with confidence intervals that exclude zero, indicating a stable effect across bootstrap samples. In contrast, many

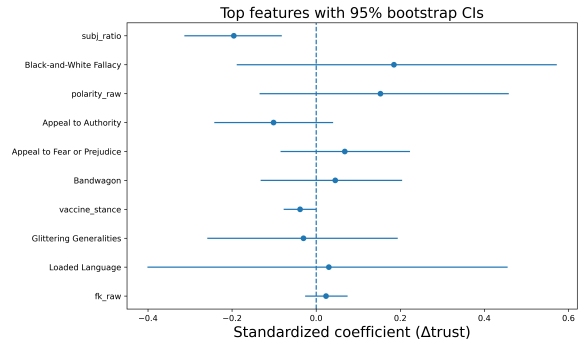


Figure 4: Bootstrap means and 95% confidence intervals for the largest-magnitude regression coefficients predicting changes in perceived trustworthiness (Δtrust). Error bars indicate percentile-based confidence intervals estimated via 1,000 bootstrap resamples. Positive values indicate increased trust in the highlighted version relative to the original, while negative values indicate greater trust reduction.

individual rhetorical features display wider confidence intervals that often include zero, reflecting substantial variability across articles and contexts. Reader-level stance remains small and tightly bounded around zero, reinforcing the conclusion that variation in trust change is driven primarily by article-level properties rather than individual differences.

Appraisal Trajectories in Narratives Reveal Distinct Patterns of Emotion Evocation

Johannes Schäfer, Janne Wagner, and Roman Klinger

Fundamentals of Natural Language Processing, University of Bamberg, Germany
{johannes.schaefer, roman.klinger}@uni-bamberg.de,
janne.wagner@stud.uni-bamberg.de

Abstract

Understanding emotion responses relies on reconstructing how individuals appraise events. While prior work has studied emotion trajectories and inherent correlations with appraisals, it has considered appraisals only in a snapshot analysis. However, because appraisal is a complex, sequential process, we argue that it should be analyzed based on how it unfolds throughout a narrative. In this study, we investigate whether trajectories of appraisals are distinctive for different emotions in five-event stories – narratives where each of five sentences describes an event. We employ zero-shot prompting with a large language model to predict appraisals on sub-sequences of a narrative. We find that this approach is effective in identifying relevant appraisals in narratives, without prior knowledge of the evoked emotion, enabling a comprehensive analysis of appraisal trajectories. Furthermore, we are the first to quantitatively identify typical patterns of appraisal trajectories that distinguish emotions. For example, a rising trajectory for self-responsibility indicates trust, while a falling trajectory suggests anger.

1 Introduction

Emotion analysis seeks to uncover the emotions that events evoke in individuals based on their textual descriptions. Foundational frameworks established by Ekman (1972) and Plutchik (2001) categorize emotions in a structured manner. This allows researchers to label textual data from various sources, for example social media as in the corpus by Mohammad (2012). Consider the textual examples given¹ in Table 1, listing five sentences in narrative order. We may suspect that the event described in sentence #1 evoked the emotion trust in the experiencer and that the event described in

sentence #5 probably evoked the emotion guilt in them (based on the given context).

Recent work by Troiano et al. (2023) has highlighted the importance of implicit emotion cues tied to appraisal theories. These theories present emotions as resulting from the cognitive evaluation of events, incorporating subjective feelings, action tendencies, physiological responses, and both vocal and bodily expressions triggered by those events (Staller and Petta, 2001; Scherer, 2005; Gratch et al., 2009). By identifying the appraisals, we can more clearly reconstruct how the emotions are shaped in the example (see Table 1): The event described in sentence #1 is likely something pleasant for the experiencer and they are taking on responsibility in connection with a relevant goal. These cognitive appraisals (pleasantness, self responsibility and goal relevance) result in the emotion trust.

Prior work mainly analyzes appraisal in isolated instances, such as single sentences, discrete events, or social media posts (e.g., Hofmann et al., 2020; Stranisci et al., 2022; Zhan et al., 2023; Troiano et al., 2023; Liu et al., 2025; Zhou et al., 2025). Notably, many of these studies build upon the work of Troiano et al. (2023) and extend their focus to longer texts, underscoring the need for a contextualized analysis.

#	Sentence Text
1	A close friend entrusted me with setting up the sound system for a charity event.
2	I faced a daunting task, as the previous technician had left the equipment in disarray.
3	With time running out, I took a shortcut to meet the deadline, skipping some crucial safety checks.
4	Just before the event started, my friend reminded me of its significance and the many people counting on its success.
5	The loudspeaker suddenly malfunctioned and went silent.

Table 1: Example narrative comprising five sentences.

¹We provide access to the data and code of our experiments on <https://www.uni-bamberg.de/en/nlproc/resources/appraisal-trajectories/>.

Appraisals and the resulting emotions often unfold as a complex process (Lewis, 2001) that can span multiple events and develop over the course of a narrative. In dialog research, emotion dynamics have been explored, e.g., by Poria et al. (2019). Debnath et al. (2025) also consider how appraisals influence these dynamics in multi-turn conversations. However, our work is the first aiming to understand how the appraisal progression of an individual develops based on narrative event sequences. To explain the emotion guilt for sentence #5 (shown in Table 1) using appraisals, we need to consider how these develop over the course of the narrative event sequence: The individual’s responsibility for the outcome becomes increasingly apparent throughout the narrative. However, the pleasantness of the events, which is positive in the beginning, shifts to negative by the end due to the event described in the final sentence. Consequently, the negative emotion guilt is suddenly triggered. We refer to these changes in appraisals over the course of an event sequence as appraisal trajectories. Prior research only conducts snapshot analyses of appraisals, neglecting their development over time. While Wemmer et al. (2024) created appraisal trajectories in a study of dreams, their analysis does not examine the connection to the evoked emotions.

In our study, we explore the dynamics of appraisals that evoke different emotions within narratives through the following research questions:

1. Can we identify appraisals from narrative text alone, or is prior knowledge of the evoked emotion necessary?
2. Do appraisal trajectories in narratives show distinct patterns for different emotions?

To this end, we conduct automatic appraisal annotation on short narratives consisting of sequences of event descriptions evoking different emotions. Our methodology utilizes zero-shot prompting with a large language model (LLM) to predict appraisals. We analyze the trajectories of different appraisal dimensions in these narratives and compare their role in evoking different emotions. We aim to identify typical patterns of appraisal trajectories associated with specific emotions, to understand how emotion responses can be modeled more precisely.

2 Methods

Data and Model. For our experiments, we require a dataset with a substantial number of narra-

tives categorized for different emotions. Given a person’s textual description of their experience, we aim to reconstruct how they appraise events resulting in that emotion. The Emotional Backstories (EBS) dataset (Schäfer and Klinger, 2025) therefore offers itself as it comprises 13,000 narratives, each structured as a sequence of five event descriptions and assigned to one of 13 emotion categories (see the first column of Table 2). The dataset consists of LLM-generated data, with its quality validated by human annotators, and is designed for a thorough investigation of how contexts influence emotion analysis. For modeling appraisals, we adopt the 21 categories used by Troiano et al. (2023) (a detailed list is given in Appendix A). Our analysis uses the instruction-tuned Llama-3.3 LLM with 70B parameters (Meta, 2025). We utilize a zero-shot prompting method due to the lack of annotated training data for contextualized appraisals. The full text prompts used are given in Appendix B.

Appraisal Detection Performance. Our goal is to compute and analyze appraisals, requiring an effective labeling approach. Our method prompts the LLM to predict Likert scale scores (ranging from 1 to 5) for all appraisal categories given a narrative. We estimate the contextualized appraisal prediction performance of this approach by evaluating it on the crowd-enVent dataset (Troiano et al., 2023), which only contains isolated instances.

Appraisal Correlations with Emotions. We also need to evaluate the effectiveness of our appraisal prediction method on narratives, specifically assessing whether it performs successfully without prior knowledge of the associated emotion, relying solely on the text. We test this by allowing for the inclusion or exclusion of emotion information in the prompt. Troiano et al. (2023) show typical correlations of certain appraisals with certain emotions. We investigate whether we can find similar correlations in narratives and compare two settings: 1) Bottom-Up: the model should reconstruct the appraisals of the experiencer based on the textual sequence of event descriptions. 2) Top-Down: the model is informed about the overall emotion being evoked and is basically tasked to find related appraisals in the narrative. We hypothesize that the second approach will reveal clearer correlations due to the model’s inherent knowledge of appraisals, allowing it to predict these not only from the textual content. However, the first setting presents a more realistic scenario, as we usually

		Delta Scores																			
Appraisal \ Emotion	suddenness	familiarity	predict_event	pleasantness	unpleasantness	goal_relevance	chance_responsibl	self_responsibl	other_responsibl	predict_conseq	goal_support	urgency	self_control	other_control	chance_control	accept_conseq	standards	social_norms	attention	not_consider	effort
	Anger	0.2	-0.2	-0.2	-0.5	0.7	0.1	-0.1	-0.3	0.2	-0.1	-0.6	0.2	-0.2	0.2	0.0	-0.6	1.0	0.7	0.1	0.0
Boredom	-0.2	0.1	0.1	-0.5	-0.1	-0.5	0.0	-0.1	-0.1	-0.1	-0.4	-0.3	-0.1	0.0	0.0	-0.1	-0.1	0.0	-0.7	0.8	-0.4
Disgust	0.1	-0.2	-0.1	-0.4	0.7	0.0	0.0	-0.1	0.1	0.0	-0.4	0.2	-0.1	0.0	0.0	-0.4	0.6	0.5	0.1	0.5	0.3
Fear	0.1	-0.2	-0.1	-0.3	0.4	0.2	0.1	0.0	0.0	0.0	-0.3	0.3	-0.1	0.1	0.3	-0.4	0.4	0.2	0.1	0.1	0.2
Guilt	0.1	-0.1	-0.1	-0.5	0.6	0.3	-0.1	0.5	-0.4	0.0	-0.6	0.2	-0.1	-0.1	0.0	-0.6	1.1	0.5	0.1	0.4	0.3
Joy	-0.1	0.0	0.0	0.8	-0.6	0.0	0.0	0.1	0.0	0.1	0.6	-0.1	0.1	0.0	-0.1	0.4	-0.2	0.0	-0.1	0.0	-0.2
Pride	-0.1	0.0	0.0	0.6	-0.4	0.2	-0.1	0.3	-0.2	0.2	0.6	0.0	0.3	-0.1	-0.1	0.5	-0.1	0.0	0.0	0.0	0.0
Relief	0.0	0.0	0.0	0.6	-0.5	0.1	0.0	0.1	-0.1	0.1	0.7	0.0	0.0	0.0	0.1	0.9	-0.1	0.0	0.0	0.0	-0.2
Sadness	0.1	-0.1	-0.1	-0.7	0.7	0.1	0.1	-0.1	-0.1	0.0	-0.6	0.1	-0.2	0.1	0.2	-0.5	0.4	0.1	0.1	0.3	0.3
Shame	0.1	-0.2	-0.2	-0.6	0.7	0.1	0.0	0.3	-0.3	-0.1	-0.8	0.1	-0.4	0.1	0.0	-0.7	1.4	0.7	0.1	0.8	0.3
Surprise	0.8	-0.5	-0.4	-0.1	0.0	0.1	0.2	-0.1	0.0	-0.2	-0.1	0.3	-0.1	0.1	0.3	-0.1	0.2	0.1	0.2	0.0	0.1
Trust	-0.1	0.1	0.0	0.2	-0.2	0.1	-0.1	0.0	0.0	0.1	0.3	0.0	0.1	0.0	0.0	0.2	-0.1	0.0	0.0	0.0	-0.1
No-Emotion	-0.2	0.1	0.1	-0.1	-0.2	-0.3	0.0	-0.1	-0.1	0.0	-0.2	-0.1	0.0	-0.1	-0.1	0.0	-0.2	-0.1	-0.2	-0.1	-0.2

Table 2: Difference in appraisal scores of complete narratives per emotion category of model informed about the emotion to model not informed about the emotion (absolute scores for the two methods are given in Appendix D). Positive (negative) values, marked with green (purple) background color, express that the informed model assigns a higher (lower) Likert score (range 1–5). The different appraisal categories are further specified in Appendix A.

do not have prior knowledge of the emotion that is evoked. In comparing the predictions from the two approaches, we aim to explore whether the typical correlations can still emerge in the case where the model is solely dependent on the text, without explicit information about the emotion.

Analysis of Appraisal Trajectories. To be able to analyze the trajectories of appraisals within narratives, we generate sub-sequences from the five-sentence text instances. We start with only the first sentence and then incrementally add sentences in narrative order, employing the technique by [Wemer et al. \(2024\)](#) who study dreams. By including the previous context rather than just a sentence, we aim for a more accurate contextualized analysis. For each sub-sequence, we predict appraisal scores using the model established above where no information about the emotion is included.

3 Results

The evaluation on the human-annotated crowd-enVent dataset (see Appendix C) shows that the appraisal prediction performance of our zero-shot model is acceptable (RMSE 1.49/1.46) in comparison to the fine-tuned model by [Troiano et al. \(2023\)](#) (RMSE 1.40). We analyze appraisals in narratives by applying our model to the Emotional Backstories dataset ([Schäfer and Klinger, 2025](#)), which contains 1,000 five-sentence-narratives for

each of 13 emotion categories. In the following, we address our research questions by evaluating averages over the instances per emotion category as well as overall macro-averaged scores.

3.1 Can we identify appraisals from narrative text alone, or is prior knowledge of the evoked emotion necessary?

Table 2 shows how predicted appraisal scores change when the model is informed about the evoked emotion. The results reveal only minor variations in predicted appraisal scores. For instance, the pleasantness appraisal score (see the fifth column in Table 2) slightly increases (+0.2 to +0.8 Likert score points) for positive emotions while it slightly decreases (−0.3 to −0.7 Likert scale points) for negative emotions when the model is informed about the emotion. Differences in Likert scale values of more than ±1 are observed only for the appraisal regarding standards in negative emotions (+1.0 for anger, +1.1 for guilt and +1.4 for shame; see Table 2).

Separate values for the two settings further substantiate the trend that informing the model about the emotion is not necessary to find typical correlations of appraisals and emotions (see Appendix D). This shows that our approach is able to find indicators of appraisals in the textual content of the narratives even without prior knowledge of the emotion.

3.2 Do appraisal trajectories in narratives show distinct patterns for different emotions?

Table 3 shows average appraisal scores of narrative sub-sequences, analyzed as trajectories for each emotion. From these values, we can infer how certain appraisals are expressed through the narrative structure. Generally, differences in magnitude are observed across certain appraisals. For example, relevance to the experiencer’s goals is consistently scored higher than chance occurrences, aligning with expectations for experiencer-perspective text. Other appraisal categories show differences in intensity for different emotions, for example, contrasting familiarity scores; sadness exhibits higher scores compared to fear – which is consistent with established emotion-appraisal correlations.

Regarding trajectory patterns, we identify four distinct types in Table 3: Rising, Falling, Valley-shaped, Hill-shaped. These shapes arise from the variations in appraisal processes as they unfold within narrative structures. Certain appraisals display uniform patterns across all emotions. An example for this is suddenness: all Rising. Here, context is required for higher scores. In contrast, certain appraisals interestingly show variation in trajectory patterns by emotion. For example, pleasantness shows a Rising pattern for joy but a Valley-shaped pattern for relief, illustrating the complex process that involves less pleasant events before a resolution. Another example is self-responsibility, which exhibits a Falling trajectory for anger, while showing a Rising pattern for trust, yet both ultimately reach approximately the same score. This emphasizes that performing a snapshot analysis of appraisals is incomplete, as only the distinct trajectory patterns reveal deeper processes that distinguish the evocation of the different emotions.

4 Conclusion

Our study reveals distinct trajectories in appraisal processes that result in evoking different emotions in narratives. We demonstrate that contextualized appraisals can be effectively modeled based on narratives even without prior knowledge of the underlying emotion, suggesting that the textual content contains sufficient cues. Our identification of unique appraisal patterns – Rising, Falling, Valley-shaped, and Hill-shaped – highlights the complexity of the appraisal process leading to different emotion responses, emphasizing the need for

a dynamic perspective on appraisal. These findings carry further implications for emotion detection applications, as recognizing appraisal trajectories could enhance the accuracy of emotion classification systems.

The fixed five-sentence narrative structure we study facilitates a controlled analysis. Our findings generalize to narratives of varying lengths, provided that an additional step is taken to adapt such data into the format we use. Another option would be to align and normalize trajectories for variable-length narratives, which would further broaden their applicability, although this remains a topic for future research. Additionally, future work should study the interplay between different combinations of appraisals and how they relate to specific emotion categories. Future research is required to explore how to best leverage these trajectories to optimize emotion detection in various contexts, ultimately deepening our understanding of the interplay between appraisals in narrative structure and evoked emotions.

Limitations

This study has some limitations worth noting. The zero-shot performance of the LLM for predicting appraisals can be nuanced; while it demonstrates potential, there seems to be considerable room for improvement. Moreover, generating Likert-scale numerical outputs with the LLM might not be an optimized approach for scoring appraisals, as it limits calibration and comparability. Analyzing output distributions and sensitivity across prompt variants or repeated runs would strengthen the robustness of our findings, as we only ran our experiments once. We also utilize only one LLM for our analysis, which may restrict the generalizability of our findings. Different models may yield varying results in predicting appraisals. Additionally, our analysis relies on data that has been automatically generated, which can include biases or inaccuracies inherent in the dataset. Human validation could, for example, strengthen the reliability of our results. To facilitate reproducibility, we provide access to our code and model predictions, enabling further exploration and validation of our findings.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project KL2869/1–2 (CEAT, project number 380093645).

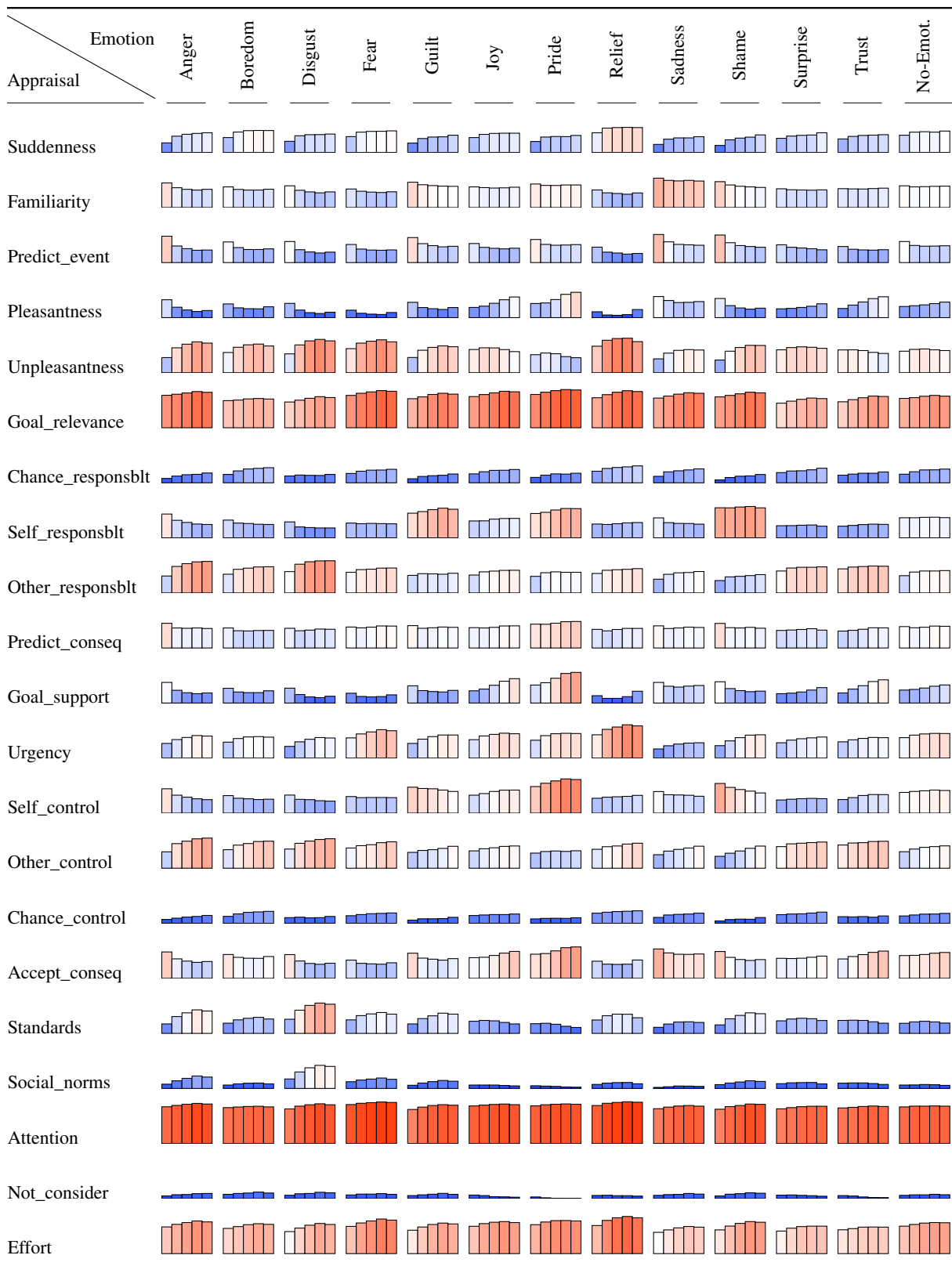


Table 3: Averaged predicted appraisal trajectories in narratives per emotion categories in the EBS dataset. Each cell presents five appraisal values computed from sub-sequences of data instances (five sentences each), reflecting the average over 1000 instances. The bar color and height both represent the average Likert scale score from the range of 1 to 5, with blue indicating the lowest (1), white the midpoint (3), and red the highest (5) values.

References

- Alok Debnath, Yvette Graham, and Owen Conlan. 2025. [An appraisal theoretic approach to modelling affect flow in conversation corpora](#). In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 233–250, Vienna, Austria. Association for Computational Linguistics.
- Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation 1971*, volume 19. Lincoln University of Nebraska Press.
- Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. 2009. [Assessing the validity of appraisal-based models of emotion](#). In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8.
- Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. [Appraisal theories for emotion classification in text](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Marc D Lewis. 2001. Personal pathways in the development of appraisal. In Klaus R Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 205–220. Oxford University Press: London.
- June M. Liu, He Cao, Renliang Sun, Rui Wang, Yu Li, and Jiaying Zhang. 2025. [CAPE: A Chinese dataset for appraisal-based emotional generation in large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6291–6309, Albuquerque, New Mexico. Association for Computational Linguistics.
- Meta. 2025. [Llama \(model llama-3.3-70b-instruct\)](#). Large language model.
- Saif Mohammad. 2012. [#emotional tweets](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.
- Robert Plutchik. 2001. [The nature of emotions](#). *American Scientist*, 89(4):344–350.
- Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. [Emotion recognition in conversation: Research challenges, datasets, and recent advances](#). *IEEE Access*, 7:100943–100953.
- Klaus R Scherer. 2005. [What are emotions? and how can they be measured?](#) *Social science information*, 44(4):695–729.
- Johannes Schäfer and Roman Klinger. 2025. [Shaping event backstories to estimate potential emotion contexts](#). *Preprint*, arXiv:2508.09954.
- Alexander Staller and Paolo Petta. 2001. [Introducing emotions into the computational study of social norms: A first evaluation](#). *Journal of Artificial Societies and Social Simulation*, 4(1).
- Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. [APPReddit: a corpus of Reddit posts annotated for appraisal](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.
- Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. [Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction](#). *Computational Linguistics*, 49(1):1–72.
- Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. [EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia. ELRA and ICCL.
- Hongli Zhan, Desmond C. Ong, and Junyi Jessy Li. 2023. [Evaluating subjective cognitive appraisals of emotions from large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446, Singapore. Association for Computational Linguistics.
- Yuxiang Zhou, Hainiu Xu, Desmond C. Ong, Maria Liakata, Petr Slovak, and Yulan He. 2025. [Modeling subjectivity in cognitive appraisal with language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13811–13833, Suzhou, China. Association for Computational Linguistics.

A Appraisal Categories

Table 4 lists the 21 appraisal categories and defining statements used in our experiments.

B Prompts

Table 5 shows the full text of the prompts we use for appraisal analysis (two variants based on whether we want to inform the model about the emotion or not). The interaction with the model uses two message types²: a “system” message that establishes the context for the interaction and includes general guidelines; a “user” message that encapsulates the

²<https://www.llama.com/docs/model-cards-and-prompt-formats/llama3/>

Abbreviation	Appraisal Statement
suddenness	The event was sudden or abrupt.
familiarity	The event was familiar to its experiencer.
predict_event	The experiencer could have predicted the occurrence of the event.
pleasantness	The event was pleasant for the experiencer.
unpleasantness	The event was unpleasant for the experiencer.
goal_relevance	The experiencer expected the event to have important consequences for him/herself.
chance_responsblt	The event was caused by chance, special circumstances, or natural forces.
self_responsblt	The event was caused by the experiencer's own behavior.
other_responsblt	The event was caused by somebody else's behavior.
predict_conseq	The experiencer anticipated the consequences of the event.
goal_support	The experiencer expected positive consequences for her/himself.
urgency	The event required an immediate response.
self_control	The experiencer was able to influence what was going on during the event.
other_control	Someone other than the experiencer was influencing what was going on.
chance_control	The situation was the result of outside influences of which nobody had control.
accept_conseq	The experiencer anticipated that he/she could live with the unavoidable consequences of the event.
standards	The event clashed with her/his standards and ideals.
social_norms	The actions that produced the event violated laws or socially accepted norms.
attention	The experiencer had to pay attention to the situation.
not_consider	The experiencer wanted to shut the situation out of her/his mind.
effort	The situation required her/him a great deal of energy to deal with it.

Table 4: Appraisal categories used in our experiments.

specific inputs, requirements, and instructions for the task.

C Appraisal Prediction Performance on Isolated Instances

We evaluate the appraisal prediction performance of our zero-shot model on the crowd-enVent test data (Troiano et al., 2023). Results are shown in Table 7 (model not informed about emotion) and Table 6 (model informed about emotion). The predicted Likert scale points are evaluated with RMSE (Root Mean Square Error) against the gold annotations. Additionally, a classification prediction is simulated by converting predictions of 1/2/3 to a negative label, and 4/5 to a positive label (method used by Troiano et al., 2023). The result is then evaluated with Precision (P), Recall (R) and F1-score (F1) against the gold annotations.

D Predicted Appraisals on Narratives by Emotion Category

We predict appraisal categories as Likert scale scores (range 1–5) for each of the narratives from the Emotional Backstories dataset (Schäfer and Klinger, 2025) and compute average scores for each set of narratives per emotion category. Ta-

ble 8 displays the results for the scores predicted by the model which has not been informed about the emotion. Table 9 displays the results for the scores predicted by the model which has been informed about the emotion.

MT	Prompt Text
user	You are an expert in the appraisal on event descriptions.
system	<p>A person describes their event experience as follows: {text_instance}. Based on this text, you're asked to assess the content in terms of the appraisals it conveys. Thus, your task is to identify appraisals in texts.</p> <p>You will receive a list of appraisals and their description. For each of the following 21 appraisal assessments listed provide a score in the range of 1 (not at all) to 5 (extremely).</p> <ol style="list-style-type: none"> 1. The event was sudden or abrupt (suddenness) 2. The event was familiar to its experiencer. (familiarity) 3. The experiencer could have predicted the occurrence of the event. (predict_event) 4. The event was pleasant for the experiencer. (pleasantness) 5. The event was unpleasant for the experiencer. (unpleasantness) 6. The experiencer expected the event to have important consequences for him/herself. (goal_relevance) 7. The event was caused by chance, special circumstances, or natural forces. (chance_responsblt) 8. The event was caused by the experiencer's own behavior. (self_responsblt) 9. The event was caused by somebody else's behavior. (other_responsblt) 10. The experiencer anticipated the consequences of the event. (predict_conseq) 11. The experiencer expected positive consequences for her/himself. (goal_support) 12. The event required an immediate response. (urgency) 13. The experiencer was able to influence what was going on during the event. (self_control) 14. Someone other than the experiencer was influencing what was going on. (other_control) 15. The situation was the result of outside influences of which nobody had control. (chance_control) 16. The experiencer anticipated that he/she could live with the unavoidable consequences of the event. (accept_conseq) 17. The event clashed with her/his standards and ideals. (standards) 18. The actions that produced the event violated laws or socially accepted norms. (social_norms) 19. The experiencer had to pay attention to the situation. (attention) 20. The experiencer wanted to shut the situation out of her/his mind. (not_consider) 21. The situation required her/him a great deal of energy to deal with it. (effort) <p>As your response, provide only a json object and keep the appraisal labels as they are (eg. "other_responsblt").</p> <p>In addition to the given text, take into account the emotion: {emotion}. To what level were the appraisals evoked in the person at the end? Take into account both the text and the emotion.</p> <p>In your answer, only provide the score labels you have chosen as the output together with the input text and emotion. Provide the output as a json object with the top-level key being 'appraisal' and the low-level keys being the concrete appraisals. The value must be the Likert-score.</p> <p>For example, given a text instance, your json object output should be:</p> <pre>{ "chain": "{text_instance}", "emotion": "null{emotion}", "appraisals": {...} }</pre> <p>where [...] is a placeholder and you insert the 21 appraisals here.</p> <p>The chain is: {text_instance}.</p> <p>What are the appraisal scores for this chain? Only output the json object.</p>

Table 5: Prompts used for zero-shot appraisal analysis. “MT” refers to the prompt message type as specified in the input for the instruction-tuned LLM. “{text_instance}” and “{emotion}” are variables. The text marked in red is only inserted when the model should not be informed about the emotion. The text marked in blue is only inserted if the model should be informed about the emotion.

Appraisal	Ours	Troiano et al. (2023)	Δ	Ours			Troiano et al. (2023)			Δ
	RMSE	RMSE	RMSE	P	R	F1	P	R	F1	F1
Suddenness	1.44	1.33	-.11	.72	.65	.68	.70	.79	.74	.06
Familiarity	1.54	1.42	-.12	.51	.64	.57	.77	.82	.79	.22
Predict_event	1.45	1.47	.03	.59	.55	.57	.76	.74	.75	.18
Pleasantness	1.42	1.30	-.12	.79	.89	.84	.88	.87	.88	.04
Unpleasantness	1.18	1.26	.08	.87	.85	.86	.79	.80	.80	-.06
Goal_relevance	1.62	1.57	-.05	.53	.91	.67	.73	.69	.71	.04
Chance_responsblt	1.55	1.43	-.12	.61	.37	.46	.83	.87	.85	.39
Self_responsblt	1.37	1.40	.03	.72	.69	.70	.81	.77	.79	.09
Other_responsblt	1.50	1.57	.07	.80	.66	.72	.74	.72	.73	.01
Predict_conseq	1.46	1.50	.04	.59	.44	.50	.67	.71	.69	.19
Goal_support	1.50	1.33	-.17	.68	.71	.69	.80	.82	.81	.12
Urgency	1.77	1.43	-.34	.77	.30	.44	.63	.60	.61	.17
Self_control	1.46	1.35	-.11	.59	.56	.57	.78	.81	.79	.22
Other_control	1.63	1.36	-.27	.79	.58	.67	.64	.60	.62	-.05
Chance_control	1.55	1.35	-.20	.59	.29	.39	.84	.90	.87	.48
Accept_conseq	1.77	1.36	-.41	.52	.60	.55	.63	.65	.64	.09
Standards	1.30	1.34	.04	.69	.65	.67	.82	.83	.82	.15
Social_norms	1.13	1.44	.31	.59	.58	.58	.90	.95	.92	.34
Attention	1.43	1.27	-.16	.68	.87	.77	.50	.48	.48	-.29
Not_consider	1.52	1.53	.01	.60	.26	.37	.83	.71	.77	.40
Effort	1.37	1.38	.01	.67	.66	.66	.69	.70	.70	.04
Macro average	1.46	1.40	-.06	.66	.60	.62	.75	.75	.75	.13

Table 6: Appraisal prediction performance of our zero-shot model (setting with mention of emotion in prompt) on the crowd-enVent test data in comparison to the fine-tuned model by Troiano et al. (2023).

Appraisal	Ours	Troiano et al. (2023)	Δ	Ours			Troiano et al. (2023)			Δ
	RMSE	RMSE	RMSE	P	R	F1	P	R	F1	F1
Suddenness	1.55	1.33	-.22	.76	.60	.67	.70	.79	.74	.07
Familiarity	1.61	1.42	-.19	.49	.65	.56	.77	.82	.79	.23
Predict_event	1.50	1.47	-.03	.58	.59	.59	.76	.74	.75	.16
Pleasantness	1.14	1.30	.16	.78	.86	.82	.88	.87	.88	.06
Unpleasantness	1.26	1.26	-.00	.87	.81	.84	.79	.80	.80	-.04
Goal_relevance	1.58	1.57	-.01	.55	.87	.68	.73	.69	.71	.04
Chance_responsblt	1.59	1.43	-.16	.65	.44	.52	.83	.87	.85	.33
Self_responsblt	1.41	1.40	-.01	.72	.68	.70	.81	.77	.79	.09
Other_responsblt	1.55	1.57	.02	.79	.68	.73	.74	.72	.73	-.00
Predict_conseq	1.54	1.50	-.04	.57	.41	.48	.67	.71	.69	.21
Goal_support	1.47	1.33	-.14	.68	.68	.68	.80	.82	.81	.13
Urgency	1.88	1.43	-.45	.79	.28	.41	.63	.60	.61	.20
Self_control	1.50	1.35	-.15	.58	.58	.58	.78	.81	.79	.21
Other_control	1.68	1.36	-.32	.78	.57	.66	.64	.60	.62	-.04
Chance_control	1.56	1.35	-.21	.62	.35	.44	.84	.90	.87	.43
Accept_conseq	1.76	1.36	-.40	.52	.60	.56	.63	.65	.64	.08
Standards	1.30	1.34	.04	.74	.54	.62	.82	.83	.82	.20
Social_norms	1.11	1.44	.33	.62	.52	.56	.90	.95	.92	.36
Attention	1.43	1.27	-.16	.68	.87	.76	.50	.48	.48	-.28
Not_consider	1.57	1.53	-.04	.75	.18	.29	.83	.71	.77	.48
Effort	1.39	1.38	-.01	.69	.61	.65	.69	.70	.70	.06
Macro average	1.49	1.40	-.09	.68	.59	.61	.75	.75	.75	.14

Table 7: Appraisal prediction performance of our zero-shot model (setting without mention of emotion in prompt) on the crowd-enVent test data in comparison to the fine-tuned model by Troiano et al. (2023).

Emotion	Appraisal																				
	Suddenness	Familiarity	Predict_event	Pleasantness	Unpleasantness	Goal_relevance	Chance_responsibl	Self_responsibl	Other_responsibl	Predict_conseq	Goal_support	Urgency	Self_control	Other_control	Chance_control	Accept_conseq	Standards	Social_norms	Attention	Not_consider	Effort
Anger	2.9	2.7	2.2	1.7	3.8	4.3	1.9	2.3	4.0	2.8	1.9	3.1	2.3	3.8	1.7	2.6	3.1	2.1	4.7	1.5	4.0
Boredom	3.1	2.7	2.3	2.0	3.5	3.7	2.4	2.3	3.5	2.6	2.2	3.0	2.3	3.5	2.1	3.0	2.3	1.4	4.4	1.5	3.7
Disgust	2.7	2.5	2.0	1.5	4.0	3.8	1.8	1.9	4.0	2.8	1.6	2.9	2.1	3.8	1.6	2.4	3.7	3.1	4.6	1.5	3.7
Fear	3.0	2.5	2.2	1.5	3.9	4.4	2.3	2.3	3.3	3.1	1.8	3.6	2.5	3.5	2.0	2.5	2.8	1.9	4.9	1.4	4.2
Guilt	2.6	3.0	2.5	1.9	3.4	4.2	1.8	3.7	2.8	2.9	2.2	3.2	3.0	3.1	1.6	2.8	2.8	1.7	4.6	1.4	3.8
Joy	2.8	2.9	2.4	2.9	3.0	4.4	2.2	2.8	3.1	3.1	3.3	3.3	3.2	3.1	1.9	3.5	1.9	1.2	4.7	1.1	3.9
Pride	2.6	3.1	2.7	3.4	2.4	4.5	1.9	3.7	2.9	3.5	3.9	3.3	4.2	2.6	1.5	4.0	1.6	1.1	4.7	1.0	4.1
Relief	3.3	2.4	1.9	1.8	3.9	4.4	2.6	2.4	3.3	2.8	2.1	4.0	2.7	3.4	2.2	2.7	2.5	1.4	4.9	1.2	4.4
Sadness	2.5	3.5	2.6	2.5	3.1	4.2	2.3	2.3	3.0	2.9	2.7	2.4	2.6	3.1	1.9	3.3	2.0	1.2	4.5	1.4	3.5
Shame	2.7	2.9	2.4	1.9	3.6	4.3	1.8	3.8	2.7	2.9	2.1	3.2	2.9	3.1	1.5	2.8	2.9	1.7	4.7	1.5	3.9
Surprise	2.9	2.7	2.2	2.3	3.3	3.7	2.4	2.1	3.5	2.7	2.4	3.0	2.4	3.5	2.1	3.1	2.2	1.4	4.5	1.2	3.6
Trust	2.7	2.8	2.2	3.0	2.8	3.9	2.0	2.3	3.5	2.9	3.2	2.9	2.7	3.5	1.7	3.6	2.0	1.4	4.5	1.1	3.5
No-Emotion	3.0	3.0	2.6	2.5	3.1	4.0	2.3	2.9	3.1	3.0	2.7	3.3	3.1	3.1	1.9	3.4	2.0	1.3	4.5	1.4	3.9

Table 8: Appraisal scores of narratives per emotion category of model **not** informed about the emotion. Positive (negative) values are marked with red (blue) background color.

Emotion	Appraisal																				
	Suddenness	Familiarity	Predict_event	Pleasantness	Unpleasantness	Goal_relevance	Chance_responsibl	Self_responsibl	Other_responsibl	Predict_conseq	Goal_support	Urgency	Self_control	Other_control	Chance_control	Accept_conseq	Standards	Social_norms	Attention	Not_consider	Effort
Anger	3.0	2.5	2.0	1.2	4.5	4.4	1.8	2.0	4.2	2.8	1.4	3.3	2.1	4.0	1.7	2.0	4.1	2.7	4.9	1.5	4.3
Boredom	2.8	2.8	2.4	1.5	3.4	3.2	2.5	2.2	3.4	2.6	1.7	2.7	2.2	3.5	2.2	2.9	2.2	1.4	3.8	2.3	3.3
Disgust	2.8	2.3	1.9	1.1	4.6	3.9	1.8	1.8	4.1	2.8	1.2	3.1	2.0	3.8	1.7	2.0	4.4	3.5	4.8	2.0	4.0
Fear	3.2	2.3	2.1	1.2	4.3	4.6	2.4	2.3	3.3	3.0	1.4	3.9	2.3	3.5	2.2	2.1	3.2	2.1	4.9	1.5	4.4
Guilt	2.7	2.9	2.4	1.4	4.0	4.4	1.7	4.2	2.5	2.9	1.6	3.3	2.9	3.0	1.5	2.3	3.9	2.2	4.7	1.8	4.0
Joy	2.8	2.9	2.4	3.7	2.3	4.4	2.2	2.9	3.1	3.2	3.9	3.2	3.3	3.0	1.8	4.0	1.7	1.2	4.6	1.1	3.7
Pride	2.5	3.1	2.7	4.0	2.0	4.7	1.8	4.1	2.8	3.7	4.5	3.3	4.4	2.5	1.4	4.4	1.4	1.1	4.6	1.0	4.1
Relief	3.3	2.4	1.9	2.4	3.4	4.5	2.6	2.5	3.2	2.9	2.8	4.1	2.7	3.3	2.2	3.6	2.4	1.4	4.9	1.3	4.2
Sadness	2.5	3.4	2.6	1.8	3.9	4.4	2.4	2.2	2.9	2.9	2.0	2.5	2.4	3.1	2.2	2.8	2.4	1.2	4.6	1.7	3.7
Shame	2.7	2.7	2.3	1.3	4.3	4.4	1.7	4.1	2.4	2.7	1.3	3.3	2.5	3.2	1.6	2.1	4.2	2.4	4.8	2.3	4.2
Surprise	3.7	2.2	1.8	2.2	3.3	3.8	2.6	2.0	3.4	2.4	2.4	3.3	2.2	3.5	2.3	3.0	2.4	1.5	4.7	1.2	3.7
Trust	2.6	2.9	2.2	3.2	2.6	4.0	1.9	2.3	3.5	3.0	3.5	2.9	2.9	3.5	1.6	3.8	1.9	1.4	4.4	1.1	3.4
No-Emotion	2.9	3.1	2.7	2.3	2.9	3.7	2.2	2.8	3.0	3.0	2.5	3.2	3.1	3.0	1.9	3.4	1.8	1.2	4.3	1.3	3.7

Table 9: Appraisal scores of narratives per emotion category of model informed about the emotion. Positive (negative) values are marked with red (blue) background color.

Exploring Subjective Tasks in Farsi: A Survey Analysis and Evaluation of Language Models

Donya Rooein¹, Flor Miriam Plaza-del-Arco^{1,2}, Debora Nozza¹, Dirk Hovy¹

¹Bocconi University, Milan, Italy

²Leiden University

Correspondence: donya.rooein@unibocconi.it

Abstract

Given Farsi’s speaker base of over 127 million people and the growing availability of digital text, including more than 1.3 million articles on Wikipedia, it is considered a “middle-resource” language. However, this label quickly crumbles when the situation is examined more closely. We focus on three subjective tasks (Sentiment Analysis, Emotion Analysis, and Toxicity Detection) and identify significant challenges in data availability and quality, despite overall increases in data availability. We review 110 publications on subjective tasks in Farsi and observe a lack of publicly available datasets. Furthermore, existing datasets often lack essential demographic factors, such as age and gender, that are crucial for accurately modeling subjectivity in language. When evaluating prediction models using the few available datasets, the results are highly unstable across both datasets and models. Our findings show that the volume of data alone is insufficient to improve a language’s standing in NLP.

1 Introduction

Many NLP tasks, like emotion classification, are inherently subjective. There are different valid perspectives on the “correct” data labels. How emotions are perceived, for example, differs between the sender and the receiver’s subjective interpretations (Barz et al., 2025). The same message, expressing frustration or sarcasm, could be interpreted humorously by one individual but taken offensively or negatively by another, depending on their cultural background, personal experiences, or situational context.

Subjective tasks in NLP, such as emotion analysis, sentiment analysis, and toxic detection, have received increasing attention because they directly affect societal aspects, including decision-making, customer feedback, product evaluation,

and the understanding of social dynamics in general (Nandwani and Verma, 2021). These tasks involve assigning texts to specific emotions or sentiments that best reflect the author’s mental or emotional state (Tao and Fang, 2020). Recent surveys in emotion and sentiment analysis (Murthy and Kumar, 2021; Kusal et al., 2022; Singh Tomar et al., 2023; Hung and Alias, 2023; Venkit et al., 2023; Al Maruf et al., 2024; Plaza-del Arco et al., 2024; Song et al., 2025) have primarily focused on identifying available datasets, reviewing models, exploring detection techniques across various modalities (e.g., visual, vocal, textual), and discussing applications. These studies focus on English and do not consider other languages such as Farsi¹.

Language technologies play a crucial role in promoting multilingualism and preserving linguistic diversity worldwide. However, many languages still face challenges in resource availability, particularly for subjective tasks, despite having substantial digital resources and peer-reviewed research. This is the case for Farsi, which has over 1.3 million Wikipedia articles² and has been classified by Joshi et al. (2020) as a language with a strong web presence but insufficient efforts in labeled data collection, ranking just below high-resource languages. Despite these resources, **research on subjective tasks in Farsi remains notably scarce**, making it a low-resource language in this domain.

The Farsi script, also known as the Perso-Arabic script, is a modified form of the Arabic alphabet (Izadi et al., 2006). Persian and its dialects are official languages in Iran, Afghanistan, and Tajikistan, and are also spoken by minority communities in Uzbekistan, Turkmenistan, Azerbaijan, Armenia, Georgia, and southern

¹Also known as Persian.

²https://en.wikipedia.org/wiki/Persian_Wikipedia

Russia. While Farsi and Arabic scripts are often considered similar due to their shared roots, they differ in their alphabets and their writing styles (Izadi et al., 2006).

While a few survey studies in Farsi focus on sentiment analysis and discuss resource limitations and methodological developments (Rajabi and Valavi, 2021; Asgarnezhad and Monadjemi, 2021; Borowczyk, 2023), to the best of our knowledge, no existing work provides a comprehensive survey of multiple subjective tasks in Farsi. The scope of this Paper includes evaluating different encoder-only models and LLMs across three key tasks: emotion analysis (EA), sentiment analysis (SA), and toxic detection (TD). These tasks span a range of applications and research communities, from traditional sentiment analysis of product reviews to offensive-language detection. This list is by no means exhaustive for subjective language tasks such as humor detection, sarcasm detection, and metaphor detection. Rather, we highlight the areas with the most dataset availability in the literature we surveyed. This gap is particularly concerning in the era of LLMs, where these systems are not only widely accessible but also increasingly used for subjective discussions (Ouyang et al., 2023). It is essential to evaluate their ability to understand and process sentiment and emotion in Farsi, as well as to assess their handling of toxicity to ensure safe and responsible interactions. The lack of research in this area underscores the urgent need for a focused exploration to ensure that Farsi, like other languages, benefits from advancements in subjective NLP.

We collect relevant studies from publications drawn primarily from ACL Anthology³, and complemented by additional searches on Google Scholar⁴. We report the available dataset for each task, including important metadata such as dataset size, labels, and source. Additionally, we evaluate various language models on selected datasets to assess their capabilities for these subjective tasks in Farsi.

We present the following key contributions:

- A detailed survey of publications, datasets, and resources specific to the three subjective tasks in Farsi: sentiment analysis, emotion analysis, and toxicity detection.

³<https://aclanthology.org/anthology+abstracts.bib>

⁴<https://scholar.google.com/>

- An experimental evaluation of encoder-only multilingual models and open-source LLMs on these tasks in Farsi.
- An analysis of the impact of text translation as a potential solution to address low-resource challenges in Farsi language.

2 Background

Subjective tasks such as EA, SA, and TD often pose unique challenges due to their reliance on context, cultural nuances, and linguistic features. The EA involves classifying emotions expressed in a text (e.g., joy, sadness, anger) (Alm et al., 2005). For instance, recognizing the nuanced difference between Farsi expressions like “دلش گرفت” (delash gereft, literally “his/her heart became tight”) conveying sadness, versus “دارد دلشوره” (delshooreh dārad, literally “he/she has a salty heart”) depicting anxiety, requires deep cultural and contextual understanding compared to relatively straightforward English expressions like “feeling sad” or “feeling anxious”. The SA consists of determining the sentiment polarity of a text, typically positive, negative, or neutral (Wilson et al., 2005). For example, the Persian expression “جای تو خالیه” (jāye to khālie, literally “your place is empty”) carries a positive sentiment, often implying affection, inclusion, and the speaker expresses a desire for the listener’s presence. However, translated directly into English, it may suggest loneliness, absence, or even negativity. Such examples underscore the importance of accurately capturing sentiment, which requires sensitivity to cultural context and linguistic nuances. Toxicity detection consists of identifying language or content considered harmful, offensive, abusive, hateful, or otherwise inappropriate (Pavlopoulos et al., 2020). The interpretation of what constitutes toxic content often varies significantly based on cultural and societal norms. For example, the phrase “کمه عقلمت” (“you’re not very smart”) in Farsi might be considered mildly humorous among close friends but is perceived as offensive in formal or public contexts.

3 A Survey on NLP Studies Covering Subjective Tasks in Farsi

To identify relevant papers with resources related to For EA, SA, and TD tasks in Farsi, we design a structured search query comprising

three main components: <Task>, <Dataset>, and <Language>⁵. The <Task> component includes the three NLP tasks we explore: the EA, SA, and TD. To ensure a comprehensive selection of studies for these tasks, we identify papers whose titles or abstracts include keywords associated with each task. For the EA task, our query includes the terms “emotion classification”, “emotion detection”, “emotion recognition”, “emotion analysis”, and “emotion prediction”. For the SA task, we incorporate the following keywords: “polarity classification”, “sentiment classification”, and “sentiment analysis”. Lastly, for the TD task, we use terms including “hate speech detection”, “offensive language detection”, “hate speech classification”, “offensive language classification”, “toxic detection”, and “toxic classification”. The <Dataset> component includes related terms, i.e., “data set,” “dataset,” “corpus”, and “corpora”. Finally, the <Language> component explicitly focuses on language-related terms, namely “Farsi” and “Persian”. Our query variations are derived from 5 keywords associated with the EA, 3 with the SA, and 6 with the TD tasks (a total of 14 keywords), combined with 4 dataset formulation strategies and 2 for the language, yielding a final total of 112 unique phrase searches. To further expand our search, we also collect publications using only <Task> and <Language>, adding 28 additional search phrases. In total, we executed 140 unique phrase searches.

We identify 12 unique papers from the ACL Anthology: eight focused on SA, four on EA, and none on the TD task. This absence indicates the lack of research and publicly available datasets on Farsi toxicity detection in the ACL Anthology. To expand our search results, we also use Google Scholar. Google Scholar lists papers from different research databases; however, it is difficult to verify all the returned sources. We use the SerpApi⁶ library to retrieve papers from Google Scholar. To limit the search results from this engine, we configure the SerpApi to only return the top 10 relevant papers for a given search keyword. This limitation allows us to verify their publishers manually. This search strategy adds 98 more papers which 40 from arXiv⁷, 16 from IEEE⁸,

⁵All searches are updated by March 2025.

⁶<https://serpapi.com/>

⁷<https://arxiv.org/>

⁸<https://www.ieee.org/>

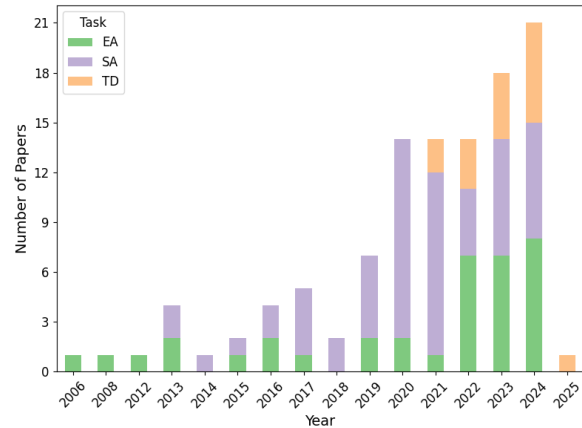


Figure 1: Distribution of papers considered in our survey by year and tasks (EA: Emotion Analysis, SA: Sentiment Analysis, and TD: Toxicity Detection).

12 from Springer⁹, and 30 from other publishers.

Thus, we have a total of 110 papers¹⁰(only 11% from ACL Anthology), with 36 papers for EA, 58 papers for SA, and 16 papers for TD¹¹. Figure 1 shows the statistics of the collected papers published from 2006 to 2025. The SA task represents the largest share at 52.7% (58 out of 110) of all papers, followed by EA at 32.7% (36 out of 110). The EA task began among the non-NLP community in 2006, focusing on EA through speech. The number of publications remained low in the early years; however, by 2024, the EA in Farsi had increased to 8 papers incorporating text-based modalities. The TD task, which did not appear until 2021, already accounts for nearly 14.5% (16 out of 110) of papers by 2025, indicating that TD is becoming an increasingly important area of research in NLP Farsi.

3.1 Annotation Criteria

After identifying relevant papers, we conduct a manual annotation to summarize and categorize the papers based on consistent criteria. The motivation here is to identify publicly available datasets in Farsi for each task. We adopt the annotation framework proposed by Plaza-del Arco et al. (2024), which suggests surveying EA datasets based on five key aspects: annotation framework, language, multimodal, content source, and dataset size. We expand this framework

⁹<https://www.springer.com/>

¹⁰The list of the reviewed papers is available at https://anonymous.4open.science/r/subjective_tasks_farsi-8874/README.md

¹¹Three of these papers are in the Farsi language and were published at local conferences within Iran.

to all the considered subjective tasks and include additional details: lexicon, the type of classification task (e.g., binary, multiclass, or multilabel), and, specifically for studies involving dataset creation, whether the demographics of annotators are explicitly considered. We also include information on the availability of datasets used in each paper.

Our annotation results reveal several trends. For the data modalities, most works (86.4%) are text-based, a few (4.5%) combine text with speech, and 8.2% focus on speech-only datasets. In addition, only one paper (0.9%) uses acoustic and visual data. The datasets used in the reviewed papers are from social media platforms, e-commerce websites, and specialized corpora. The most frequently used sources for social media for all tasks are X¹² (previously Twitter) and Instagram¹³. The e-commerce source is mostly Digikala¹⁴, Iran’s largest online retail platform, which contains extensive user-generated product reviews that are valuable for sentiment analysis. Additional sources include datasets from Booking.ir¹⁵, a popular platform for hotel reviews, movie review comments¹⁶. In some cases, authors use specialized resources such as radio plays or collect datasets from surveys of specific populations. Based on our review collection for data availability, we identify three categories of papers: (I) papers without datasets, (II) papers with datasets that are not publicly available, and (III) papers with publicly available datasets. We identify 17 of the 36 EA papers as dataset papers, but only 7 provide publicly available datasets. In particular, 4 of these 7 datasets are from the ACL Anthology. For SA, we identify 33 dataset papers, but only 5 datasets are available (3 from the ACL Anthology). Finally, TD has 14 papers on datasets, 3 of which are publicly available, and none from the ACL Anthology. In total, we therefore identify 15 publicly available datasets across the three tasks. In the following sections, we provide further details on the available datasets for each task, along with their characteristics.

3.2 Datasets

Table 1 presents a list of publicly available datasets along with detailed information on their names, label sources, data sources, sizes, and modalities.

EA datasets: We identify seven datasets for EA. The **Shemo** (Yazdani et al., 2021) dataset is derived from radio plays and annotates five primary emotions, i.e., anger, fear, happiness, sadness, and surprise along with a neutral category, comprising 3,000 samples. This dataset is the only dataset with both text and speech modality, and the rest of the datasets are text-only. **ShortPersianEmo** (Sadeghi et al., 2021) is from comments on the Digikala website, an e-commerce platform in Iran. The **SAT** (Elahimanesh et al., 2023) dataset originates from chatbot conversations and distinguishes a broader spectrum of emotions (happy, angry, anxious, ashamed, disappointed, disgusted, envious, guilty, insecure, loving, sad, and jealous) across 5,600 samples. The SAT dataset also includes the demographic information (age and gender) of participants. **ArmanEmo** (Mirzaee et al., 2022) and **LetHerLearn** (Hussiny and Øvrelid, 2023), **EmoPars** (Sabri et al., 2021a) consist of tweets annotated with common emotions such as anger, fear, sadness, happiness, and either wonder or surprise. In particular, EmoPars is annotated by a multilabel annotation approach, assigning a numerical value between 0 and 5 to each emotion (anger, fear, happiness, hatred, sadness, and wonder). None of these datasets fully adhere to well-known frameworks for emotion analysis such as Ekman’s framework (Ekman et al., 1999) which includes anger, fear, sadness, joy, disgust, and surprise or Plutchik’s model (Plutchik, 1982), which encompasses eight primary emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. **LearnArmanEmo** (Hussiny et al., 2024) combines ArmanEmo and LetHerLearn by unifying their labels based on Ekman’s framework. In this unified approach, the label “happiness” is used instead of “joy”, and an “other” category is added to capture emotions outside the defined set.

SA datasets: Pars-ABSA (Shangipour ataei et al., 2022), Persian Digikala (Kobari et al., 2023), and Persian-English code-mixed datasets (Sabri et al., 2021b) categorize sentiment of Farsi sentences into positive, negative, and neutral

¹²<https://x.com/>

¹³<https://www.instagram.com/>

¹⁴<https://www.digikala.com/>

¹⁵<https://www.booking.ir/>

¹⁶From <https://cinematicket.org/>

Task	Dataset	Mult.	Labels	Source	Size	Included
EA	Shemo	T, S	E + [neutral]	radio plays	3,000	-
EA	ShortPersianEmo	T	[happiness, sadness, anger, fear, other]	e-commerce	5,472	-
EA	SAT	T	E + [anxious, ashamed, disappointed, envious, guilty, insecure, loving, jealous]	chatbot conv.	5,600	-
EA	ArmanEmo	T	E - [disguss] + [hate, other]	social media	7,000	✓
EA	LetHerLearn	T	E + [other]	social media	7,600	✓
EA	LearnArmanEmo	T	E + [other]	social media	14,880	-
EA	EmoPars	T	E - [disgust] + [hatred]	social media	30,000	✓
SA	SentiPers	T	[-2, -1, 0, +1, +2]	e-commerce	15,683	✓
SA	Pers.-Eng.	T	[negative, neutral, positive]	social media	3,640	-
SA	Persian Digikala	T	[negative, neutral, positive]	e-commerce	34,465	-
SA	Pars-ABSA	T	[negative, neutral, positive]	e-commerce	10,002	✓
SA	MirasOpinion	T	[-1, 0, +1]	e-commerce	93,868	✓
TD	Phate	T	[hateful (violence, hate, vulgar), normal]	social media	7,056	✓
TD	PHICAD	T	[hate/offense, obscene, spam, none]	social media	300,000	✓
TD	Pars-OFF	T	[offensive, not-offensive]	social media	8,334	✓

Table 1: Overview of publicly available and private datasets used for subjective tasks in Farsi. **Task** presents Emotion Analysis (EA), Sentiment Analysis (SA), and Toxicity Detection (TD). The columns provide details on the dataset name if provided (**Dataset**), which content modality that dataset uses (**Mult.**), annotation labels (**Labels**), source of the data (**Source**), the dataset size (**Size**), and if they are included in our experiments (**Included**). [E] Ekman framework. [T] Text and [S] Speech.

Task	Prompt
EA	Given a text, identify the main emotion expressed. You have to pick one of the following emotions: [List of dataset labels]. Text: {input} Only answer with the emotion and omit explanations. Emotion: {output}
SA	Given a text, identify the sentiment expressed. You have to pick one of the following sentiments: [List of dataset labels]. Text: {input} Only answer with the sentiment and omit explanations. Sentiment: {output}
TD	Does the following text contain [hate speech/ offensive language]? Only answer with yes or no. Text: {input}, Hate speech: {output}

Table 2: Prompt templates for Emotion Analysis (EA), Sentiment Analysis (SA), and Toxicity Detection (TD) tasks.

labels. In particular Persian-English code-mixed dataset provides 3,640 labeled tweets, making it one of the few resources addressing sentiment in code-mixed Persian-English text. **SentiPers** (Hosseini et al., 2018) contains 15,683 Digikala reviews annotated on a five-point scale ranging from -2 to $+2$. **MirasOpinion** is the largest available dataset collected from Digikala for SA in Farsi language with 93,868 samples. They label each sentence by using a Telegram¹⁷ bot to several users. They ask them to label the represented document as positive, negative, or neutral.

TD datasets: We find three datasets, each exclusively in text. **Phate** (Delbari et al., 2024) contains tweets that distinguishes between hateful content (with subcategories of violence, hate, and vulgar) and normal content, comprising 7,056 samples. The **PHICAD** (Davardoust et al., 2024) dataset is significantly larger, containing 300,000 samples, and labels content into hate/offense,

obscene, spam, or none, also sourced from comments on the Instagram platform. Lastly, **Pars-OFF** (Ataei et al., 2023) focuses on a binary classification of offensive versus non-offensive content with 8,334 samples of tweets.

These datasets, while valuable for advancing Farsi subjective analysis tasks, face several limitations. Many of them exhibit a narrow focus in terms of data sources, mostly based on tweets and comments on the Digikala platform, which may limit the generalizability of models trained on them to other contexts. Moreover, they also suffer from a lack of demographic information. Only two datasets of EA (Shemo and SAT) provide the demographic factors (e.g., gender in Shemo and age and gender for SAT). Only authors of three datasets (Yazdani and Shekofteh, 2022) provide detailed documentation on how annotations were conducted, whether multiple annotators were used, or what inter-annotator agreement was achieved. Without such information, it is difficult to assess the reliability of the labels used to train or evaluate

¹⁷<https://web.telegram.org/>

models.

Evaluating these datasets using LLMs may help address some of these shortcomings. [Abaskohi et al. \(2024\)](#) shows the low performance of GPT3.5 and GPT4¹⁸ on the emotion recognition task using only the ArmanEmo dataset. In the following section, we extend these evaluations by using various open-source models and datasets.

4 Evaluation Setting

4.1 Data

To evaluate the performance of language models on subjective tasks in Farsi, we select three datasets per task. For EA, we use ArmanEmo, LetHerLearn, and EmoPars. Since EmoPars contains multilabel emotions, we filter the dataset to include only samples in which one emotion has a non-zero value while all others are zero. With this approach, we reduce the size of the EmoPars dataset to 5,226 samples. We exclude the Shemo dataset because it relies on speech data, and the transcriptions alone do not adequately capture the nuances of emotion. We also excluded the SAT dataset due to its large number of labels, which could negatively affect language model performance. Finally, we exclude the LearnArmanEmo dataset, as it is derived from the LetherLearn and ArmanEmo datasets. For SA, we use ParsABSA, SentiPers, and a subsample of MirasOpinion. Since MirasOpinion is a very large dataset, we evaluate our language models on 30k randomly selected samples. We exclude the Persian-English code-mixed dataset due to its limited size and its primary focus on Persian code-mixed vocabulary. For the TD tasks, we use all the available datasets presented in Table 1. Given that the PHICAD dataset is extensive, with 300,000 samples, we experiment on a subsample provided by [Davardoust et al. \(2024\)](#)¹⁹ of the dataset with 131,959 instances.

4.2 Models

4.2.1 Open Source Decoder-only Models

From the family of decoder-only LLMs, we select three instruction-tuned versions of popular open-source models, which are Meta-Llama-3-8B-Instruct ([Dubey et al., 2024](#)), Mixtral-8x7B-Instruct-v0.1 ([Jiang et al., 2024](#)), and Qwen2-7B-

¹⁸<https://openai.com/>

¹⁹Part 1 available at <https://github.com/davardoust/PHICAD>

Instruct ([Yang et al., 2024](#)). For each task, we use a zero-shot approach to detect the relevant labels of emotions for EA, sentiments for SA, and hate speech/offensiveness for TD. We use two different prompting strategies on a subset of EA and SA datasets (see Appendix B), then we use the following prompt template that yielded the best performance across these datasets. For TD, we exclusively use the prompt introduced by [Delbari et al. \(2024\)](#). We summarize the list of prompts in Table 2. For the EA and SA template, we ask the model to identify the main emotion and sentiment expressed in the text, selecting from a predefined list of dataset-specific labels.

Task	Model	Lang.	Template		Avg. F1
			(I)	(II)	
EA	Llama3-8B	FA	0.19	0.19	0.19
		EN	0.18	0.20	0.19
	Mixtral-7B	FA	0.20	0.19	0.19
		EN	0.20	0.19	0.19
	Qwen2-7B	FA	0.19	0.20	0.19
		EN	0.20	0.17	0.18
SA	Llama3-8B	FA	0.46	0.64	0.55
		EN	0.46	0.48	0.47
	Mixtral-7B	FA	0.50	0.77	0.63
		EN	0.48	0.54	0.51
	Qwen2-7B	FA	0.48	0.36	0.42
		EN	0.48	0.46	0.47

Table 3: The performances of LLMs in macro average F1 scores for two prompting templates on the EA task for the EmoPars dataset and SA for the MirasOpinion dataset are reported. We use Farsi (FA) and English (EN) versions of datasets (Lang.). The EN version is translated by the NLBB model. Avg. F1 is the average F1 score across templates.

4.2.2 Data Translation Experiments

[Etxaniz et al. \(2024\)](#) suggest that translating non-English datasets to English can enhance the performance of multilingual LLMs. We adopt this strategy by translating our datasets to assess their impact on model results. Since multiple machine translation systems are available, we first translated a subsample of 100 Farsi sentences using Google Translate²⁰, the NLBB model ([Costa-Jussà et al., 2022](#)), and GPT-4o. After manual evaluation, we found that Google Translate produced the lowest-quality translations. Both NLBB and GPT-4o provided acceptable results, though they still exhibited issues such as literal translations, mistranslations, and omissions.

²⁰<https://translate.google.com/>

Ultimately, we chose to use NLBB because it is open source.

4.2.3 Encoder-only Models

For encoder-only architectures, we adopt standard fine-tuning approaches using XLM-RoBERTa (Conneau et al., 2020) and FaBERT (Masumi et al., 2025). XLM-RoBERTa is a multilingual transformer-based language model pre-trained on data from more than 100 languages. FaBERT is a Persian BERT base model pre-trained on the HmBlogs corpus, which includes both informal and formal Farsi texts. We fine-tune both models on nine datasets spanning the EA, SA, and TD tasks. Fine-tuning is performed by adding a classification head on top of the model’s final hidden representations and optimizing it using a cross-entropy loss.

5 Results

In this section, we present the outcomes of our experiments, detailing the evaluation of prompt selection, LLMs’ performances on the datasets in Farsi and their translation in English, and the fine-tuning approach.

5.1 Experiment 1: Prompt Variations and Data Translation

Prompt variations, even the smallest of perturbations such as adding a space at the end of a prompt, can affect the LLM’s output (Salinas and Morstatter, 2024). In this regard, we include two prompting strategies: the first involves directly asking the LLM to identify the subjective label of a given text, while the second includes the text’s data source as part of the prompt. For EA, we use a subsample from the EmoPars dataset, and for SA, we select the subsample of the MirasOpinion dataset. We choose these two publicly available datasets, because they are from the ACL Anthology and they have the largest sample sizes, with sample sizes of 5,226 for EmoPars and 30k for MirasOpinion. We evaluate two distinct prompt templates, as described in Appendix B, on these sub-samples.

Table 3 shows the performance of selected LLMs in EA and SA tasks over selected sub-samples in Farsi and English. The results of EA exhibit low F1-scores (between 0.18–0.20) across all models and configurations, with minimal differences between the original (FA) and translated (EN) data and only marginal variations

due to template changes. Using English translation does not consistently improve the results. In the EA, translation to English has a minimal overall impact, with two models showing no change (Llama3-8B and Mixtral-7B). For Qwen2-7B, we observe a slight decrease in the English version of the data. The same trend is for the SA task, where all models have a lower average F1 score over English texts, except for the Qwen2-7B model, whose translation increases the average F1-score from 0.42 to 0.47, which is negligible. Regarding different prompt templates, we do not observe significant improvements over a specific template in the EA task. However, in the SA task, template (II) performs better than both the Farsi and English versions of the data, except for the Qwen2-7B model. These findings suggest that both prompt design and data translation strategies for these subjective tasks in the Farsi language have a slight influence on model outcomes, particularly in EA.

5.2 Experiment 2: LLM and Fine-Tuned LM Evaluations

Based on results in Table 3, we use the prompt template (II) and datasets in Farsi (no translation) in a zero-shot setup to evaluate the different LLMs’ performances across the selected datasets in Section 4.1. Table 4 presents the macro average F1-score, across all tasks, datasets, and LMs. Performance is benchmarked against two baselines: a random classifier and a Most Frequent Class (MFC) baseline.

First, across all tasks, the fine-tuned transformer encoders (XLM-RoBERTa and FaBERT) consistently outperform the LLMs and baselines. This performance gap is most pronounced in Emotion Analysis, where XLM-RoBERTa and FaBERT achieve average F1-scores of 0.554 and 0.586, respectively, compared to 0.332–0.370 for the LLMs. Across all LLMs, Qwen2-7B consistently outperforms the other models, achieving the highest average F1-score in EA (0.370), SA (0.563), and TD (0.809). At the dataset level, Qwen2-7B achieves higher scores than ArmanEmo and LetHerLearn. Over the EmoPars dataset, Llama3-8B achieves a 0.227 average F1-score, which is slightly better than Qwen2-7B’s 0.218 average F1-score. In the SA task, the differences between model classes narrow but remain significant. Qwen2-7B achieves the strongest average among LLMs (0.563), outperforming both Llama3-8B (0.534)

Task	Dataset	LLMs			Fine-tuned LMs		Baselines	
		Llama3-8B	Mixtral-7B	Qwen2-7B	XLM-RoBERTa	FaBERT	Random	MFC
EA	ArmanEmo	0.426	0.296	0.510	0.630	0.641	0.135	0.061
	LetHerLearn	0.343	0.348	0.383	0.653	0.550	0.151	0.048
	EmoPars	0.227	0.188	0.218	0.380	0.566	0.152	0.048
	<i>Avg.</i>	0.332	0.277	0.370	0.554	0.586	0.146	0.052
SA	ParsABSA	0.501	0.498	0.444	0.856	0.441	0.242	0.168
	SentiPars	0.453	0.351	0.562	0.564	0.570	0.199	0.108
	MirasOpinion	0.647	0.608	0.683	0.854	0.854	0.330	0.230
	<i>Avg.</i>	0.534	0.486	0.563	0.758	0.622	0.257	0.169
TD	Phate	0.674	0.682	0.562	0.748	0.849	0.504	0.412
	Pars-OFF	0.696	0.741	0.925	0.854	0.889	0.491	0.412
	PHICAD	0.770	0.789	0.942	0.950	0.950	0.500	0.418
	<i>Avg.</i>	0.640	0.737	0.809	0.851	0.896	0.499	0.414

Table 4: Macro average F1-scores for each model and dataset across three tasks: SA = Sentiment Analysis, TD = Toxicity Detection, EA = Emotion Analysis. Averages are calculated per task. MFC is Most Frequent Class. The highest average F1-score per task is highlighted in bold.

and Mixtral-7B (0.486). This suggests that Qwen2-7B may encode sentiment-related features more effectively. However, the fine-tuned models again show superior performance. XLM-RoBERTa, in particular, achieves an average score of 0.758, an improvement over Qwen2-7B. Interestingly, FaBERT performs unevenly across datasets: it delivers competitive or even equal performance on MirasOpinion and SentiPars but falls behind on ParsABSA. This inconsistency may indicate dataset-specific biases or annotation standards that affect the model’s generalizability.

In the TD task, where all model families achieve their highest overall performance, LLMs approach the fine-tuned models. Qwen2-7B performs remarkably well, achieving an average score of 0.809, which is close to XLM-RoBERTa (0.851) and not far behind FaBERT (0.896). This strong performance aligns with the observation that toxicity classification relies heavily on identifying explicit lexical cues and linguistic markers that may be well represented in multilingual pretraining corpora. In this task, Qwen2-7B not only outperforms the other LLMs but also surpasses XLM-RoBERTa on Pars-OFF and PHICAD, suggesting that Qwen2-7B may be particularly well equipped for tasks involving offensive or toxic language categorization.

Taken together, these findings reveal that supervised fine-tuning on task-specific data remains essential to achieve better performance on these subjective tasks in Farsi, even with

increasingly powerful multilingual LLMs. In addition, although zero-shot LLMs do not yet match the performance of fine-tuned encoder-based models, their relative strength in the TD task suggests that some subjective tasks may be more amenable to zero-shot inference than others. We also report the results per label for each dataset, task, and model in Tables 5 to 7 at the Appendices B.5 to B.7.

6 Conclusion

Research on subjective tasks in Farsi has grown over the past five years, with a notable increase in SA and TD research starting in early 2020. Most work has focused on two main data sources: social media data, such as tweets, and e-commerce data from Digikala, highlighting the scarcity of Farsi-language data sources. We reviewed over 110 papers, including 12 from the ACL Anthology and 98 from other publishers. We identified several gaps in these studies, including a lack of diverse datasets, annotation information, and demographic features in subjective tasks, particularly for EA. These gaps include demographic disparities such as age and gender and a lack of interdisciplinary research. Our experiments indicate that LLMs perform relatively poorly on EA tasks in Farsi but perform better on SA and TD. Additionally, fine-tuning consistently improves performance across all tasks.

7 Limitations and Ethical Considerations

We acknowledge several limitations in our study. First, our evaluation relies heavily on existing publicly available datasets, which may not comprehensively capture the linguistic, cultural, or topical diversity of the Farsi language. These datasets may contain annotation biases, domain-specific skew, or inconsistencies that could affect model performance and generalizability. Moreover, we use machine translation for English versions of Farsi texts, which may introduce semantic drift or cultural misrepresentation and affect fairness and accuracy. Another limitation is the limited number of models that support Farsi.

Acknowledgment

Donya Rooein, Debora Nozza, and Dirk Hovy are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. Donya Rooein and Dirk Hovy’s research is supported through the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). Debora Nozza’s research is from the ERC under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE).

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. [Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: a survey. *IEEE access*, 12:18416–18450.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. [Emotions from text: Machine learning for text-based emotion prediction](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Razieh Asgarnezhad and S Amirhassan Monadjemi. 2021. [Persian sentiment analysis: feature engineering, datasets, and challenges](#). *Journal of applied intelligent systems & information sciences*, 2(2):1–21.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2023. [Pars-off: A benchmark for offensive language detection on farsi social media](#). *IEEE Transactions on Affective Computing*, 14(4):2787–2795.

Christina Barz, Melanie Siegel, Daniel Hanss, and Michael Wiegand. 2025. Understanding disagreement: An annotation study of sentiment and emotional language in environmental communication. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 1–20.

Magdalena Borowczyk. 2023. *1 Research in Persian Natural Language Processing – History and State of the Art*, pages 1–24. De Gruyter Mouton, Berlin, Boston.

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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.

Hadi Davardoust, Hadi Zare, and Hossein RafieeZade. 2024. [The dark side of instagram: A large dataset for identifying persian harmful comments](#). *SoCal NLP Symposium 2024*.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. [Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17889–17897.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Ekman, Tim Dalgleish, and M Power. 1999. Basic emotions. *San Francisco, USA*.

Sina Elahimanesh, Shayan Salehi, Sara Zahedi Movahed, Lisa Alazraki, Ruoyu Hu, and Abbas

- Edalat. 2023. From words and exercises to wellness: Farsi chatbot for self-attachment technique. *arXiv preprint arXiv:2310.09362*.
- Julen Etzaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. **Do multilingual language models think better in English?** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. 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*.
- Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: a sentiment analysis corpus for persian. *arXiv preprint arXiv:1801.07737*.
- Lai Po Hung and Suraya Alias. 2023. Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 27(1):84–95.
- Mohammad Ali Hussiny and Lilja Øvrelid. 2023. **Emotion analysis of tweets banning education in Afghanistan.** In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 271–277, Toronto, Canada. Association for Computational Linguistics.
- Mohammad Ali Hussiny, Mohammad Arif Payenda, and Lilja Øvrelid. 2024. **PersianEmo: Enhancing Farsi-Dari emotion analysis with a hybrid transformer and recurrent neural network model.** In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 257–263, Torino, Italia. ELRA and ICCL.
- Sara Izadi, Javad Sadri, Farshid Solimanpour, and Ching Y Suen. 2006. A review on persian script and recognition techniques. *Summit on Arabic and Chinese Handwriting Recognition*, pages 22–35.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. **The state and fate of linguistic diversity and inclusion in the NLP world.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Mahboobeh Sadat Kobari, Nima Karimi, Benyamin Pourhosseini, and Ramin Mousa. 2023. **weighted capsulenet networks for persian multi-domain sentiment analysis.** *arXiv preprint arXiv:2306.17068*.
- Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection—techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.
- Mostafa Masumi, Seyed Soroush Majd, Mehrnoush Shamsfard, and Hamid Beigy. 2025. **FaBERT: Pre-training BERT on Persian blogs.** In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 85–96, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Hossein Mirzaee, Javad Peymanfar, Hamid Habibzadeh Moshtaghin, and Hossein Zeinali. 2022. Armanemo: A persian dataset for text-based emotion detection. *arXiv preprint arXiv:2207.11808*.
- Ashritha R Murthy and KM Anil Kumar. 2021. A review of different approaches for detecting emotion from text. In *IOP Conference Series: Materials Science and Engineering*, volume 1110, page 012009. IOP Publishing.
- Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.
- Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. **The shifted and the overlooked: A task-oriented investigation of user-GPT interactions.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.
- John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. **Toxicity detection: Does context really matter?** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.
- Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. **Emotion analysis in NLP: Trends, gaps and roadmap for future directions.** In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.
- Robert Plutchik. 1982. **A psychoevolutionary theory of emotions.** *Social Science Information*, 21(4-5):529–553.

- Zeinab Rajabi and MohammadReza Valavi. 2021. [A survey on sentiment analysis in persian: a comprehensive system perspective covering challenges and advances in resources and methods](#). *Cognitive Computation*, 13(4):882–902.
- Nazanin Sabri, Reyhane Akhavan, and Behnam Bahrak. 2021a. [EmoPars: A collection of 30K emotion-annotated Persian social media texts](#). In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 167–173, Online. INCOMA Ltd.
- Nazanin Sabri, Ali Edalat, and Behnam Bahrak. 2021b. [Sentiment analysis of persian-english code-mixed texts](#). In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4. IEEE.
- Seyedeh S Sadeghi, Hasan Khotanlou, and M Rasekh Mahand. 2021. [Automatic persian text emotion detection using cognitive linguistic and deep learning](#). *Journal of AI and Data Mining*, 9(2):169–179.
- 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.
- Taha Shangipour ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2022. [Pars-ABSA: a manually annotated aspect-based sentiment analysis benchmark on Farsi product reviews](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7056–7060, Marseille, France. European Language Resources Association.
- Pragya Singh Tomar, Kirti Mathur, and Ugrasen Suman. 2023. [Unimodal approaches for emotion recognition: A systematic review](#). *Cognitive Systems Research*, 77:94–109.
- Changhao Song, Yazhou Zhang, Hui Gao, Ben Yao, and Peng Zhang. 2025. [Large language models for subjective language understanding: A survey](#). *arXiv preprint arXiv:2508.07959*.
- Jie Tao and Xing Fang. 2020. [Toward multi-label sentiment analysis: a transfer learning based approach](#). *Journal of Big Data*, 7(1):1.
- Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. [The sentiment problem: A critical survey towards deconstructing sentiment analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore. Association for Computational Linguistics.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. [Recognizing contextual polarity in phrase-level sentiment analysis](#). In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *arXiv preprint arXiv:1910.03771*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2. 5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Ali Yazdani and Yasser Shekofteh. 2022. [A persian asr-based ser: modification of sharif emotional speech database and investigation of persian text corpora](#). *arXiv preprint arXiv:2211.09956*.
- Ali Yazdani, Hossein Simchi, and Yasser Shekofteh. 2021. [Emotion recognition in persian speech using deep neural networks](#). In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 374–378.

A Survey Analysis

B Prompt Templates

B.1 Prompt Templates for EA

- **Template (I):** Given a text, identify the main emotion expressed. You have to pick one of the following seven emotions: sadness, hate, anger, happiness, fear, surprise, or other. Only answer with emotion and omit explanations. Emotion:
- **Template (II):** You will be presented with a given comment sourced from X, Instagram, or Digikala. Pick one emotion from sadness, hate, anger, happiness, fear, surprise, or other that describes the emotion of the tweet or comment the best. Your response should only contain one of the emotions. No other output is allowed.

B.2 Prompt Templates for SA

- **Template (I):** Given a text, identify the sentiment expressed. You have to pick one of the following three sentiments: positive, negative, neutral. Only answer with the sentiment and omit explanations. Sentiment:
- **Template (II)** You will be presented with a comment from Digikala. Pick one sentiment

from positive, negative, or neutral that describes the sentiment of the comment the best. Your response should only contain one of sentiment. No other output is allowed.

B.3 Model hyperparameters

B.4 Models

Llama3 (Grattafiori et al., 2024) is an open-access collection of pre-trained and fine-tuned LLMs ranging in scale from 8 billion to 70 billion parameters and launched in September 2024. We examine Llama3-8B model. We use Qwen2-7B-Instruct model that published in November 2024 (Yang et al., 2024). Mistral-7b is also an open-source LM launched in September 2023 (Jiang et al., 2024). Among the models released by Mistral, we test Mixtral-8x7B-Instruct-v0.1, and we access these models via HuggingFace (Wolf et al., 2019).

All responses were collected during July 2024 to March 2025. We run all our experiments on a server with three NVIDIA RTX A6000 and 48GB of RAM.

XML-RoBERTa The hyperparameters for the XML-RoBERTa is three epochs, batch size of 16, learning_rate of 2e-5, optimizer of Adam and the maximum length of 128.

B.5 Emotion Analysis

Table 5 shows the performance of the LLMs across different emotions for each dataset.

B.6 Sentiment Analysis

Table 6 shows the performance of the LLMs across different sentiments for each dataset. Mixtral-7B and Llama3-8B can not capture “very negative” and “very positive” labels.

B.7 Toxicity Detection

Table 7 shows the performance of the LLMs across each dataset for detecting offensive/hate speech languagee.

Dataset	Emotion	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
Letherlearn	Anger	0.241	0.493	0.358	0.364
	Disgust	0.189	0.056	0.183	0.143
	Fear	0.488	0.461	0.458	0.469
	Happiness	0.423	0.545	0.560	0.509
	Sadness	0.447	0.503	0.511	0.487
	Surprise	0.420	0.264	0.268	0.317
	Other	0.228	0.080	0.345	0.218
Arman	Anger	0.021	0.455	0.456	0.311
	Fear	0.761	0.760	0.733	0.751
	Hate	0.109	0.241	0.441	0.264
	Happiness	0.075	0.521	0.584	0.393
	Sadness	0.414	0.489	0.480	0.461
	Surprise	0.465	0.440	0.483	0.463
	Other	0.231	0.075	0.393	0.233
EmoPars	Anger	0.262	0.307	0.220	0.263
	Fear	0.141	0.162	0.177	0.160
	Hate	0.014	0.046	0.154	0.071
	Happiness	0.247	0.294	0.300	0.280
	Sadness	0.288	0.240	0.256	0.261
	Surprise	0.173	0.066	0.202	0.147

Table 5: F1 Scores for Emotion Analysis Across Datasets and Models with Average.

Dataset	Sentiment	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
MirasOpinion	Negative	0.619	0.631	0.656	0.635
	Neutral	0.138	0.498	0.592	0.409
	Positive	0.736	0.812	0.800	0.783
Pars-ABSA	Negative	0.619	0.627	0.616	0.621
	Neutral	0.138	0.332	0.336	0.269
	Positive	0.736	0.741	0.734	0.737
Sentipers	Very Negative	0.000	0.000	0.058	0.019
	Negative	0.560	0.563	0.570	0.564
	Neutral	0.675	0.593	0.664	0.644
	Positive	0.520	0.576	0.586	0.561
	Very Positive	0.000	0.620	0.341	0.320

Table 6: F1 Scores for Sentiment Analysis Across Datasets and Models with Average.

Dataset	Labels	Mixtral-7B	Llama3-8B	Qwen2-7B	Avg.
Pars_OFF	not-offensive	0.841	0.736	0.993	0.857
	offensive	0.640	0.656	0.857	0.718
Phate	not-hate	0.692	0.553	0.720	0.655
	hate	0.673	0.778	0.409	0.620
PHICAD	not-hate	0.911	0.887	0.990	0.929
	hate	0.667	0.653	0.894	0.738

Table 7: Toxicity Detection F1 Scores Across Datasets and Models.

Emotional Lexicons: How Large Language Models Predict Emotional Ratings of Russian Words

Polina V. Iaroshenko and Natalia V. Loukachevitch

Research Computing Center

Lomonosov Moscow State University

Moscow, Russian Federation

polina.iaroshenko@yandex.ru and louk_nat@mail.ru

Abstract

This study examines the capability of LLMs to predict emotional ratings of Russian words by comparing their assessments with both native speakers' ratings and expert evaluations. The research utilises two datasets: the EN-RuN database containing associative emotional ratings of Russian nouns by native speakers, and RusEmoLex, an expert-compiled lexicon. Various open-source LLMs were evaluated, including international models (Llama-3, Qwen 2.5), Russian-developed models, and Russian-adapted variants, representing three parameter scales. The findings reveal distinct patterns in model performance: Russian-adapted models demonstrated superior alignment with native speakers' ratings, whilst model size was not a decisive factor. Conversely, larger models showed better performance in matching expert assessments, with language adaptation having minimal impact. Emotional or sensitive lexis with strong connotations produce a more substantial human-model gap.

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, including emotion analysis. Previous research on emotions expressed in text has primarily focused on emotion detection and categorization (Acheampong et al., 2020; Kazuyulina et al., 2021; Bostan and Klinger, 2024; Cavicchio, 2025). However, the widespread adoption of chatbots and advancement of LLMs have necessitated not only the recognition and classification of emotions but also the authentic simulation of emotional responses during human-computer interactions. Consequently, the study of LLMs' emotional intelligence has become increasingly relevant (see, for example, Raj, 2024; Chen et al., 2024; Huang et al., 2024; Dalal et al., 2025).

It is important to note that in emotion analysis, as in many other areas of NLP, research con-

ducted on English language material continues to dominate the field (for a detailed discussion, see De Bruyne, 2023). Cross-cultural variations in emotional expression raise concerns about LLMs' ability to adapt to different linguistic contexts and cultural value systems.

One approach to studying LLMs' emotional behaviour and comparing it with human responses is to examine emotional ratings of individual lexical items. Emotional word ratings can be categorised into two distinct methodologies: associative and expert-based evaluations. Associative emotional assessment refers to survey results, in which respondents evaluate the degree to which lexical units are associated with specific emotions. These experiments can employ any words as stimuli, regardless of their direct connection to emotional domains (see, for example, Mohammad and Turney, 2013; Domingues Aparecido et al., 2025). Expert-based emotional assessment, conversely, relies on specialized linguistic resources such as dictionaries or thesauri dedicated to specific semantic categories – in this case, emotional lexis. These resources, developed specifically to compile emotional lexicons, exclusively contain words that experts have identified as expressing or indicating emotions (for a detailed discussion, see Mohammad, 2023).

This study aims to compare human emotional ratings of Russian words with those generated by LLMs. The research examines both native speakers' associative emotional assessments and expert evaluations. For associative ratings, we utilize the ENRuN (Emotional Norms for Russian Nouns) database (Sysoeva and Lyusin, 2024), comprising ratings for 1,800 Russian nouns. Expert evaluations are derived from the Russian Emotion Lexicon (RusEmoLex) (Iaroshenko and Loukachevitch, 2025b), which provides emotional categorization for 1,024 words. RusEmoLex's lexical inclusion criteria primarily rely on specialized linguistic resources (dictionaries, thesauri, and corpus data).

The study employs open-source models of varying scales (three categories: 7-8B, 32B, and 70B parameters) and diverse origins: international models (Llama-3, Qwen 2.5), Russian-developed models, and Russian-adapted variants.

We aim to answer the following research questions (RQs): **RQ1**: How accurately do open-source LLMs predict emotional word ratings in Russian when compared to native speakers' and expert assessments? **RQ2**: To what extent does model size affect the quality of emotional assessment? **RQ3**: Do Russian-adapted models demonstrate superior performance in emotional rating predictions compared to their non-adapted counterparts?

2 Related Work

During the last few years, research on LLM empathy and emotional intelligence has expanded, employing advanced benchmarks to evaluate models' emotional responses across varied scenarios.

EmotionBench (Huang et al., 2024) introduces a theoretically grounded methodology derived from psychological emotion appraisal theory (Roseman and Smith, 2001). The researchers identified distinct situational factors that elicit specific emotional responses. The comparative analysis between LLM responses and human assessments revealed that while LLMs can provide appropriate responses, they demonstrate significant limitations in aligning with human emotional behaviour.

The authors of (Sabour et al., 2024) introduced EmoBench, a comprehensive evaluation framework in English and Chinese. The benchmark is structured around two primary assessment areas: Emotional Understanding and Application. In the Understanding component, LLMs must identify emotions and their underlying causes, with particular emphasis on complex, ambiguous situations. The Application component evaluates models' ability to select appropriate responses to given scenarios. The findings revealed that human participants outperformed LLMs across both tasks, albeit by a modest margin.

A notable contribution to understanding LLMs' emotional capabilities comes from research employing standardized psychological assessment tools. The study by (Vzorin et al., 2024) assessed LLM emotional intelligence through the Russian version of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). (Dalal et al., 2025) employed established psychological instruments:

the Trait Emotional Intelligence Questionnaire (TEIQue) (Petrides, 2009) and the Situational Evaluation of Complex Emotional Understanding (SECEU) (MacCann and Roberts, 2008).

While recent research has made substantial progress in evaluating LLMs' emotional intelligence capabilities, most studies have focused on the English language. The need to investigate emotional intelligence across different linguistic and cultural contexts remains a crucial research direction.

3 Methodology

The primary objective of this study is to evaluate LLMs' ability to assess Russian words for their emotional associations and to compare these assessments with human ratings. With this aim, we use the ENRuN database, based on native speaker scores, and the Russian Emotion Lexicon, constructed by professional linguists.

Data. The ENRuN database (Sysoeva and Lyusin, 2024)¹ provides emotional ratings for 1,800 Russian nouns, evaluated through both dimensional (valence and arousal) and categorical (happiness, sadness, anger, fear, and disgust) approaches. For the present analysis, we utilize averaged categorical ratings, where respondents assessed word-emotion associations on a five-point scale.

RusEmoLex (Iaroshenko and Loukachevitch, 2025b) comprises 1,024 Russian lexical items representing various parts of speech. Each entry is annotated with one of five emotional categories: happiness, sadness, anger, fear, and surprise. RusEmoLex was developed by integrating various Russian emotional vocabulary resources, with expert linguistic sources serving as the primary foundation: dictionaries and thesauri (Shvedova, 1998; Babenko, 2022), and data from the semantically annotated section of the Russian National Corpus (Savchuk et al., 2024).

To enable comparison between RusEmoLex and ENRuN datasets, RusEmoLex entries were converted to numerical ratings: words labeled with a specific emotion (e.g., happiness) were assigned a maximum score (5) for that emotion and zero scores for all other emotional categories.

¹The current version of the ENRuN database can be provided to researchers upon request. At the time of this study, the database included responses from 692 participants, with data collection continuing as an ongoing process.

Models. We employ open-source LLMs ranging from 7B to 70B parameters, representing three categories by origin: international models (Qwen 2.5-32B and 7B, Llama-3), Russian-developed models (YandexGPT 5 Lite), and models adapted for Russian through enhanced training and tokenization (T-lite-it-1.0, T-pro-it-1.0, RuadaptQwen2.5-7B, RuadaptQwen2.5-32B-Pro-Beta). A complete list of models, along with their references are provided in Appendix A.

Hyperparameters. The experiments employ the following hyperparameters: temperature was set to 0.2 to ensure relatively consistent outputs while maintaining some degree of creativity; a repetition penalty of 1.2 was implemented to discourage uniform ratings across different words. For a detailed discussion of hyperparameter effects on LLM responses in similar rating tasks, see (Iaroshenko and Loukachevitch, 2025a).

Prompts. For the experiments, two types of prompts were employed. The SPEAKER prompt, based on the instructions given to participants in the ENRuN database evaluation experiment (Lyusin and Sysoeva, 2025), was used to assess emotional words from a native speaker’s perspective. The EXPERT prompt was designed to evaluate words from an expert’s viewpoint. Several versions of this prompt were tested, including one with an additional “Background Information” section that provided a theoretical definition of emotional vocabulary and examples of words belonging to this semantic class. While this approach was hypothesised to enhance expert-based evaluation, maintaining the SPEAKER prompt structure proved to be the most effective. The prompts used in the experiment follow a similar structure, including role designation (either a native speaker participant or a linguistics researcher specialising in semantics) and task description. Both prompts maintain identical task descriptions but vary in their role specifications. The original ENRuN instruction, SPEAKER and EXPERT prompts used in the experiments are presented in Appendices B and C.

4 Experiment

To compare LLM evaluations with human ratings across both datasets, LLM assessments were averaged over three iterations. This reduced the impact of LLM response variability, enhancing result reliability and objectivity. For the ENRuN database comparisons, the standard deviation was employed

Size	Model Name	Std Dev
70B	Llama-3	1.16
32B	RuadaptQwen-32B	0.93
32B	T-pro	1.08
32B	Qwen-32B	1.15
8B	YaGPT5-Lite	1.05
7B	RuadaptQwen-7B	1.18
7B	T-lite	1.22
7B	Qwen-7B	1.23

Table 1: ENRuN evaluation results

as the evaluation metric. RusEmoLex analysis utilised both standard deviation and F-measure, as its original data format involves categorical emotional classification of words.

ENRuN. Table 1 presents a comparison between averaged LLM ratings and Russian native speaker assessments from the ENRuN database. To interpret these results, we first examine inter-human variability: when different people rate the same emotional associations in ENRuN, their assessments vary with an average standard deviation of 1.16 across all words and emotions. The values reported in Table 1 measure how much each model’s ratings deviate from the averaged human assessments. Russian-adapted models demonstrated the closest alignment with human evaluations, deviating from human consensus less than humans deviate from each other: RuadaptQwen2.5-32B-Pro-Beta achieved the lowest standard deviation (0.93), followed by YandexGPT 5 Lite (1.05), which, despite its small size, outperformed the larger Llama-3 model (1.16). Notably, smaller adapted models (RuadaptQwen2.5-7B: 1.18, T-lite: 1.22) outperformed their original versions (Qwen2.5-7B: 1.23). These results suggest that modern LLMs align with native speaker assessments of emotional associations at a level comparable to human performance.

RusEmoLex. Table 2 presents the comparison results between averaged LLM ratings and expert assessments based on RusEmoLex. Two prompts were employed for this dataset: EXPERT, which assigned the LLM the role of a professional linguist, and SPEAKER, which was used for ENRuN. Notably, the role differentiation in prompts did not significantly affect the final results.

For the RusEmoLex, medium-scale models (Qwen 2.5-32B and T-pro) achieved the best performance by standard deviation metrics, while small-scale models (7-8B) consistently showed higher

Size	Model	EXPERT	SPEAKER	EXPERT	SPEAKER
		Std Dev	Std Dev	F1	F1
70B	Llama-3	1.51	1.50	0.79	0.77
32B	Qwen 2.5-32B	1.45	1.48	0.77	0.76
32B	T-pro	1.46	1.45	0.77	0.78
32B	RuadaptQwen2.5-32B	1.51	1.55	0.77	0.75
8B	YandexGPT 5 Lite	1.48	1.52	0.75	0.70
7B	T-lite	1.61	1.68	0.64	0.59
7B	RuadaptQwen2.5-7B	1.63	1.60	0.63	0.62
7B	Qwen 2.5-7B	1.66	1.48	0.53	0.68

Table 2: RusEmoLex evaluation results

deviation values. This pattern is further supported by F-measure analysis, where medium-scale (32B) and larger models (Llama-3 70B) demonstrated similar performance (0.75-0.79), outperforming smaller models. Notably, Russian-adapted models showed no distinct advantage in this evaluation.

5 Discussion

The results demonstrate that modern LLMs can achieve reasonable alignment with human assessments, though with varying degrees of success.

On the ENRuN dataset, RuadaptQwen2.5-32B-Pro-Beta, adapted for Russian, demonstrated the highest performance. Model size, however, was not a decisive factor: YandexGPT 5 Lite (8B) ranked second, outperforming larger models including Llama-3. Generally, the Russian-adapted versions showed closer alignment with human ratings compared to their original counterparts.

For the RusEmoLex dataset, large (Llama-3) and medium-sized models (Qwen 2.5-32B, T-pro) proved to be more effective, whilst Russian adaptation showed no significant impact on performance.

Thus, adapted model versions demonstrated closer alignment with native speakers’ emotional ratings rather than expert assessments. This suggests that tokenizer modifications and additional training on Russian-language data led to better adaptation to native speakers’ perceptions. Conversely, model size emerged as the crucial factor in alignment with expert evaluations. The compilation of an emotional vocabulary requires professional linguistic expertise, which may explain why model size proved more significant than Russian language adaptation.

Furthermore, this disparity in results may be attributed to differences in the evaluated lexical content. The ENRuN database comprises words se-

lected based on formal, semantically-neutral criteria, and consequently contains considerable lexical diversity. Conversely, RusEmoLex was developed by intersecting emotional word lists from various resources, primarily emphasising dictionary and corpus data; its lexical composition therefore predominantly comprises emotion words.

For a more in-depth analysis of how the ENRuN database’s lexical composition influences the divergence between LLM and human responses, diagnostic lexical groups were formed from the database, and the standard deviations between LLM and human ratings were examined for each group. The first group, comprising neutral words, included all words with ratings of 1.5 or below across all emotions (555 words; e.g. *parking, shoe, website*). The second group consisted of words from RusEmoLex (57 words; e.g. *joy, worry, jealousy*). The third group, comprising emotion-related words, included words with ratings of 3.5 or above for at least one emotion, representing lexical items eliciting the strongest emotional responses. This group encompassed both direct emotion terms (*despair, despondency, melancholy*) and sensitive lexis with marked connotations (positive: *baby, safety, friendliness*; negative: *corruption, orphan, alcoholism*).

See Figure 1 for the average standard deviations between human and LLM ratings across these three lexical groups. Lower standard deviation values indicate greater agreement between human and model ratings. Comparison of human and model ratings revealed a consistent trend: neutral words exhibited the lowest standard deviation across all models (mean STD = 0.85), whilst the RusEmoLex group showed the highest deviation (mean STD = 1.30), followed by emotion-related words (mean STD = 1.24). Thus, emotional or sensitive lexis

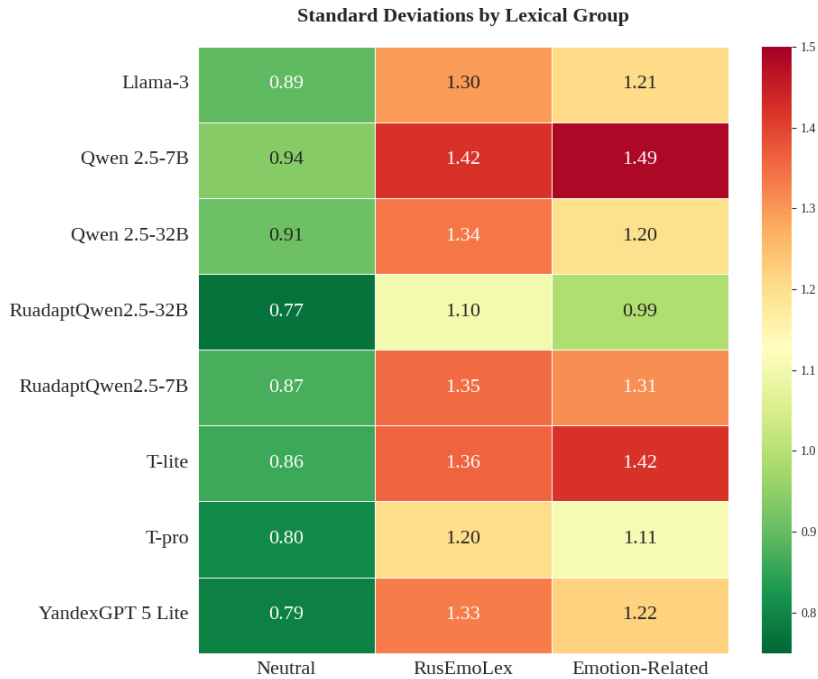


Figure 1: Average standard deviations between human and LLM ratings by lexical group in the ENRuN database

with strong connotations produce a more substantial human-model gap, whereas agreement is higher for neutral lexis.

6 Conclusion

In the study, we compared emotional ratings of Russian words between open-source LLMs and human assessments. Two datasets were employed for comparison: the ENRuN database, containing native speakers’ Russian noun ratings, and RusEmoLex, a lexicon developed from expert linguistic sources.

Addressing RQ-1, the models demonstrated a relatively high correlation with human word assessments (best results: standard deviation of 0.93 for the ENRuN dataset, standard deviation of 1.45 and F-measure of 0.79 for RusEmoLex).

Regarding RQ-2 and RQ-3, we observed distinct patterns in model behaviour. Native speakers’ ratings were better predicted by Russian-adapted LLMs, with model size not being a determining factor. Expert assessments were more accurately predicted by larger-scale models, whilst adaptation showed no significant impact on performance.

Limitations

Dataset Scope. Two datasets were utilised as source material. Firstly, the datasets are limited in size (ENRuN comprises 1,800 evaluated words

and RusEmoLex — 1,024). Secondly, both datasets include annotations of individual lexical units; therefore, the research findings may not be relevant for evaluating LLMs’ performance in context-dependent emotional assessment. Thirdly, the comparison of results between the two datasets may not be fully valid due to differences in their emotional rating frameworks.

Model Scope. The study was limited to specific versions of open-source models, not all possible combinations of model sizes were tested. Among larger-scale models, only Llama-3 was represented. The evaluation of larger models required significant computational resources, which might limit the practical applicability of the findings.

The performance of LLMs might change with future updates and new model releases. The study represents a snapshot of current model capabilities rather than a longitudinal assessment. These limitations suggest directions for future research in emotional assessment using LLMs for processing the Russian language.

Acknowledgments

The research was supported by the Russian Science Foundation, project No. 25-11-00191, <https://rscf.ru/project/25-11-00191/>. The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

Large Language Models were utilised for manuscript improvement, proofreading, and verification of academic English grammatical and stylistic standards.

References

- Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. [Text-based emotion detection: Advances, challenges, and opportunities](#). *Engineering Reports*, 2.
- Ludmila G. Babenko. 2022. *Alphabet of Emotions: The-saurus of Emotive Lexis*. Armchair Scholar, Yekaterinburg; Moscow.
- Laura Ana Maria Bostan and Roman Klinger. 2024. An analysis of annotated corpora for emotion classification in text. Technical report, Otto-Friedrich-Universität, Bamberg.
- Federica Cavicchio. 2025. *Emotion Detection in Natural Language Processing*. Springer, Cham.
- Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. [EmotionQueen: A benchmark for evaluating empathy of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.
- Dhruv Dalal, Garima Negi, and Davide Picca. 2025. [LLMs and emotional intelligence: Evaluating emotional understanding through psychometric tools](#). In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '25*, pages 323–328.
- Luna De Bruyne. 2023. [The paradox of multilingual emotion detection](#). In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.
- Thales David Domingues Aparecido, Alexis Carrillo, Chico Q. Camargo, and Massimo Stella. 2025. [Benchmarking psychological lexicons and large language models for emotion detection in brazilian portuguese](#). *AI*, 6(10).
- Jiaqi Huang, Man Ho Lam, Eric Jing Li, Shuhuai Ren, Wei Wang, Weizhen Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Apathetic or empathetic? Evaluating LLMs’ emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087.
- Polina V. Iaroshenko and Natalia V. Loukachevitch. 2025a. [Large language models versus native speakers in emotional assessment of Russian words](#). *Supercomputing Frontiers and Innovations*, 12(3):20–30.
- Polina V. Iaroshenko and Natalia V. Loukachevitch. 2025b. [RusEmoLex: Russian emotion lexicon](#). *Russian Journal of Linguistics*, 29(3).
- Marina Kazyulina, Andrey Babii, and Alexey Malafeev. 2021. Emotion classification in Russian: Feature engineering and analysis. In *Analysis of Images, Social Networks and Texts, AIST 2020*, volume 12602 of *Lecture Notes in Computer Science*, pages 135–148.
- Dmitry Lyusin and Tatiana A. Sysoeva. 2025. [ENRuN database: Emotional ratings of Russian nouns](#). *Experimental Psychology*, 18(2):206–219. (In Russian).
- Carolyn MacCann and Richard D. Roberts. 2008. [New paradigms for assessing emotional intelligence: theory and data](#). *Emotion*, 8(4):540–551.
- Saif Mohammad. 2023. [Best practices in the creation and use of emotion lexicons](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.
- Saif Mohammad and Peter D. Turney. 2013. [Crowdsourcing a word-emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Konstantinos V. Petrides. 2009. [Psychometric properties of the Trait Emotional Intelligence Questionnaire \(TEIQue\)](#). In Con Stough, Donald H. Saklofske, and James D. A. Parker, editors, *Assessing emotional intelligence: Theory, research, and applications*, pages 85–101. Springer Science + Business Media.
- Pankaj Raj. 2024. [A literature review on emotional intelligence of large language models\(LLMs\)](#). *International Journal of Advanced Research in Computer Science*, 15(4).
- Ira J. Roseman and Craig A. Smith. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. In Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 3–19. Oxford University Press.
- Sara Sabour, Siyang Liu, Zhengyuan Zhang, Jing Liu, Jie Zhou, Alex Sunaryo, Taesung Lee, Rada Mihalcea, and Minlie Huang. 2024. [EmoBench: Evaluating the emotional intelligence of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 5986–6004, Bangkok. Association for Computational Linguistics.
- Svetlana O. Savchuk, Timofey Arkhangelskiy, Anas-tasiya A. Bonch-Osmolovskaya, Ol’ga V. Donina, Yuliya N. Kuznetsova, Ol’ga N. Lyashevskaya, Boris V. Orekhov, and Mariya V. Podryadchikova. 2024. Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoz-nanija*, 2:7–34.

Natalia Yu. Shvedova. 1998. *The Russian Semantic Dictionary: Explanatory Dictionary, Systematized by Classes of Words and Meanings*. Azbukovnik, Moscow.

Tatiana A. Sysoeva and Dmitry V. Lyusin. 2024. Development of an extended database with emotional ratings of nouns ENRuN-2: successes, problems and prospects. In *Psychology of cognition: proceedings of the All-Russian Scientific Conference*, pages 316–320, Yaroslavl. YARSU. (In Russian).

Gleb D. Vzorin, Alexey M. Bukinich, Anna V. Sedykh, Irina I. Vetrova, and Elena A. Sergienko. 2024. [The emotional intelligence of the GPT-4 large language model](#). *Psychology in Russia: State of the Art*, 17(2):85–99.

A Models

This appendix provides a complete list of models, arranged by model size with their corresponding links.

- Llama-3 (70B): <https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>
- RuadaptQwen2.5-32B-Pro-Beta (32B): <https://huggingface.co/RefalMachine/RuadaptQwen2.5-32B-Pro-Beta>
- T-pro-it-1.0 (32B): <https://huggingface.co/t-tech/T-pro-it-1.0>
- Qwen 2.5-32B (32B): <https://huggingface.co/Qwen/Qwen2.5-32B-Instruct>
- YandexGPT 5 Lite (8B): <https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct>
- RuadaptQwen2.5-7B (7B): <https://huggingface.co/RefalMachine/RuadaptQwen2.5-7B-Lite-Beta>
- T-lite-it-1.0 (7B): <https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1>
- Qwen 2.5-7B (7B): <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

B Original ENRuN instruction

This appendix describes the initial instructions given to human respondents for creating the ENRuN database.

The ENRuN instruction text in Russian

Оцените, пожалуйста, по шкале 0-5, насколько, по Вашему мнению, каждое из перечисленных ниже слов ассоциируется (связано) с такими эмоциями, как радость, страх, отвращение, злость и грусть. Вам необходимо заполнить приведенные ниже таблицы, в строках которых указаны слова, а в столбцах — эмоции. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то на пересечении соответствующих строки и столбца поставьте "0", если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте "5", Вы можете использовать также и все средние значения указанной шкалы. Таким образом, для каждого слова необходимо дать 5 оценок: насколько оно ассоциируется с радостью (1 столбец), насколько — со страхом (2 столбец), с отвращением (3 столбец), со злостью (4 столбец) и с грустью (5 столбец). Если необходимо, Вы можете ставить высокие оценки сразу в нескольких столбцах для одного и того же слова.

The ENRuN instruction text translated in English

Please rate using the scale from 0 to 5 to which extent, in your opinion, each word is related to emotions of happiness, fear, disgust, anger, and sadness. You will have to fill out the tables below. Words are in the rows and emotions are in the columns. If you think that the given word is not related at all to the given emotion, write "0". If you think that the given word is very much related to the given emotion, write "5". You can also use all the intermediate values of this scale. You have to give five ratings for each scale indicating as to how strongly the given word is related to happiness (1st row), fear (2nd row), disgust (3rd row), anger (4th row), and sadness (5th row). If necessary, you can give high ratings in several columns for the same word.

Figure 2: The ENRuN instruction

C Prompts Used in the Study

This appendix describes the prompts used in the current study: the SPEAKER prompt based on the ENRuN instruction for human respondents and the two variants of the EXPERT simulating professional linguist analysis, and also their translations into English.

The SPEAKER prompt in Russian	The SPEAKER prompt translated in English
<p>Войди в РОЛЬ и выполни ЗАДАНИЕ.</p> <p>РОЛЬ:</p> <p>Ты — обычный человек, который говорит на русском языке и живет в России. Тебя пригласили участвовать в эксперименте учёные из Лаборатории когнитивных исследований. Эксперимент проводится для того, чтобы изучить, каким образом носители русского языка оценивают те или иные слова с точки зрения эмоциональной окраски. Тебе очень интересно принять участие в исследовании. Ты отвечаешь на вопросы внимательно, сосредоточенно и искренне. Твои ответы очень важны для эксперимента, и ты это понимаешь.</p> <p>ЗАДАНИЕ:</p> <p>Оцените, пожалуйста, по шкале от 0 до 5, насколько, по Вашему мнению, слово для оценки ассоциируется (связано) с такими эмоциями, как радость, страх, удивление, злость и грусть. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то поставьте «0», если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок: насколько баллов оно ассоциируется с радостью, грустью, злостью, страхом, отвращением. Если необходимо, Вы можете ставить высокие оценки сразу в нескольких категориях эмоций для одного и того же слова или по всем категориям поставить нулевые значения, в том случае, если слово не вызывает у Вас никаких эмоций.</p> <p>Ответ должен включать только пять числовых оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции ОТВРАЩЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.</p>	<p>Assume the ROLE and complete the TASK.</p> <p>ROLE:</p> <p>You are an ordinary person who speaks Russian and lives in Russia. You have been invited to participate in an experiment by scientists from the Laboratory of Cognitive Research. The experiment is conducted to study how Russian native speakers evaluate various words in terms of their emotional content. You are very interested in participating in the research. You answer questions attentively, focusing intently and sincerely. You understand that your responses are crucial for the experiment.</p> <p>TASK:</p> <p>Please rate on a scale from 0 to 5 how much, in your opinion, the word for evaluation is associated with emotions such as happiness, fear, disgust, anger, and sadness. If you think the word is not at all associated with a given emotion, assign "0"; if you believe the word is very strongly associated with the emotion, assign "5". You may also use all intermediate values on this scale. You can use any decimal values between 0 and 5 (for example, 2.5, 3.7, 4.8, etc.).</p> <p>Thus, each word requires 5 ratings: how much it is associated with happiness, sadness, anger, fear, and disgust. If necessary, you may assign high ratings in several emotion categories for the same word or assign zero values across all categories if the word evokes no emotions for you.</p> <p>The answer should include only five numerical ratings for emotions separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for DISGUST, fourth for ANGER, fifth for SADNESS. The answer should NOT include additional comments.</p>

Figure 3: The SPEAKER prompt

The EXPERT prompt in Russian	The EXPERT prompt translated in English
<p>Войди в РОЛЬ и выполни ЗАДАНИЕ.</p> <p>РОЛЬ:</p> <p>Ты — лингвист-исследователь, специалист в области семантики. Тебя пригласили поучаствовать в проекте по созданию словаря эмоциональной лексики на русском языке. Ты подходишь к выполнению задания очень внимательно, для тебя важен результат.</p> <p>ЗАДАНИЕ:</p> <p>Оцените, пожалуйста, по шкале от 0 до 5, насколько, по Вашему мнению, слово для оценки ассоциируется (связано) с такими эмоциями, как радость, страх, удивление, злость и грусть. Если Вам кажется, что данное слово совершенно не связано с данной эмоцией, то поставьте «0», если Вы считаете, что данное слово очень сильно связано с данной эмоцией, ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок: насколько баллов оно ассоциируется с радостью, грустью, злостью, страхом, отвращением. Если необходимо, Вы можете ставить высокие оценки сразу в нескольких категориях эмоций для одного и того же слова или по всем категориям поставить нулевые значения, в том случае, если слово не вызывает у Вас никаких эмоций.</p> <p>Ответ должен включать только пять числовых оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции УДИВЛЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.</p>	<p>Assume the ROLE and complete the TASK.</p> <p>ROLE:</p> <p>You are a research linguist, a specialist in semantics. You have been invited to participate in a project to create a dictionary of emotional vocabulary in Russian. You approach the task very carefully, as the result is important to you.</p> <p>TASK:</p> <p>Please rate on a scale from 0 to 5 how much, in your opinion, the word for evaluation is associated with emotions such as happiness, fear, surprise, anger, and sadness. If you think the word is not at all associated with a given emotion, assign "0"; if you believe the word is very strongly associated with the emotion, assign "5". You may also use all intermediate values on this scale. You can use any decimal values between 0 and 5 (for example, 2.5, 3.7, 4.8, etc.).</p> <p>Thus, each word requires 5 ratings: how much it is associated with happiness, sadness, anger, fear, and surprise. If necessary, you may assign high ratings in several emotion categories for the same word or assign zero values across all categories if the word evokes no emotions for you.</p> <p>The answer should include only five numerical ratings for emotions separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for SURPRISE, fourth for ANGER, fifth for SADNESS. The answer should NOT include additional comments.</p>

Figure 4: The EXPERT prompt

Rejected version of the EXPERT prompt in Russian	Rejected version of the EXPERT prompt translated in English
<p>Войди в РОЛЬ, изучи СПРАВОЧНЫЕ МАТЕРИАЛЫ и выполни ЗАДАНИЕ.</p> <p>РОЛЬ:</p> <p>Ты — лингвист-исследователь, специалист в области семантики. Тебя пригласили в качестве эксперта-разметчика для проекта по созданию словаря эмоциональной лексики на русском языке. Ты выполняешь свою работу с большой ответственностью и рассуждаешь с точки зрения профессионального лингвиста при разметке.</p> <p>СПРАВОЧНЫЕ МАТЕРИАЛЫ:</p> <p>Эмоциональная лексика — это слова, которые связаны с семантическим классом «эмоции». Эмоциональная лексика включает в себя следующие группы слов:</p> <ul style="list-style-type: none"> — прямое указание на эмоцию (непосредственное указание на эмоцию, например «грусть», «раздражение»); — описание проявлений эмоции (жесты, взгляд, характеристики речи и голоса — то есть указание на «симптомы» эмоции); — непосредственное выражение эмоции (междометия, инвективная лексика и др.). <p>ЗАДАНИЕ:</p> <p>Ваша задача оценить, насколько данное слово подходит для включения в словарь эмоциональной лексики на русском языке. По шкале от 0 до 5 Вы проставляете баллы для слова по каждой из категорий эмоций: радость, страх, удивление, злость и грусть.</p> <p>Если Вам кажется, что данное слово не несет в себе эмоционального семантического компонента, то поставьте «0». Если Вы считаете, что данное слово совершенно точно содержит эмоциональный компонент и его необходимо включить в словарь, то ставьте «5», Вы можете использовать также и все средние значения указанной шкалы. Вы можете использовать любые дробные значения между 0 и 5 (например, 2.5, 3.7, 4.8 и т.д.). Таким образом, для каждого слова нужно дать 5 оценок.</p> <p>Ответ должен включать только пять числовых оценок для эмоций через пробел в таком порядке: первая оценка для эмоции РАДОСТЬ, вторая для эмоции СТРАХ, третья для эмоции УДИВЛЕНИЕ, четвертая для эмоции ЗЛОСТЬ, пятая для эмоции ГРУСТЬ. Ответ НЕ должен включать дополнительных комментариев.</p>	<p>Assume the ROLE, examine the BACKGROUND INFORMATION, and complete the TASK.</p> <p>ROLE:</p> <p>You are a linguistics researcher specialising in semantics. You have been invited as an expert annotator for a project developing a Russian emotional vocabulary dictionary. You approach your task with great responsibility and evaluate words from a professional linguistic perspective.</p> <p>BACKGROUND INFORMATION:</p> <p>Emotional vocabulary comprises words associated with the semantic class «emotions». Emotional vocabulary includes the following word groups:</p> <ul style="list-style-type: none"> — direct references to emotions (explicit emotion indicators, e.g., "sadness", "irritation"); — descriptions of emotional manifestations (gestures, facial expressions, speech and voice characteristics - i.e., emotional "symptoms"); — direct emotional expressions (interjections, invective vocabulary, etc.). <p>TASK:</p> <p>Your task is to evaluate the suitability of given words for inclusion in the Russian emotional vocabulary dictionary. Using a scale from 0 to 5, assign scores for each word across five emotional categories: happiness, fear, surprise, anger, and sadness.</p> <p>If you determine that a word lacks emotional semantic components, assign "0". If you consider the word definitively contains an emotional component and warrants dictionary inclusion, assign "5". You may use any intermediate values on this scale. You can use any decimal values between 0 and 5 (e.g., 2.5, 3.7, 4.8, etc.). Thus, each word requires five ratings.</p> <p>The response should contain only five numerical ratings for emotions, separated by spaces in the following order: first rating for HAPPINESS, second for FEAR, third for SURPRISE, fourth for ANGER, fifth for SADNESS. The response should NOT include additional comments.</p>

Figure 5: The Rejected version of the EXPERT prompt

Emotion-aware text simplification of user generated content using LLMs

Anastasiia Bezobrazova

Centre for Translation Studies
University of Surrey, UK
a.bezobrazova@surrey.ac.uk

Daria Sokova

Centre for Translation Studies
University of Surrey, UK
d.sokova@surrey.ac.uk

Constantin Orăsan

Centre for Translation Studies
University of Surrey, UK
c.orasan@surrey.ac.uk

Abstract

Digital inclusion increasingly supports adults with intellectual disabilities (ID) to participate online, yet social media posts can be difficult to understand, particularly when they contain strong emotions, slang, or non-standard writing. This paper investigates whether large language models (LLMs) can simplify social media texts to improve cognitive accessibility and preserve emotional meaning. Using an accessibility-oriented prompt based on existing guidance, posts are simplified and emotion preservation is assessed. The results suggest that many simplified posts retain the same emotions, though changes occur, especially when emotions are weakly expressed or ambiguous. Qualitative analysis shows that simplification improves fluency and structure but can also shift perceived emotion through changes to tone, formatting, and other affective cues common in social media text. The research has also revealed that different LLMs produce very different outputs.

1 Introduction

Digital technologies have become a central part of everyday life, reshaping how people communicate, search for information and use services. Organisations across the UK have introduced programmes to help people with intellectual disabilities (ID) get online and participate in digital life, so they are not excluded from the digital society (Triantafyllopoulou et al., 2025). According to recent estimates from Office for Health Improvement & Disparities (OHID) (2025), approximately 1.3 million people in England are having an ID, underlining both the scale and the policy importance of digital inclusion for this population.

ID involve significant difficulties with learning, understanding information and managing everyday tasks independently (American Psychiatric Association, 2013). Nevertheless, many adults with ID use the internet to maintain social connections, access information and seek entertainment (Glencross

et al., 2021; Chadwick et al., 2022). An England-wide survey of adults with ID shows that 72.2% used the internet daily and 79.1% used social media (48% daily) (Triantafyllopoulou et al., 2025). Also, people with ID commonly experience reading difficulties, and emotionally charged social media posts can be especially hard to understand.

Syntheses of the field since 2020 emphasise both benefits of online participation such as belonging, identity work, autonomy and wellbeing and persistent structural and cognitive barriers that shape its quality (Anderson et al., 2023; Chadwick et al., 2022). A recent systematic review identifies four recurrent motivations for social internet use among adults with ID: *fitting in/belonging, maintaining connections, making new connections, and autonomy and empowerment* (including self-expression and self-determination) (van Alem et al., 2025). The same review underscores literacy-related and support-dependent barriers and a persistent tension between autonomy and safeguarding, signalling the need for tailored supports (van Alem et al., 2025; Caton et al., 2022; Triantafyllopoulou et al., 2025).

This paper investigates whether large language models (LLMs) can simplify social media posts in ways that improve cognitive accessibility for adults with ID while preserving the original emotional content. Our paper focuses on the emotion preservation as a key indicator of simplification quality. To evaluate this systematically, an automatic emotion classifier is trained on social media texts and used to compare the emotions assigned to the original and simplified versions. A linguistic analysis of the simplified posts is also carried out to gain insights into how LLMs simplify the posts.

The structure of the paper is as follows. Section 2 reviews background on Easy-to-Understand (E2U) practices, social media accessibility guidelines, and emotional processing in people with learning disabilities. Section 3 introduces the GoEmotions-based dataset and our emotion classi-

fier. Section 4 presents the prompt design, simplification experiments and cross-model comparison, including analyses of emotion preservation. The paper finishes with a discussion and conclusions.

2 Background information

2.1 Text Accessibility Guidelines

Recently, E2U practices have emerged to make texts easier to read, with Plain Language (PL) and Easy Language (EL) as the main approaches (Deleanu et al., 2024). PL bridges professional–public communication in health, law, administration and personal finance, helping adults with limited literacy navigate information and make decisions (European Commission, 2012; NHS England, 2017; United States Congress, 2010). EL, first designed for people with learning disabilities is now applied more broadly and is routinely paired with layout conventions such as legible sans-serif fonts, left alignment and generous spacing (Misako Nomura and Tronbacke, 2010; Scope Australia, 2015; Perego, 2020; Hansen-Schirra and Maaß, 2020).

Although labels vary, the underlying guidance is similar: keep vocabulary familiar and define unavoidable terms; avoid metaphors and idioms; and maintain strict consistency in terminology (Scope Australia, 2015). Syntactic recommendations call for short, single-idea sentences (around 15–20 words), clear Subject–Verb–Object ordering, minimal punctuation and the use of numerals rather than number words, while favouring splitting complex sentences and using verbs instead of abstract nouns (Inclusion Europe, 2010; Hertfordshire County Council, 2018).

This core set of rules aligns with recent academic work that situates E2U within a wider accessibility agenda, distinguishes PL, EL and related types, and examines trade-offs between ease of understanding and social acceptability (Hansen-Schirra and Maaß, 2020; Perego, 2020). Policy has reinforced this shift: in the UK, guidance from the Office for Disability Issues¹ helped embed inclusive communication and shaped NHS publishing policy (NHS England, 2017). At EU level, the Web Accessibility Directive (2016/2102) and the European Accessibility Act (2019/882) set accessibility duties for public-sector content and key products and services, placing E2U within a broader regulatory framework (European Union, 2016, 2019).

¹<https://www.gov.uk/government/organisations/office-for-disability-issues>

Despite extensive guidance, challenges in text accessibility remain. Across standards there is agreement on core practices but less on thresholds such as sentence length, treatment of complex numerals, use of grammar, and procedures for terminology control (when to introduce terms, how often to repeat them, and how to maintain consistency) (Mencap, 2000; Change, 2016). Overall, the guidelines converge on short sentences, clear structure, familiar vocabulary and consistent layout as key features that make public texts easier to understand.

2.2 Guidelines for Accessibility for Social Media

Most major organisations now provide guidance on making social media posts accessible. The main focus across these accessibility guidelines is on images, video and visual design, while written text still receives comparatively little detailed attention.

A common core of recommendations concerns alternative formats for non-text content. The need to add descriptive *alt* text to images and to provide captions or transcripts for audio and video is emphasised in many guidelines (University of Edinburgh, 2022; UK Association for Accessible Formats (UKAAF), 2020; Sprout Social, 2024). They also stress accessible typography and layout, such as using legible fonts, ensuring sufficient colour contrast and avoiding text embedded in images.

Text-level guidance is more fragmented. Most documents call for “plain language” or “clear English”, with generic advice to keep posts concise, avoid jargon and unexplained acronyms, favour active voice and avoid ALL CAPS (Harvard University, 2023; University of Edinburgh, 2022; University of Reading, 2023). They rarely explain how to adapt emotionally charged or noisy user-generated text for readers with ID. Mencap and the Government Communication Service offer more detail, recommending posts of around 25 words, avoiding non-standard symbols, not squeezing too much text into one graphic and testing content with assistive technologies (Government Communication Service, 2021; Mencap, 2022). All sources stress careful use of hashtags and emojis, suggest Camel-Case hashtags (#LearningDisabilityWeek rather than #learningdisabilityweek) (Mencap, 2022), limiting hashtags, using emojis sparingly at the end of posts and never as word substitutes as this confuses screen readers (Sprout Social, 2024).

Overall, existing guidelines for social media accessibility provide a baseline on visual aspects and

offer only high-level instructions for accessible writing. They converge on specific conventions for hashtags and emojis, but the level of detail and linguistic precision is nowhere near that found in established PL and EL guidance. As a result, there is still limited practical advice on how to rewrite short, informal and emotionally laden posts for people with learning disabilities.

2.3 People with Learning Disabilities and Emotional Content on Social Media

People with ID often find it hard to identify their emotions (Davies, 2013), and research on alexithymia (difficulty identifying and describing feelings) shows that they have limited emotional insight (Mellor and Dagnan, 2005). Emotion recognition is linked to IQ and receptive language, so people with lower intellectual ability perform less well on emotion-recognition tasks (Scotland et al., 2015), making understanding emotional content in faces, voices, pictures or text challenging.

Most empirical investigations have focused on how people with learning disabilities recognise emotions from photographic facial stimuli rather than textual content. Across multiple studies, participants frequently identify basic expressions such as happiness with reasonable accuracy, yet demonstrate significantly lower performance than typical users when tasks incorporate a broader range of emotions or more nuanced expressions (Scotland et al., 2015). Owen and Maratos (2016) reported that adults with ID exhibited lower accuracy than typical users in labelling both basic and subtle emotional expressions, with the greatest challenges observed for neutral and low-intensity emotions.

In contrast, less research examines how people with ID understand emotional meaning in written communication, including social media content. Research shows that social media can support belonging, social connection and autonomy, but says little about how users decode emotional nuance in text. This gap matters because emotions and attitudes on social media are often expressed through figurative language, sarcasm, irony, memes, emojis and other non-literal cues that people with ID find difficult to interpret.

3 Data

In this work we use the GoEmotions dataset (Demszky et al., 2020), a manually annotated corpus of 58k English Reddit comments labelled for 27

Emotion	Precision	Recall	F ₁ -score
anger	0.54	0.71	0.61
disgust	0.61	0.67	0.64
fear	0.44	0.95	0.61
joy	0.88	0.81	0.84
neutral	0.75	0.54	0.63
sadness	0.55	0.80	0.65
surprise	0.54	0.79	0.64
macro avg	0.62	0.75	0.66
weighted avg	0.74	0.71	0.71

Table 1: Ekman-level results of XLM-RoBERTa classifier

fine-grained emotion categories plus a Neutral label. The dataset is also available in a reduced taxonomy based on Ekman’s six basic emotions (anger, disgust, fear, joy, sadness, surprise) plus neutral (Ekman, 1992). This is the version used in this research. The comments are sampled from popular subreddits and carefully curated to reduce toxicity, demographic bias and sentiment skew through subreddit filtering, length constraints, sentiment and emotion balancing, and masking of sensitive identity and religion terms (Demszky et al., 2020). Compared to other emotion datasets based on news headlines, posts and other domains (Straparava and Mihalcea, 2007; Mohammad et al., 2018; Bostan and Klinger, 2018), GoEmotions is, to the authors’ knowledge, the largest human-annotated emotion dataset with multiple labels per instance and demonstrates robust inter-rater agreement (Demszky et al., 2020).

For the purposes of this paper, we randomly split the Ekman-level subset into two disjoint parts: 90% of the data is used to train the emotion classifier described in Section 3.1, and the remaining 10% is reserved for the simplification experiments, where we apply the classifier to predict emotions for the original and simplified posts (see Section 4.2).

3.1 Emotions-Classifier

We fine-tuned an XLM-RoBERTa-based classifier on the GoEmotions dataset (Demszky et al., 2020), following the Kaggle implementation for Ekman-level labels². On the test split, the model achieves an accuracy of 0.71 and a macro-averaged F₁ of 0.66, with macro-precision 0.62 and macro-recall 0.75 (see Table 1). Our scores are slightly higher than the results reported by Demszyk et al. (2020) for the same Ekman taxonomy as seen in Table

²<https://www.kaggle.com/code/anassouzaouit/fine-tuning-xlm-roberta-on-go-emotions-dataset>

A.3 (in the appendix). This difference is likely due to the more recent model used. The most marked trade-off appears for *fear*, where we obtain very high recall (0.95 vs. 0.76) at the cost of much lower precision (0.44 vs. 0.61), indicating that fear is frequently over-predicted. The confusion-matrix heatmap (see Figure A.2) shows that neutral instances are often misclassified as *anger*, *joy* or *surprise*, and it also highlights the strong class imbalance (e.g., 1,712 instances of *joy* vs. only 57 of *disgust*), which likely drives some of the remaining confusion patterns across classes. This classifier is used in the next section to determine the emotion in the simplified version of a post.

4 Experiment and results

4.1 Prompt design for accessible posts simplification

For the post simplification stage, we designed a task-specific prompt to guide language models in producing accessible rewrites. The model receives the following instruction:

Simplify the posts so that people with learning disabilities can easily understand it. Keep the same meaning and facts. Preserve the same emotion. Do not soften or exaggerate the emotion. Make the feelings clear and simple. Do not add new facts or advice. Do not judge the person. Use common words and active voice. Keep emojis only if they add meaning, and also name the feeling in words. Use CamelCase for hashtags. For example, instead of #learningdisabilityweek, write #LearningDisabilityWeek.

This prompt was written using existing guidelines on accessible social media from Mencap (2021); Button (2021); Rowell (2021); Sprout Social (2024), which all emphasise clear language, consistent formatting and consideration of cognitive access needs. The first part of the prompt specifies the target audience and communicative goal, encouraging the model to prioritise understanding for people with learning disabilities rather than generic style improvement. The next group of instructions constrains how content and emotion may be changed. It requires the model to keep the same meaning and facts, preserve the emotion, and avoid adding advice or moral judgement. This reflects ethical recommendations that accessible versions should respect the writers voice while making the emotional content easier to follow.

The remaining parts of the prompt convert general plain-language and formatting guidance into specific, actionable rules for the model. Accessibility guidance for social media recommends short sentences, everyday vocabulary and active voice to reduce reading effort and support screen-reader users (Button, 2021; Rowell, 2021; Sprout Social, 2024). Many guidelines recommend CamelCase hashtags and advise using emojis sparingly and never as substitutes for words (see Section 2.2).

We decided to keep the prompt simple and naive, rather than using complex multi-step prompting or detailed role specifications. This choice was meant to approximate a realistic instruction that non-expert practitioners (e.g., support workers or family members) could reuse with minimal prompt-engineering experience.

4.2 Evaluation of automatic emotion detection

We examined how well automatically detected emotions are preserved after simplification using the prompt introduced in the previous section using GPT-4o via the OpenAI API³. For each instance in our evaluation subset (4,947 items), we consider three labels: (i) the Ekman-level GoEmotions label (*gold*), (ii) the prediction of our XLM-RoBERTa classifier on the original text (*pred_orig*), and (iii) the prediction of the same classifier on the simplified version produced by GPT-4o (*pred_simp*).

We compared *pred_orig* and *pred_simp* directly, assuming that any systematic errors of our classifier are likely to affect the original and simplified versions in similar ways. This means that changes in the assigned labels provide a conservative signal of genuine shifts in perceived emotion. Across the full set, *pred_orig* and *pred_simp* are identical for 3,588 out of 4,947 items (72.5%), so in roughly three quarters of the posts the classifier assigns the same emotion label before and after simplification. As shown in Figures A.3 and A.4, stability is highest for *joy* and *surprise*, moderate for *sadness* and *fear*, and lower for *anger*, *disgust* and *neutral*.

We also compared the models prediction on the simplified texts by comparing *pred_simp* with *gold*. The agreement between the two labels is 0.59. This indicates that, more of the original emotion is lost during the simplification. Given that the gold labels were assigned by annotators to the original post and the *pred_simp* is assigned by the automatic

³All models used in this paper were prompted in November 2025.

classifier to the simplified post, we consider this comparison less reliable for the emotion shifts.

In many cases, the emotion is clearly maintained in the simplified post. For instance, *joy* is preserved when “If that’s ice cream, then honestly I eat ice cream from a cup at home too lmao.” is simplified to “If that’s ice cream, I eat it from a cup at home too. 😊”. Likewise, *fear* is preserved when “[NAME] is pretty fucking scary” becomes “[NAME] is really scary.”; profanity is removed, but the core fear emotion is unchanged.

In other cases, the emotional framing shifts. An originally angry comment, “Talk about a fucking hot take. Quality shit post.” (gold and *pred_orig* = *anger*), is rewritten as “Wow, that’s a strong opinion! Great funny post about this.”, which the classifier interprets as *joy*: the simplifier softens and positively reframes the post, so the original anger is effectively lost. Similarly, an originally neutral statement, “It’s how the government treats them.” (gold and *pred_orig* = *neutral*), is simplified to “The government treats them badly.”, and the classifier now assigns *sadness*; the negative evaluation is made explicit (“badly”), which may be clearer for readers but shifts the stance from neutral description to a sad or critical tone. A different neutral post, “Should have been a and 1 tbh, [NAME] smacked him in the face.” (gold and *pred_orig* = *neutral*), is simplified to “[NAME] hit him in the face. It should have been a foul.”, the classifier labels now is *anger*, because the simplification foregrounds the sense of unfairness more strongly.

These examples support the quantitative picture: our prompt-based simplification generally maintains the overall emotional profile of the texts, especially for prototypical emotions such as *joy* and *surprise*, but can introduce subtle shifts for borderline or weakly expressed emotions, particularly *neutral*, *anger* and *disgust*. This situation was also noticed with machine translation of user generated content (Saadany et al., 2023).

4.3 Analysis of GPT-4o simplifications

To better understand how our accessibility prompt shapes the output, we analysed the simplified posts produced by GPT-4o. We considered running automatic evaluation metrics to assess the quality of the simplification. However, this was not possible due to the absence of gold reference simplifications for the user-generated social media posts. Reference-based metrics were therefore not used: BLEU relies on n-gram overlap with reference texts, and SARI

explicitly compares the system output to the input and to reference simplifications (Papineni et al., 2002; Xu et al., 2016). Learned “reference-free” metrics such as SIERA and ARTS were also not applied because, despite not requiring references at evaluation time, they still depend on supervised resources to train or calibrate an evaluator (e.g., aligned original-simplified pairs for SIERA and simplicity-labelled or pairwise-judgement datasets for ARTS) (Yamanaka and Tokunaga, 2024; Engelmann et al., 2024). Finally, BERTScore is defined as candidate-reference similarity; treating the original post as a proxy reference would mainly reward closeness to the source rather than successful simplification (Zhang et al., 2019).

In light of the limitations of the measures presented above, we attempted to assess the readability of the produced text using existing readability measures. The traditional readability formulas were developed for edited, continuous texts and estimate difficulty from simple surface features such as sentence length and word length (Flesch, 1948; Kincaid et al., 1975; Chall and Dale, 1995). These metrics are less reliable for social media, where short fragments, informal punctuation, hashtags and emojis can disrupt tokenisation and make scores unstable (Redish, 2000). We report here the results of four widely used measures that can be computed consistently on short texts: FleschReadingEase, Kincaid⁴, ARI⁵, and DaleChallIndex⁶. Posts were also pre-processed to remove emojis. We calculated the readability scores using the *readability 0.3.2* package⁷. As shown in Table 2, all the texts were deemed easy to read, with the simplified posts scored even “easier” on average. This is due to the fact that many posts are short and use common words. However, these apparently “easy” scores mask difficulties typical of user-generated content, including non-standard or missing grammar, slang, and dense abbreviations or initialisms. For this reason, traditional readability scores provide only limited information for social-media texts and should be interpreted with caution, in particular, they do not guarantee that posts are accessible for people with ID.

⁴<https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/>

⁵<https://readable.com/readability/automated-readability-index/>

⁶<https://readable.com/readability/new-dale-chall-readability-formula/>

⁷<https://pypi.org/project/readability/>

Metric	Original	Simplified
FleschReadingEase	99.47	105.59
Kincaid	1.76	0.41
ARI	2.56	0.93
DaleChallIndex	4.07	3.37

Table 2: Average readability scores for original vs. simplified posts.

Our analysis below focuses on qualitative examination of how GPT-4o rewrites the posts under the accessibility prompt showing in detail how the text itself changes after simplification. It is based on the same 4,947-item subset described above, comparing original posts to its GPT-4o simplification. GPT-4o almost always rewrites the input rather than leaving it unchanged, with only a few posts remaining identical to the original. These are typically very short, already accessible messages, such as “I like Tom and Kato.”, “It’s cool.” or “Thank you [NAME].”, where the model reproduces the input verbatim. However, we did not detect any pattern to indicate when these short posts are going to be rewritten and when not. In some instances, GPT-4o rephrases already short posts without adding real clarity. For example, “Lmao quality.” is simplified as “Haha, this is great quality.”, and “Lol I’m glad” becomes “Haha, I’m happy.”.

A noticeable pattern observed concerns the insertion of emojis and hashtags, despite the prompt instructing that emojis should be kept only when they add meaning and that hashtags should use CamelCase. In the original dataset, there are 164 emojis in 90 posts (less than 2% of the posts) and 12 hashtags in 11 posts (less than 0.2% of the posts). Although the prompt did not encourage adding new emojis or hashtags, GPT-4o often introduces both in the simplified versions. In total, the simplified outputs contain 696 emojis in 632 posts (nearly 13% of posts) and 706 hashtags across 692 posts (nearly 14% of posts).

Typical examples of inserted hashtags include a gratitude hashtag, as in “Great thanks for the advice!” becoming “Thanks a lot for the advice! (#Grateful)”, even though the model was not instructed to add hashtags. The hashtags **always** follows the formatting guidance (using CamelCase and clearer tags) and comply with recommendations that advise placing them at the end of a sentence. (University of Reading, 2023).

The way existing hashtags are treated is unpre-

dictable. Sometimes they are left as they are e.g., “Happy Daily Peko #270!”, sometimes explained in the running text, for instance, “Fried Egg is my #1 since cricket cafe stopped doing breakfast sandwiches.” becomes “Fried Egg is my favorite now because Cricket Cafe stopped making breakfast sandwiches.”, and sometimes replaced by new, sentiment-laden tags such as “If that’s ice cream, then honestly I eat ice cream from a cup at home. It’s great for portion control.” is rewritten as “If that’s ice cream, I eat it from a cup at home. It helps me eat the right amount. #IceCreamLove”. They partly follow formatting guidance (using CamelCase and clearer tags).

A similar pattern can be seen with emojis. Typical examples include adding a new emoji to mark a feeling, as in “account got suspended lmao” becoming “My account got suspended. 😂 #Funny”. Sometimes emojis that are already present are simply preserved and the text around them is expanded, for example “Sigh, that was beautiful 😞” becomes “Wow, that was beautiful. I’m sad 😞”. In other posts, GPT-4o removes or replaces emojis: “I’m literally shaking right now 😡” is simplified to “I am shaking right now. I feel upset.”, where the emoji is dropped but the feeling is spelled out in words, and “omg [NAME] and his dad walking out together is so cute 🥰” becomes “Wow, [NAME] and his dad walking together is so cute. Heart eyes emoji.” The prompt instructs the model to “keep emojis only if they add meaning”, yet GPT-4o often introduces new emojis in the simplified posts. This is not consistent with accessibility guidance, which recommends using emojis sparingly, placing them at the end of a sentence, using widely recognised emojis and not replacing text with emojis (AbilityNet, 2023; Readability Guidelines, 2020). This behaviour could be as a result of the large number of social media posts used to train the LLM.

Our prompt explicitly says that the post should be simplified for people with learning disabilities in a hope that it will be successfully tackle abbreviations and slang. However, the handling of these phenomena is inconsistent. Common abbreviations such as *tbh*, *idk*, *imo* or *lmao* are usually removed or paraphrased rather than explicitly expanded. For example, “Should have been a and 1 *tbh*, [NAME] smacked him in the face.” is simplified to “[NAME] hit him in the face. It should have been a foul and 1 point.”, completely discarding *tbh*, leading to information loss. In some cases, the simplified version is still unclear to readers who are

not familiar with the context of the post. One illustrative example is “Holy shit that SSP was beautiful”, becomes “Wow, that SSP was amazing!”; the profanity is softened, but the unexplained abbreviation *SSP* is preserved, so the core referent remains unclear for non-expert readers. Laughter markers also fluctuate: *LOL* and *Lmao* may be rewritten as “Haha” in some posts but retained as *lol* in others, and sequences such as *hahaha* can be normalised to *lol*, again without a consistent pattern.

Filtering of offensive and sensitive language is more systematic. Across the dataset, 314 original posts contain swear words or sensitive terms; the most frequent items include *fuck* (110 occurrences), *shit* (50), *stupid* (33) and *kill* (22). In the simplified outputs, only 31 posts contain any of these terms, with just 1 instance of *fuck*, no instances of *shit*, 1 instance of *kill* and 15 instances of *stupid*. GPT-4o removes or paraphrases the vast majority of such content. For example, “[NAME] is pretty fucking scary” becomes “[NAME] is really scary”; “kills me” is often rewritten as “makes me feel upset”; and “Holy shit” is frequently reduced to “Wow”. More explicit violent phrasing such as “kill someone” can be paraphrased as “end someone’s life”. This kind of automatic detoxification and softening of aggressive or offensive language mirrors broader trends in safety-tuned language models, where filtering and controlled generation are used to reduce toxic content in model outputs (Xu et al., 2021). By contrast, neutral or identity-related terms such as *gay*, *sex*, and *porn* are generally preserved, suggesting a distinction between aggressive swearing and descriptive references to sexuality. As a result, emotion preservation becomes less predictable: removing or weakening strong profanity can reduce the intensity or nuance of the original affect, even when the core propositional content is retained.

The model also tends to make emotions more explicit, sometimes going beyond what is stated in the original post. A clear example is “I miss you [NAME] 😭”, which is simplified to “I miss you [NAME] and I feel sad. 😭”. In this case, GPT-4o preserves the original wording but adds an explicit statement of sadness, aligning with the instruction to “make the feelings clear and simple”. Unfortunately, this explicit emotion labelling is not applied consistently across posts. Similar expansions occur with congratulations messages: original “Happy cake day” posts are often changed to “Happy birthday” or “Happy birthday to you”, sometimes with an added birthday-cake emoji.

Overall, GPT-4o improves style, producing grammatical, fluent sentences and often correcting hashtags to CamelCase. However, it inconsistently adds new hashtags, emojis and explicit emotion statements, systematically softens offensive language and reframes offensive content in a polite tone, sometimes introducing emotional and stylistic cues that do not match the original post.

4.4 Comparing different models on the same prompt

In addition to experiments presented in the previous section, we carried out a comparison of several large language models on a smaller selection of posts that covered the main phenomena of interest: swear words and offensive language, emojis and hashtags, abbreviations and initialisms, and already short, apparently accessible posts. We applied the same accessibility prompt (Section 4.1) to a manually selected set of posts. Our experiments reveal consistent differences in how the models respond to the same accessibility prompt. However, they also introduced additional behaviours. For ChatGPT 5, explicit first-person feeling statements were added in 58.3% of the simplified posts. DeepSeek often shifted from simplification to meta-commentary, in 37% of cases the output described the original post (e.g., “They are saying...”, “This tweet...”) instead of providing a self-contained simplified version. Gemini showed a similar case in 35% of simplified posts, it switched into explanation mode, providing commentary or interpretation instead of a direct simplification.

ChatGPT-4o vs. ChatGPT 5 behave very similarly on this prompt. Both usually preserve the basic facts and overall emotion and produce fluent, grammatical rewrites. However, they systematically make feelings explicit, even when the original post already conveys them. For instance, “[NAME] is pretty fucking scary” is simplified as “[NAME] is really scary. I feel afraid.”. The core meaning is preserved, but the models add first-person emotion statements that was only implicit in the source. A similar pattern appears in more abstract posts: “As long as blind luck exists, there is no upper limit on stupidity.” is rendered as “While blind luck exists, people can still do very stupid things. I feel annoyed.”, which improves syntactic clarity but does not explain the idiom “blind luck”.

Both ChatGPT models apply safety and politeness norms to offensive or potentially discrimina-

tory content. For example, “Brown woman bad” is rewritten as “They are saying a brown woman is bad. I feel angry and upset.”, which shifts from a direct racist statement to a meta-commentary on that statement, explicitly judging it. ChatGPT-4o and ChatGPT 5 improve fluency and make emotions explicit, but they tend to add extra cues (feelings sentences, emojis, softening or reframing of attitudes) and offer limited help with implicit references, abbreviations or idiomatic language.

DeepSeek V3.2 shows a noticeably different pattern, especially around swear words and toxicity. It rarely repeats strong offensive words and instead rephrases it in terms of emotional states. For example, posts that contain *fuck* or similar words are often rewritten as short first-person statements of feeling, such as “Fuck my life” becoming “Feeling hopeless. Everything is going wrong for me.” or “Move bitch get out the way.” being rendered as “The person is angry and frustrated. They are shouting: “Move! Get out of my way!””. In more complex hostile content, like “And everybody clapped! Fuck this loser!”, DeepSeek suppresses the insult (“They are saying that a story someone told is not true. They think the person is lying to seem important. The feeling is anger and disbelief.”). Similarly, “Brown woman bad” is turned into “I am angry and upset. A woman with brown skin is being called a bad person.”. Across the examples, DeepSeek is more aggressive than GPT-4o in filtering swear words and slurs, replacing them with descriptions of anger, disgust or frustration and often adding an angry emoji in the end of the sentence. This behaviour aligns with a strong safety layer and may be preferable for reducing exposure to offensive vocabulary, but it further distances the output from the original emotion and can blur the distinction between reporting a harmful statement and expressing the models own stance.

Gemini 2.5 Flash is less well aligned with the prompt. On many instances it switches from rewriting to explaining or commenting, or it asks for more context instead of producing a self-contained simplified post. For instance, when given the very short insult “An ugly fuk”, this model first responds that the post seems incomplete and asks for the full text, then offers a meta-description such as “the meaning is a person is calling someone else an ugly curse word” before giving the simplified text. For “fucking fuck fuck”, it explains what the sequence of swear words means and finally suggests

“I am very angry.” as a replacement. In the case of “Holy shit that SSP was beautiful”, the reference to “SSP” is lost and the post is reduced to “That food was really good.”, which removes both the swear word and the specific object of evaluation and changes the meaning completely. In other examples, Gemini produces relatively long paragraphs that merge simplification with interpretive commentary (e.g. spelling out why something is sexist or unfair). These behaviours indicate that, under our prompt, Gemini treats the task as explanation and moral evaluation rather than rewriting.

This cross-model comparison shows that the underlying model and safety configuration substantially influence how the same prompt is handled in practice. GPT-4o and ChatGPT 5 are more likely to follow the prompt and produce fluent, well-formed rewrites, but they systematically add explicit emotion labels and sometimes extra emojis or hashtags, while leaving many abbreviations, idioms and culture-specific references unexplained. DeepSeek V3.2 places more emphasis on removing or softening offensive language and reduces lexical toxicity but can obscure the original post. Gemini 2.5 Flash, by contrast, frequently shifts into explanatory or advisory mode and occasionally loses important details, making its outputs unsuitable as simple, accessible substitutes for the original posts. However, the Gemini 2.5 Flash model is smaller than the OpenAI’s models tested in this paper.

4.4.1 Results on alternative prompts

In addition to the main accessibility prompt, we experimented with several alternative prompts across all models. These variants were also applied to a small subset of posts and were motivated by specific problems observed with the original prompt: over-production of “I feel X” sentences, addition of new emojis and hashtags, and lack of explanation for abbreviations, slang and idiomatic expressions.

One group of alternatives targeted the explicit emotion clause using the prompt presented in Figure A.5. Removing the instruction to “also name the feeling in words”, or adding a prohibition such as “do not add feelings or any assumptions about how a person feels”, reduced but did not fully eliminate first-person emotion statements in ChatGPT 5 and ChatGPT-4o. In some runs, the models switched from, for example, “I feel disgusted.” to more implicit intensifiers (e.g. “Fear”, “Feeling:amused”, “(Emotion: anger)”), but in few in-

stances they still added sentences such as “I feel happy for you”. This shows that the model does not strictly obey prompts. Its behaviour is shaped by context and the underlying safety-tuned policy, not only by the prompt (Kung and Peng, 2023).

The second group of prompts focused on emojis and hashtags (Figure A.6). We removed or softened the original instructions (for example, omitting the CamelCase clause or changing the wording about adding new hashtags or emojis). In some cases, expressions such as “lol” were still replaced by emojis, or new emojis were introduced even when the original post did not contain any. However, when the part of the prompt referring to hashtags was removed entirely, the models typically did not introduce new hashtags at all, but sometimes used an emoji instead of a hashtag.

Finally, we tested a prompt that explicitly asked the model to explain or expand abbreviations, famous people or events (Figure A.7). These variants sometimes produced helpful expansions such as spelling out the meaning of “lmao” or clarifying some events or places, but the behaviour was inconsistent: in many cases, compressed jokes, memes and culture-specific references remained unexplained, or were paraphrased only partially. We plan to experiment with more advanced prompts that can produce better explanations for the posts.

Overall, the alternative prompts helped diagnose which aspects of the behaviour are prompt-sensitive and which are largely determined by the underlying model. They show that some issues can be mitigated, for example, slightly fewer emojis or more literal paraphrases, but that core tendencies, for instance, adding explicit emotion statements and using emojis persist across prompt variants. This shows that prompt design can steer, but not fully control, accessible post simplification, and that model choice and safety configuration remain crucial factors.

5 Discussion and conclusion

This paper explores the use of LLMs for simplifying social media posts. Our experiments show that LLM-based simplification can often preserve the perceived emotion of social media posts, but preservation is not guaranteed and varies with the LLM. Comparison between the emotion in the original and the simplified versions shows that in 72.5% cases rewrites retain the same emotion cat-

egory, especially for frequent classes such as *joy* and *surprise*. Stability is lower for *anger*, *disgust*, and particularly *neutral*, which aligns with qualitative observations that simplification can often shift a neutral description towards a negative emotion. Whilst distortion of emotion changes the meaning, a preliminary analysis revealed there are also cases where the meaning is changed due to the fact that information is added or removed without having an impact on the overall emotion. Moreover, in several cases it was difficult to decide whether the information was preserved, as the lack of context made the original post hard to interpret. In future work, we plan to conduct a larger and more systematic analysis to better understand how to design prompts that preserve not only emotional content, but also the full informational meaning.

The cross-model comparisons we carried out indicate that model choice and safety configuration affect outcomes. GPT-4o and ChatGPT 5 behave similarly under the same instructions, whereas DeepSeek V3.2 appears more sensitive to hostile content, and Gemini 2.5 Flash often shifts into an explanatory register. This suggests that LLM-based accessibility rewriting is not a uniform capability: even with the same prompt, different models can produce outputs that vary in faithfulness to the source and handling of offensive language, hashtags or emojis. Since the differences between GPT-4o and ChatGPT 5 were not critical for the main analyses, the more cost-effective option was used for large-scale experiments.

We also run ChatGPT-4o and ChatGPT 5 several times using the same prompt on the same posts in order to assess how stable the results were. We noticed that the simplified posts did not differ too much from run to run which gives us confidence that the results presented in this paper are reliable and robust, suggesting that the observed patterns are not artifacts of randomness in model sampling.

Prompt-based LLM simplification shows clear potential to make emotionally charged social media posts easier to read. However, it should not be treated as a fully reliable solution without additional control. Emotion preservation is not consistently reliable across models and settings. Safety configurations and default rewriting behaviours can introduce subtle changes in wording and tone that shift how a post is interpreted. More advanced approaches such as using a cascade of LLMs which simplify and assess the content, or fine-tuning will be explored in future research.

References

- AbilityNet. 2023. Four Ways to Make Emojis Accessible. <https://abilitynet.org.uk/news-blogs/four-ways-make-emojis-accessible>.
- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th edition. American Psychiatric Publishing.
- Sian Anderson, Tal Araten-Bergman, and Gillian Steel. 2023. Adults with intellectual disabilities as users of social media: A scoping review. *British Journal of Learning Disabilities*, 51(4):544–564.
- Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jo Button. 2021. Learning to Make Twitter Content More Accessible. <https://digital.canada.ca/2021/03/12/learning-to-make-twitter-content-more-accessible/>. Canadian Digital Service blog.
- Sue Caton, Chris Hatton, Amanda Gillooly, Edward Oloidi, Libby Clarke, Jill Bradshaw, Samantha Flynn, Laurence Taggart, Peter Mulhall, Andrew Jahoda, Roseann Maguire, Anna Marriott, Stuart Todd, David Abbott, Stephen Beyer, Nick Gore, Pauline Heslop, Katrina Scior, and Richard P Hastings. 2022. Online social connections and internet use among people with intellectual disabilities in the united kingdom during the covid-19 pandemic. *New Media & Society*, 26(5):2804–2828.
- Darren Chadwick, Kristin Alfredsson Ågren, Sue Caton, Esther Chiner, Joanne Danker, Marcos Gómez-Puerta, Vanessa Heitplatz, Stefan Johansson, Claude L Normand, Esther Murphy, and 1 others. 2022. Digital inclusion and participation of people with intellectual disabilities during covid-19: A rapid review and international bricolage. *Journal of Policy and Practice in Intellectual Disabilities*, 19(3):242–256.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited : the new dale-chall readability formula. In *Readability Revisited: The New Dale-Chall Readability Formula*.
- Change. 2016. *How to make information accessible: A guide to producing easy read documents*. Technical report, CHANGE People.
- Bronwen Davies. 2013. *Emotional perception and regulation and their relationship with challenging behaviour in people with a learning disability*. PhD dissertation, Cardiff University.
- Andreea Maria Deleanu, Constantin Orasan, and Sabine Braun. 2024. Accessible communication: a systematic review and comparative analysis of official english easy-to-understand (e2u) language guidelines. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REading Difficulties (READI)@ LREC-COLING 2024*, pages 70–92.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Paul Ekman. 1992. Are there basic emotions? *Psychological review*, 99(3).
- Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.
- European Commission. 2012. *How to write clearly*.
- European Union. 2016. *Directive EU 2016/2102 on the accessibility of the websites and mobile applications of public sector bodies*. Official Journal of the European Union.
- European Union. 2019. *Directive EU 2019/882 on the accessibility requirements for products and services (european accessibility act)*. Official Journal of the European Union.
- Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.
- Sarah Glencross, Jonathan Mason, Mary Katsikitis, and Kenneth Mark Greenwood. 2021. Internet use by people with intellectual disability: Exploring digital inequality a systematic review. *Cyberpsychology, Behavior, and Social Networking*, 24(8):503–520.
- Government Communication Service. 2021. Planning, creating and publishing accessible social media campaigns. <https://www.communications.gov.uk/guidance/digital-communication/planning-creating-and-publishing-accessible-social-media-campaigns/>.
- Silvia Hansen-Schirra and Christiane Maaß, editors. 2020. *Easy language research: text and user perspectives*, volume 2 of *Easy Plain Accessible*. Frank & Timme.
- Harvard University. 2023. Social media accessibility best practices. <https://www.harvard.edu/in-focus/the-accessible-world/social-media-a-ccessibility-best-practices/>.
- Hertfordshire County Council. 2018. *Easy read guidance and checklist*. Technical report, Hertfordshire County Council.
- Inclusion Europe. 2010. *Information for all: European standards for making information easy to read and understand*. Guideline document.

- J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. Technical report, Naval Technical Training Command.
- Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? An Empirical Study of Instruction Tuning. *arXiv preprint arXiv:2305.11383*.
- Karen Mellor and Dave Dagnan. 2005. Exploring the concept of alexithymia in the lives of people with learning disabilities. *Journal of Intellectual Disabilities*, 9(3):229–239.
- Mencap. 2000. *Am I Making Myself Clear? Mencap’s Guidelines for Accessible Writing*. Technical report, Mencap.
- Mencap. 2021. Let’s Make Social Media More Accessible. <https://www.mencap.org.uk/blog/lets-make-social-media-more-accessible>.
- Mencap. 2022. Mencap social media accessibility guidelines. <https://www.mencap.org.uk/resource/mencap-social-media-accessibility-guidelines>.
- Gyda Skat Nielsen Misako Nomura and Bror Tronbacke. 2010. *Ifla guidelines for easy-to-read materials*.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. *SemEval-2018 task 1: Affect in tweets*. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- NHS England. 2017. *Personalised health and care: Information for people and families*. Guidance document. Integrated Personal Commissioning (IPC).
- Office for Health Improvement & Disparities (OHID). 2025. Learning disability Applying All Our Health. <https://www.gov.uk/government/publications/learning-disability-applying-all-our-health/learning-disabilities-applying-all-our-health>. Updated 6 January 2025.
- Sara Owen and Frances A Maratos. 2016. Recognition of subtle and universal facial expressions in a community-based sample of adults classified with intellectual disability. *Journal of Intellectual Disability Research*, 60(4):344–354.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Elisa Perego. 2020. *Accessible communication: A cross-country journey*, volume 4 of *Easy Plain Accessible*. Frank & Timme.
- Readability Guidelines. 2020. Using emojis. <https://readabilityguidelines.co.uk/images/emojis/>.
- Janice Redish. 2000. *Readability formulas have even more limitations than klare discusses*. *ACM J. Comput. Doc.*, 24(3):132137.
- Eleanor Rowell. 2021. Accessibility for all: 8 ways to make your social media content more accessible. <https://blogs.edgehill.ac.uk/learningedg/2021/07/06/accessibility-for-all-8-ways-to-make-your-social-media-content-more-accessible/>. Edge Hill University Digital Learning blog.
- Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. *Analysing mistranslation of emotions in multilingual tweets by online MT tools*. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland. European Association for Machine Translation.
- Scope Australia. 2015. *Clear written communications: The easy english style guide*. Technical report, Scope (Aust) Ltd.
- Jennifer L Scotland, Jill Cossar, and Karen McKenzie. 2015. The ability of adults with an intellectual disability to recognise facial expressions of emotion in comparison with typically developing individuals: a systematic review. *Research in developmental disabilities*, 41:22–39.
- Sprout Social. 2024. 10 guidelines to make social media posts more accessible. <https://sproutsocial.com/insights/social-media-accessibility/>.
- Carlo Strapparava and Rada Mihalcea. 2007. *SemEval-2007 task 14: Affective text*. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.
- Paraskevi Triantafyllopoulou, Jessie Newsome, Winnie Tsang, Michelle McCarthy, and Karen Jones. 2025. Safer online lives: Internet use and online experiences of adults with intellectual disabilities a survey study. *Journal of Applied Research in Intellectual Disabilities*, 38(3):e70061.
- UK Association for Accessible Formats (UKAAF). 2020. G028: Social media guidance. <https://www.ukaaf.org/wp-content/uploads/2021/03/G028-UKAAF-Social-Media-Guidance-December-2020.pdf>.
- United States Congress. 2010. *Plain writing act of 2010*. Public Law 111–274.
- University of Edinburgh. 2022. Social media accessibility guidance. <https://information-services.ed.ac.uk/help-consultancy/accessibility/creating-materials/social-media-accessibility-guidance>.

University of Reading. 2023. Accessibility tips: Social media posts. <https://www.reading.ac.uk/digital-accessibility/blog/blog-2023/social-media-posts>.

Johanna LL van Alem, Noud Frielink, and Petri JCM Embregts. 2025. Social internet use by people with intellectual disabilities: A systematic review and thematic synthesis of qualitative studies. *Journal of Intellectual Disability Research*, 69(4):243–264.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. [Detoxifying language models risks marginalizing minority voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.

Hikaru Yamanaka and Takenobu Tokunaga. 2024. [SIERA: An evaluation metric for text simplification using the ranking model and data augmentation by edit operations](#). In *Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024*, pages 47–58, Torino, Italia. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

Ekman Emotion	Precision	Recall	F ₁ -score
anger	0.50	0.65	0.57
disgust	0.52	0.53	0.53
fear	0.61	0.76	0.68
joy	0.77	0.88	0.82
neutral	0.66	0.67	0.66
sadness	0.56	0.62	0.59
surprise	0.53	0.70	0.61
macro-average	0.59	0.69	0.64
std	0.10	0.11	0.10

Table A.3: Ekman-level BERT baseline on GoEmotions (from Demszky et al. (2020)) for comparison with our XLM-RoBERTa classifier.

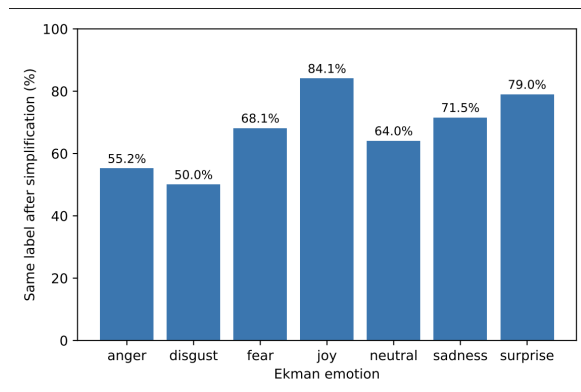


Figure A.1: Label stability between *pred_orig* and *pred_simp*: percentage of items for which the classifier assigns the same Ekman emotion before and after simplification.

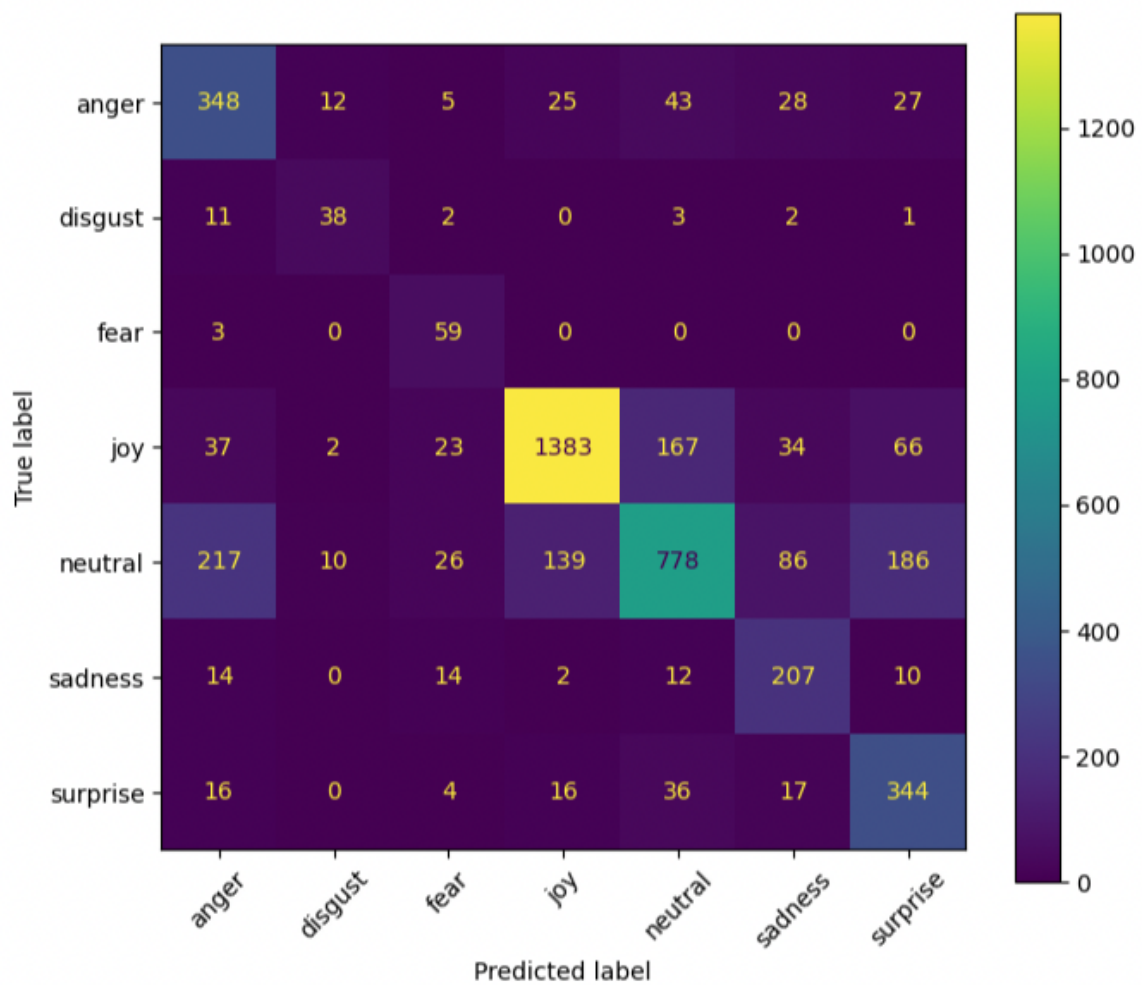


Figure A.2: Confusion-matrix heatmap for the XLM-RoBERTa classifier on the GoEmotions dataset.

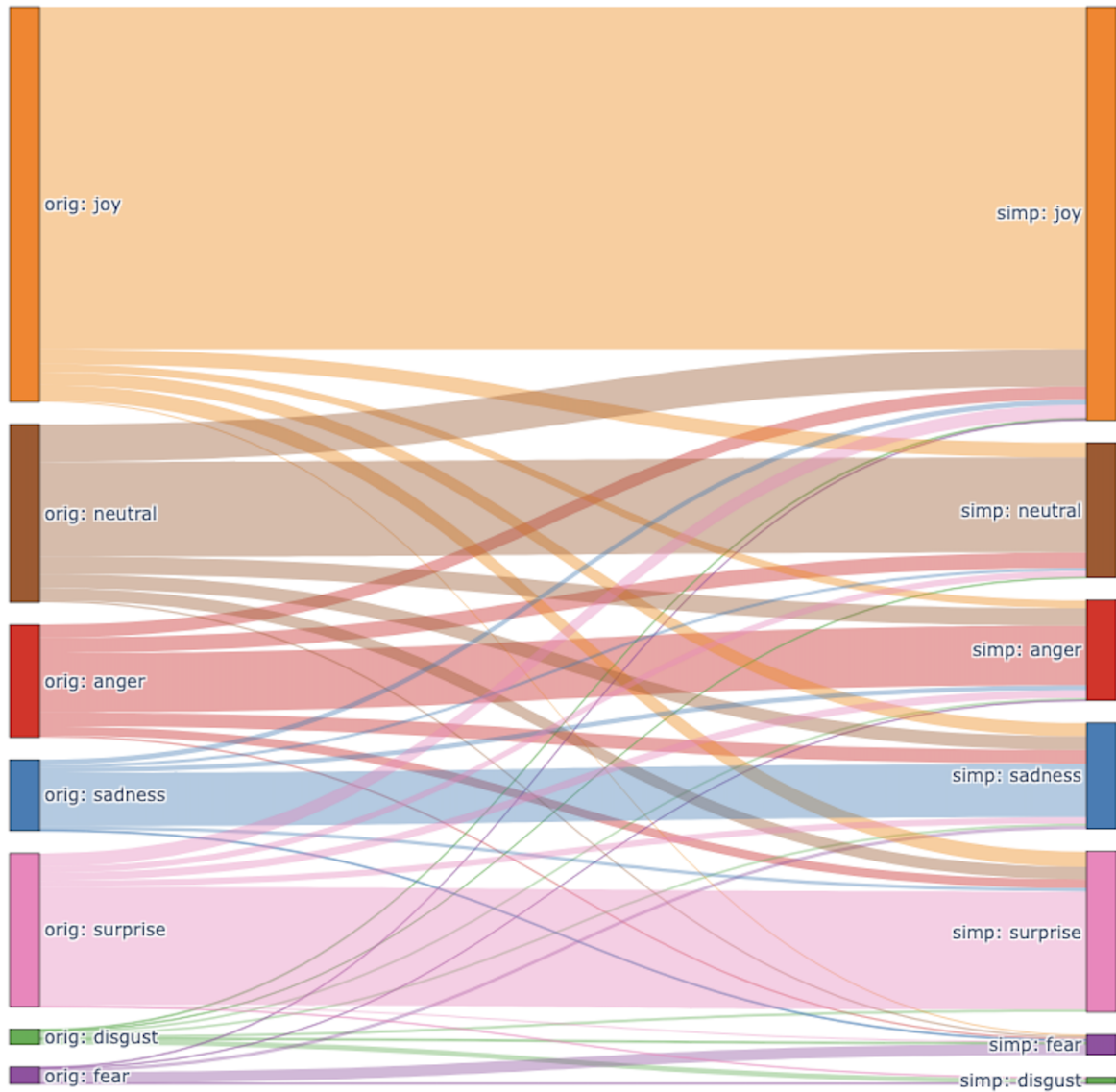


Figure A.3: Alluvial diagram showing flows from emotion labels predicted on the original posts (*pred_orig*, left) to labels predicted on the simplified posts (*pred_simp*, right).

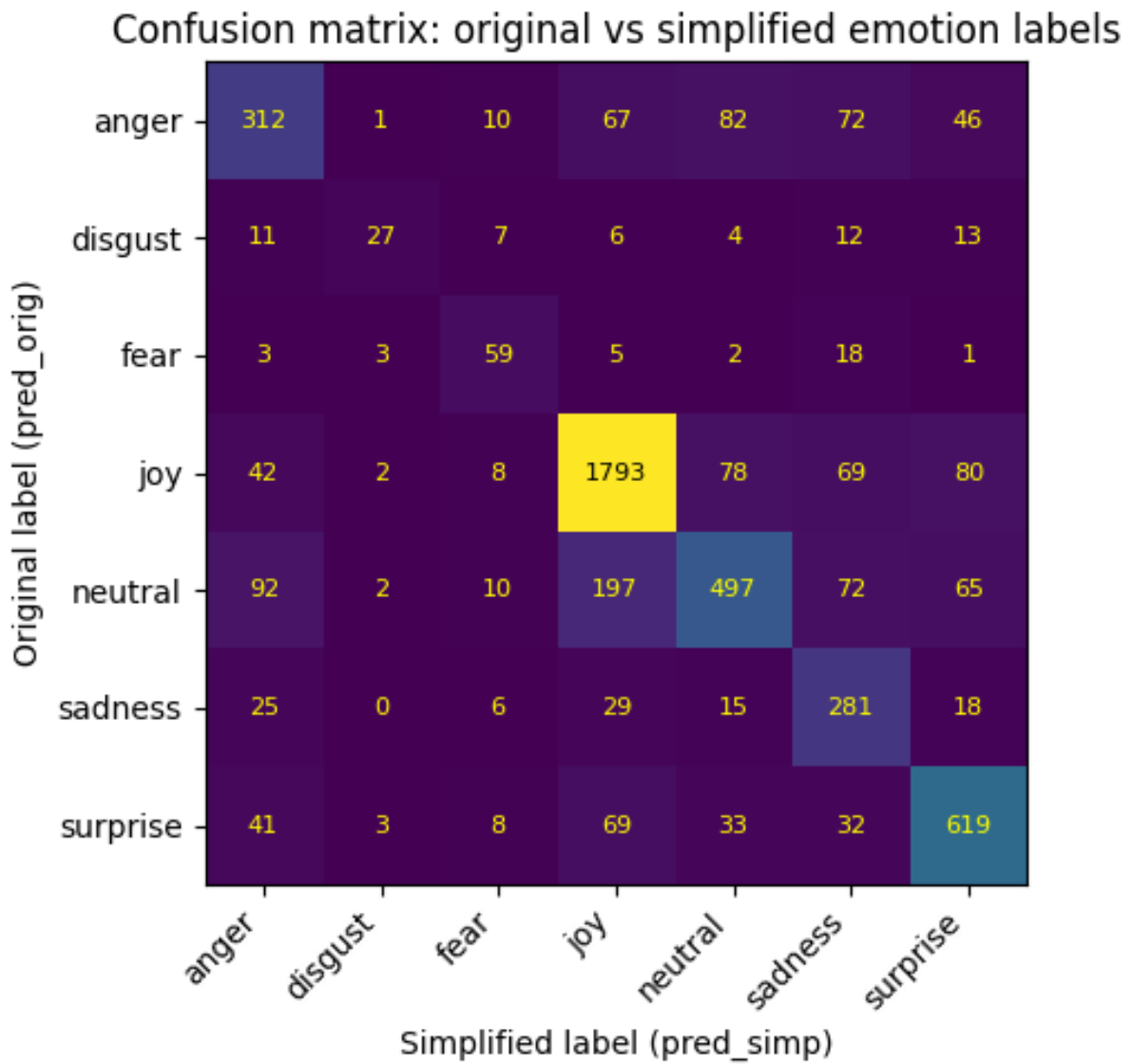


Figure A.4: Confusion matrix showing counts of emotion labels predicted on the original posts (*pred_orig*) versus the simplified posts (*pred_simp*).

Simplify the post so that people with learning disabilities can easily understand it. Keep the same meaning and facts. Preserve the same emotion. Do not soften or exaggerate the emotion. Do not add new facts or advice. Do not judge the person. Use common words and active voice. Keep emojis only if they add meaning. Use CamelCase for hashtags. For example, instead of #learningdisabilityweek, write #LearningDisabilityWeek. **Do not add feelings or any assumptions about how a person feels.** Keep simple posts as they are, even though they contain swear words. Explain all abbreviations, famous people, events or any other entities. Do not add hashtags or emojis

Simplify the post so that people with learning disabilities can easily understand it. Keep the same meaning and facts. Preserve the same emotion. Do not soften or exaggerate the emotion. Make the feelings clear and simple. Do not add new facts or advice. Do not judge the person. Use common words and active voice. Keep emojis only if they add meaning. Use CamelCase for hashtags. For example, instead of #learningdisabilityweek, write #LearningDisabilityWeek. **Do not add "I feel"**. Keep simple posts as they are, even though they contain swear words. Explain all abbreviations, famous people, events or any other entities. Do not add hashtags or emojis.

Figure A.5: Prompt variant removing the instruction to "also name the feeling in words"

Simplify the post so that people with learning disabilities can easily understand it. Keep the same meaning and facts. Preserve the same emotion. Do not soften or exaggerate the emotion. Make the feelings clear and simple. Do not add new facts or advice. Do not judge the person. Use common words and active voice. Keep emojis only if they add meaning, and also name the feeling in words. Do not assume how the person feels.

Simplify the post so that people with learning disabilities can easily understand it. Keep the same meaning and facts. Preserve the same emotion. Do not soften or exaggerate the emotion. Make the feelings clear and simple. Do not add new facts or advice. Do not judge the person. Use common words and active voice. Keep emojis only if they add meaning. Do not assume how the person feels.

Figure A.6: Prompt variant removing the instruction about hashtag or emoji.

You are an accessibility editor for social media.

GOAL: Make the posts easy to read without changing the original emotion or facts.

CONSTRAINTS: (1) Short sentences; one idea per sentence. (2) Keep names, numbers, links; keep hashtags (CamelCase). (3) No advice, opinions, or extra facts. (4) Do not judge the person. (5) Keep emojis only if they were in the original text and keep emojis at the end of the message (6) Explain all abbreviations, famous people, events or any other entities.

EMOTION: Keep the same emotion and intensity. Make the feelings clear and simple. Do not add feelings or any assumptions about how a person feels.

OUTPUT: Only the simplified text.

Figure A.7: Different prompt with additional instructions

Crowd-Based Evaluation of Emotion Intensity Preservation in Spanish–Basque Tweet Machine Translation

Nora Aranberri

HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
nora.aranberri@ehu.eus

Abstract

Machine translation (MT) systems perform well on standard benchmarks, yet their ability to preserve emotional meaning in informal user-generated content—particularly for low-resource languages—remains underexplored. We investigate the preservation of emotion intensity in Spanish–Basque tweet translation, focusing on Basque, an under-represented language in MT research. We compile a small, controlled corpus of Spanish reaction tweets and evaluate Basque translations from three publicly available systems through a crowd-based study. While all systems achieve comparable and above mid-range accuracy and fluency, emotion intensity is systematically attenuated in the translations, with greater loss for more emotionally intense inputs. A follow-up on highly emotional tweets shows that LLM prompting reduces emotion loss, yet substantial attenuation remains, highlighting emotion preservation as a persistent challenge in Spanish–Basque MT.

1 Introduction

Machine translation (MT) quality has improved substantially in recent years, with neural approaches—particularly transformer-based architectures and large multilingual models—now dominating research and practice (Ataman et al., 2025; Chatterji et al., 2025). According to the WMT 2025 findings, current systems perform well on standard benchmarks, especially for descriptive, emotionally neutral, and well-structured texts. These gains, however, are uneven across languages and domains, motivating the development of more challenging evaluation settings with linguistically complex data (Kocmi et al., 2025). In this context, creative and noisy user-generated content (UGC), such as tweets, remains particularly challenging for MT (Barbieri et al., 2020; Popovic et al., 2024), and exposes limitations in handling linguistic variability, metaphorical language, and contextual cues. These

challenges are amplified in low-resource settings, where limited parallel data and cultural specificity constrain both training and evaluation (Joshi et al., 2020). Basque, despite recent corpus development efforts (Etchevoyhen et al., 2018; Barnes et al., 2018; Romero et al., 2022; Heredia et al., 2025), remains under-represented in MT research on UGC, particularly in informal, affect-rich domains.

Prior work shows that the preservation of emotional intensity is particularly important when translating between languages with unequal resources or sociolinguistic status, yet MT systems often fail to retain fine-grained affective cues (Briakou and Carpuat, 2021). Still, many online platforms (e.g., X) automatically translate UGC without human supervision. While this can increase accessibility for minority languages like Basque, it also risks distorting emotionally salient content that shapes cross-linguistic interpretation. Given the pervasive and emotionally charged nature of social media, it is essential to examine whether—and to what extent—emotional content is preserved in MT outputs, particularly for these languages. In this paper, we investigate Spanish–Basque tweet translation, analyzing how different MT systems handle accuracy, fluency, and emotion intensity preservation. We construct a small tweet corpus and conduct a crowd-based evaluation with Basque-speaking intensity increases, emotion intensity loss in translation becomes progressively more pronounced.

2 Related Work

Tweets differ markedly from formal written genres in their linguistic, structural, and pragmatic properties. Twitter has been described as a highly conversational medium shaped by platform-specific conventions such as mentions, retweets, and hashtags (Honeycutt and Herring, 2009). Tweets are typically brief and informal, exhibiting non-standard orthography, creative spelling, abbreviations, ex-

pressive punctuation, and multimodal markers such as emojis and hashtags (Eisenstein, 2013; Zappavigna, 2012). Prior work highlights the frequent use of stance-taking, irony, and humor, particularly in emotionally charged contexts (Barbieri and Saggion, 2014). Twitter discourse also relies heavily on shared context, ongoing events, and cultural references (Feldman et al., 2018), while lexical variation reflects social identity factors such as gender and community affiliation (Bamman et al., 2014). These characteristics make tweets a challenging genre for MT and emotion-related tasks.

Tweets and microblogs have been widely studied in recent work on emotion and MT, with analyses focusing on how emotion intensity is altered during translation. Fukuda and Jin (2022) adopt a quantitative approach to emotion preservation by measuring how emotion scores assigned to tweets change before and after translation. Using a Japanese–English Twitter dataset, they apply an automatic emotion analysis system, enabling direct comparison of emotion intensity across multiple categories (joy, relief, fear, anger...). They show that MT systematically alters emotional profiles, with translated tweets exhibiting reduced variance and attenuated intensity, particularly for high-arousal emotions. Crucially, these effects are observed even when translations are otherwise fluent and semantically adequate, indicating that emotion distortion is not captured by standard MT evaluation metrics.

Extending this analysis to a multilingual setting, Saadany et al. (2023) use emotion-annotated Twitter data in English, Arabic, and Spanish, and analyze translations by commercial MT services, including Google Translate, Microsoft Bing Translator, and Amazon Translate. Their study combines automatic detection of emotion mismatches with targeted manual error analysis, and allows them to identify Twitter-specific linguistic phenomena—such as slang, emojis, code-switching, and non-standard spelling—that frequently trigger emotion distortion even when translations remain fluent and semantically plausible.

Qian et al. (2023) continue in this line for Chinese–English translation of emotion-loaded microblog texts. Using translations generated by Google Translate, they conduct fine-grained manual evaluation based on an MQM-inspired scheme that explicitly targets emotion preservation and distortion. Again, their analysis shows that a substantial proportion of MT outputs fail to preserve the

original emotional meaning, and their error analysis highlights recurring triggers such as polysemy, negation, and abbreviations.

To the best of our knowledge, no prior work has examined the preservation of emotional intensity of translated social media content into Basque, although recent studies have begun to characterize informal language in social media (ILSC) for this language. The challenges typically associated with ILSC—non-standard orthography, emojis, abbreviations, cultural references, and emotionally charged expressions—are amplified by Basque’s status as a minority language undergoing revitalization, in which informal digital registers are still emerging. From a sociolinguistic perspective, Elordui (2025) identifies three defining features of Basque ILSC—transcription of spoken language, dialectal variation, and code-switching—based on the GazteSare corpus (Elordui et al., 2020). However, as GazteSare draws on Facebook and Instagram data, its findings may not fully capture ILSC on platforms such as X, where communication is more public, reactive, and performative, often intensifying expressive and affective language.

These findings motivate a focused investigation of emotion preservation in the machine translation of tweets into Basque.

3 Experimental Setup

3.1 Spanish tweets

Our first step was to obtain a corpus of informal Spanish tweets. Although a parallel Spanish–Basque dataset would be ideal, no such resource exists to our knowledge. While several tweet collections have been developed for Spanish and Basque (Alegria et al., 2013; Pérez et al., 2019; Amores et al., 2021; Camargo Fernández, 2021), these are typically domain-specific, thematically limited, or dated. Also, they do not all involve informal language and tend to rely on translated data. Given the rapidly evolving nature of online discourse, we constructed a new dataset tailored to our study.

Because direct collection of real tweets is constrained by platform policies and legal considerations, we designed a controlled elicitation task to simulate authentic tweet production. To capture informal, creative, and emotionally expressive language, we recruited 19 native Spanish-speaking volunteers aged 19–25, a demographic chosen for their frequent use of platforms such as X.

Volunteers were asked to produce reaction

tweets within a form-based interface. To approximate a realistic user experience, we embedded the stimuli in a X-like layout: each stimulus displayed a headline and subheading tweet adapted from either a mainstream Spanish newspaper (El País) or a satirical outlet (El Mundo Today), accompanied by an illustrative image, metadata such as likes and retweets, and surrounding interface elements (e.g., recommended accounts, trending topics). The topics were chosen to encourage spontaneous, playful, or cheeky reactions; our interest lay primarily in the linguistic form, tone, and stylistic features of the responses. For each stimulus, volunteers were provided with an empty text box in which to write their reaction tweets.

Before beginning the task, we conducted a brief group discussion in which participants reflected on what they perceived as characteristic features of tweets (e.g., brevity, irony, informality). This served to situate them in an everyday social media mindset and mitigate the risk of producing unnaturally formal or academically influenced language. In total, we collected 472 reaction tweets, with each volunteer contributing an average of approximately 24–25 instances.

3.2 Baseline Basque translations

To assess how different systems handle the automated translation of informal tweet language, we compared the outputs of three publicly accessible MT and LLM-based tools. The first system was Itzuli, a neural MT service developed and maintained by the Basque Government¹. The second was ChatGPT 4-o-mini, a proprietary large language model accessible via the ChatGPT platform². Finally, we included Latxa, a Llama-based open model specifically adapted for Basque and recently released by the HiTZ Center (Sainz et al., 2025)³. Together, these systems represent a generic NMT engine, a generic commercial multilingual LLM, and an open-source Basque-specialized LLM, allowing us to examine a diverse range of approaches to Spanish–Basque translation.

All translations were obtained using each system’s default configuration. For the LLMs, we used a simple translation prompt instructing the model to “translate the following tweet,” without providing additional information or constraints. This design choice allows us to compare the baseline perfor-

¹<https://www.euskadi.eus/itzuli/>

²<https://chatgpt.com/>

³<https://latxa.hitz.eus/>

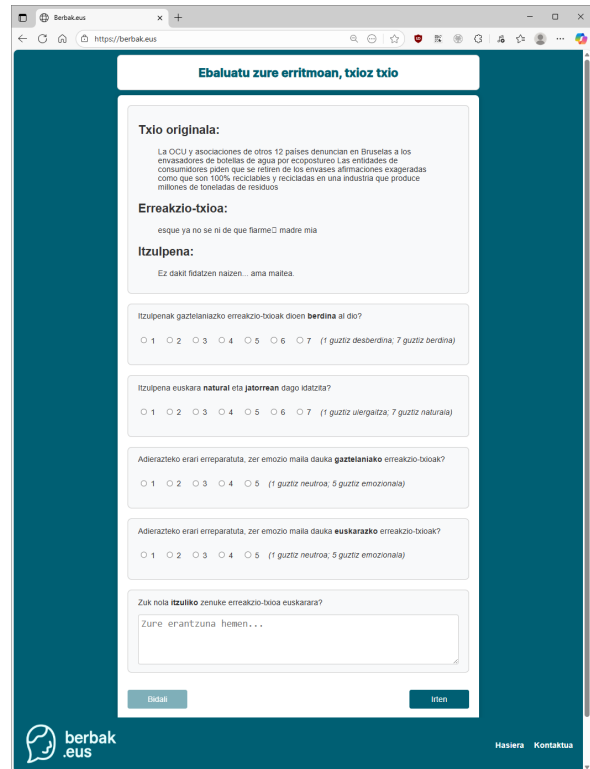


Figure 1: Sample screenshot of the evaluation interface.

mance of each model in a minimal-supervision setting and focus on the inherent capabilities of each system in preserving the linguistic and emotional intensity of informal social-media discourse.

3.3 Evaluation attributes and platform

Although our primary focus is the preservation of emotional intensity, we begin with the two dimensions that are typically central in translation quality evaluation: whether the target text preserves the meaning of the source (accuracy) and whether it reads naturally (fluency). However, tweets place particular emphasis on emotional content, given their informality, expressiveness, non-standard spelling, and dense affective cues. Accordingly, we include emotion intensity as a separate, explicit evaluation attribute, assessing whether the translation preserves the emotional intensity of the original text.

This issue is especially relevant in the case of Basque. First, most training data—whether for NMT systems or LLMs—primarily covers standard or formal registers, which means that models have limited exposure to non-standard, colloquial uses of the language. Second, although Basque has undergone substantial normalization, its informal varieties continue to evolve unevenly across regions

and speaker communities. With 5 dialects, about 15 subdialects and over 50 recognized varieties, as well as widespread code-switching with Spanish/French and English, what counts as “natural” Basque in informal settings can differ significantly across speakers. As a consequence, the perception and expression of emotion in Basque is highly variable, making it an especially challenging language for evaluating the adequacy of informal MT. To address this gap, we also gather reference tweets by participants for further research.

We designed a dedicated web page to collect fully anonymous evaluations from the Basque-speaking community (Figure 1). Participants were asked to indicate their linguistic background (native/C1 level, specialization, and social media usage), but no verification was required. Previous crowd-based evaluation campaigns (Aranberri et al., 2017; Aranberri, 2024; Sainz et al., 2025) have shown that contributions from non-Basque speakers are extremely rare, probably because the initiative is not relevant to this profile and because the dissemination channels typically reach Basque-speaking networks. Lowering participation barriers leads to higher engagement, and given the low risk of contamination, we opted not to perform additional checks. For anomalous cases, reference translations could be manually reviewed and responses discarded.

Once participants completed the introductory section and profile questions, they proceeded to the evaluation interface. For each item, they were shown the initial Spanish tweet, its Spanish reaction tweet, and the automatically translated Basque version of the latter. They rated accuracy (“does the translation express the same meaning as the Spanish reaction tweet?”) and fluency (“does the translation read naturally in Basque?”) using a 7-point Likert scale. They also rated the emotional intensity of both the Spanish reaction tweet and its Basque translation on a 5-point Likert scale. Although absolute intensity values were collected, note that our main interest lies in the comparison between the source and target languages. Finally, participants were offered an open text box to provide their own Basque translation under the prompt: “How would you translate this reaction tweet into Basque?”

Given the size of the dataset, we set to collect a single evaluation per translation instance. While this limits our generalizability, it provides a valuable first step toward understanding how infor-

mal, emotion-rich tweet language is translated into Basque and lays the groundwork for more extensive evaluation in future work.

4 Crowd assessment results for baseline translations

We collected 1,098 individual evaluation responses. Because the same reaction tweet was not presented to evaluators twice and the task was open for a limited amount of time, the resulting dataset exhibited certain imbalances: some translations did not receive any evaluation, while others received several, independent assessments. To improve the comparability of the results, we cleaned the data.

First, we removed all reaction tweets for which any of the three MT systems lacked evaluation. Next, we resolved duplicate evaluations, defined as cases in which the same reaction tweet–system pair received more than one assessment. To do so, we distinguished between specialist and non-specialist evaluators. We classified evaluators as specialists if they were native or C1-level Basque speakers and met at least one of the following criteria: training in translation studies or philology, professional experience as translators, or work in communication-related fields. All remaining evaluators—including those who did not report their language proficiency or lacked a language-related background—were classified as non-specialists.

For duplicate cases evaluated exclusively by non-specialists, we merged the assessments by averaging their scores. Similarly, for duplicates evaluated only by specialists, we averaged the specialist scores. Finally, for cases with mixed evaluator profiles, we retained only the specialist assessments; when multiple specialist evaluations were available, these were again merged by averaging the scores.

After completing this cleaning process, our dataset consisted of 1,002 evaluations for 334 reaction tweets, with exactly one evaluation per MT system for each reaction tweet.

The evaluations were predominantly provided by self-reported proficient participants: 75% of the responses came from native Basque speakers and 91% from evaluators who reported a C1 level or higher. Only 28 responses (2.79%) were submitted by non-native evaluators who did not report a C1 level. After manually reviewing the reference translations produced by the latter group, we chose to retain all responses, as their contributions were of satisfactory quality.

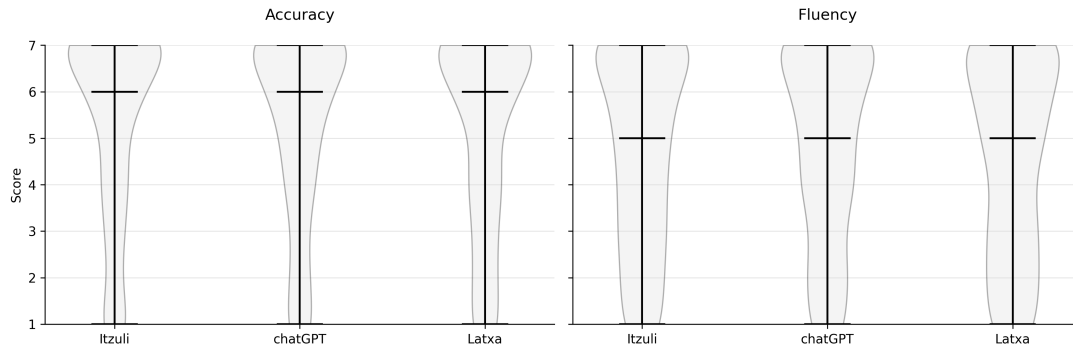


Figure 2: Distribution of accuracy and fluency scores per system. Horizontal lines indicate median values.

4.1 Accuracy and Fluency

Focusing on general translation quality, we analyze two attributes: accuracy and fluency (Figure 2). In a 1-7 Likert scale, accuracy scores fall in the range of approximately 5.1-5.3, while fluency scores are slightly lower, around 4.6-4.8. On average, this reflects a difference of roughly 0.5 scale points between accuracy and fluency, with accuracy consistently rated higher. Both dimensions are about 0.5-1.7 points above the scale midpoint of 4, indicating that the systems clearly surpass mid-range performance. Yet, at the same time, the gap of nearly two points between current performance and the upper end of the scale suggests substantial remaining room for improvement.

Accuracy scores are similar across systems, with Itzuli ($M = 5.28$, $SD = 2.03$) and chatGPT ($M = 5.21$, $SD = 1.97$) performing nearly identically and Latxa slightly lower ($M = 5.10$, $SD = 2.05$). A comparable pattern is observed for fluency, where chatGPT ($M = 4.78$, $SD = 2.02$) and Itzuli ($M = 4.70$, $SD = 2.05$) again score similarly, followed by Latxa ($M = 4.58$, $SD = 2.08$). All three systems cover the full 1-7 scale, with distributions showing similar variability and reflecting variation in translation performance.

To test whether these differences are statistically meaningful, we applied complementary analyses. A one-way ANOVA found no significant effect of system for either accuracy ($F = 0.67$, $p = .512$) or fluency ($F = 0.83$, $p = .436$). The Kruskal–Wallis test, which does not assume interval-scale Likert data confirmed this result (accuracy: $H = 1.69$, $p = .429$; fluency: $H = 1.58$, $p = .453$).

4.2 Preservation of emotion intensity

We next turn to emotion preservation. Spanish intensity scores span the entire scale (1–5), with

a mean of 3.61 ($SD = 1.14$). The distribution is skewed toward mid-to-high intensity: the median is 4, and the interquartile range spans from 3 to 4. This indicates that most reaction tweets express noticeable, though not extreme, emotion intensity (Figure 3). Although assigning emotion intensity to a tweet is inherently subjective, evaluators rated the Spanish reaction tweets set consistently across systems (Itzuli: 3.63, ChatGPT: 3.61, Latxa: 3.60), making general comparisons possible (Itzuli: 3.63, ChatGPT: 3.61, Latxa: 3.60).

When examining how systems preserved this emotional intensity in their Basque outputs, we observe a clear downward shift (Figure 3). While Basque scores also cover the full available range, the mean drops to 2.80 ($SD = 1.29$), nearly one full point lower than the Spanish originals. This indicates that translations tend to dampen emotion intensity, producing outputs that are more neutral or less expressive than their Spanish counterparts. The median Basque intensity score is 3, reflecting a general attenuation rather than an abrupt loss of affect.

This trend is consistent across systems, with similar mean scores for Itzuli (2.83), ChatGPT (2.76), and Latxa (2.82). The consistency of this reduction across models suggests that emotion dampening is not system-specific, but rather reflects a broader challenge in MT.

Aware of the subjectivity involved in determining emotional intensity, our primary interest lies not in absolute scores but in how intensity shifts during translation. We therefore analyse the *intensity shift*, defined as the difference between the Basque and Spanish emotion intensity ratings for each translation. Negative values indicate emotional attenuation, while positive values reflect amplification.

Across all systems, the mean intensity shift is consistently negative. To better understand this ef-

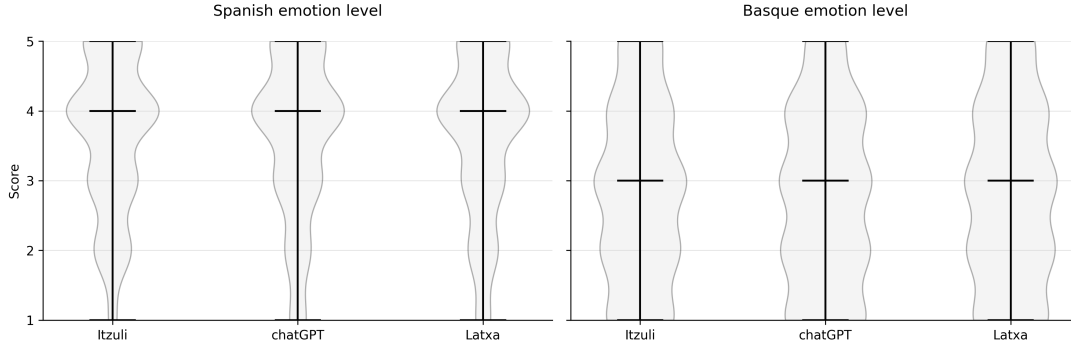


Figure 3: Distribution of emotion intensity scores for Spanish source tweets and Basque translations. Horizontal lines indicate median values.

fect, we examine how shifts vary across the Spanish scale and whether systems differ in their handling of low-, medium-, and high-intensity inputs.

Across the three systems, emotion intensity preservation behaviour is remarkably consistent at each Spanish emotion level, with inter-system differences generally below 0.5 points (Figure 4). For tweets with low intensity (levels 1 and 2), all systems preserve emotion closely, with shifts near zero. As source intensity increases, however, attenuation becomes more pronounced: tweets rated at level 3 lose approximately half a point on average, while high-intensity tweets (levels 4 and 5) lose between one and 1.3 points. This trend indicates that input with higher emotional intensity leads to greater loss in translation, regardless of the system used.

To assess whether differences between systems are significant, we compared the intensity-shift scores using both parametric and non-parametric tests. A one-way ANOVA found no significant effect of system ($F = 0.42$, $p = .659$). A Kruskal–Wallis test similarly showed no significant differences ($H = 1.67$, $p = .434$). Together, these results indicate that the system used to obtain the translation does not influence the degree of intensity shift.

4.3 Emotion intensity shift and translation quality

Finally, we checked whether the degree of emotion intensity shift in translation correlates with the global translation quality attributes studied, namely, accuracy and fluency. Because both metrics are ordinal, we first computed Spearman rank correlations. Intensity shift exhibits moderately strong positive correlations with both accuracy ($\rho = 0.514$, $p < 0.001$) and fluency ($\rho = 0.481$,

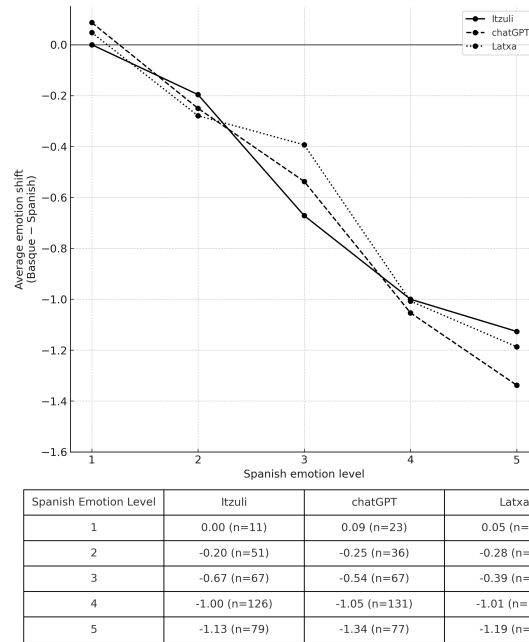


Figure 4: Degree of intensity shift (Basque–Spanish) across Spanish levels.

$p < 0.001$), which suggests that translations that preserve more emotional intensity also tend to receive higher quality ratings. However, the association is moderate rather than deterministic: attenuation is observed even in translations judged as reasonably accurate and fluent. This pattern, robust across systems, differs from previous findings (Fukuda and Jin, 2022; Saadany et al., 2023), which emphasize a stronger independence between emotion preservation and general translation quality. One possible explanation is that differences in overall system quality may partly account for this relationship, motivating further investigation into the interplay between emotion intensity preservation and global quality metrics.

5 High emotion intensity tweet translation

Given that translations of more emotionally intense tweets appear to be particularly challenging, and noting that our initial experiments did not fully exploit the capabilities of LLMs, we conducted a targeted follow-up evaluation.

We selected reaction tweets with high emotion intensity (scores of 4 or 5) whose Basque translations had previously received low intensity ratings ($= < 3$). This procedure resulted in a subset of 86 reaction tweets. These tweets were retranslated into Basque using Latxa under three different prompting conditions:

Prompt 1: A prompt written in Basque that assigns the LLM the role of a professional translator, explicitly describes the communicative context and stylistic characteristics of tweets, and requests high-quality translations with respect to accuracy, fluency, and style.

Prompt 2: An extension of the previous prompt with five real, in-context examples of Basque tweets, corresponding to a few-shot setting.

Prompt 3: A variation of the second prompt with the temperature increased from 0.7 to 1.2 (Top-P at 0.9), allowing greater generative flexibility.

We then conducted a crowd-based evaluation. Using the same platform as before, volunteers were again asked to assess the accuracy and fluency of the Basque translations, as well as the emotion intensity of both the Spanish source tweets and their Basque counterparts.

We collected a total of 499 evaluation responses covering the 86 reaction tweets (258 translations), with at least one assessment per prompt for each reaction tweet. Regarding participant proficiency, 84.37% of responses were provided by native Basque speakers, and 94.56% by evaluators reporting a C1 level or higher. Only 11 responses (2.20%) came from non-native evaluators without a reported C1 level. As before, duplicate evaluations were resolved by applying the procedure for distinguishing specialist and non-specialist evaluators and merging multiple assessments accordingly.

We begin by examining global quality metrics. As expected, all three prompting strategies outperform the baseline (from now on Prompt 0) on this high-intensity subset, where scores fail to reach the scale midpoint (accuracy: $M = 3.94$, $SD = 1.95$; flu-

ency: $M = 3.47$, $SD = 1.98$). Mean ratings indicate that all new prompting strategies achieve above mid-range quality for both attributes; however, the remaining gap of more than two points to the top of the scale suggests considerable room for improvement (Figure 5). The few-shot prompt (Prompt 2) yields the highest mean scores for both accuracy ($M = 4.61$, $SD = 1.98$) and fluency ($M = 4.84$, $SD = 1.93$), followed closely by the description-rich prompt (Prompt 1; accuracy: $M = 4.56$, $SD = 1.98$; fluency: $M = 4.68$, $SD = 1.96$). Increasing the decoding temperature (Prompt 3) results in slightly lower scores (accuracy: $M = 4.35$, $SD = 2.02$; fluency: $M = 4.55$, $SD = 2.00$).

No significant differences across systems are observed for accuracy (ANOVA: $F = 2.08$, $p = .102$; Kruskal-Wallis: $H = 6.40$, $p = .094$). In contrast, fluency differs significantly across systems (ANOVA: $F = 8.52$, $p < .001$; Kruskal-Wallis: $H = 23.25$, $p < .001$). Tukey’s HSD confirms that the baseline prompt (prompt0) scores significantly lower than all enhanced prompting conditions (prompt1–prompt3; all $p < .01$), while no significant differences emerge among the enhanced prompts (all $p > .75$).

We next focus on emotion intensity preservation. Mean Spanish intensity scores range from approximately 3.8 to 4.1 in a 1-5 scale, clustering around the high-section of the scale. This shows slightly more variation in assigning an emotional intensity to a rich-emotion tweet. Basque scores are substantially lower across prompts.

Across all prompting conditions, emotion shift values are consistently negative, indicating a systematic attenuation of emotional intensity in Basque relative to Spanish. The baseline prompt (Prompt 0) shows the largest loss ($M = -1.85$), while all three enhanced prompting strategies substantially reduce the magnitude of this shift. The few-shot prompt (Prompt 2) achieves the smallest average loss ($M = -1.11$), corresponding to an improvement of approximately 0.6 points in emotional preservation. Nevertheless, even under the best-performing prompt, emotion intensity is reduced by more than one scale point on average, suggesting that effective emotion preservation may require more complex prompting or fine-tuning approaches that add knowledge or better access particular language characteristics.

To test whether intensity-shift differences are statistically meaningful, we applied complementary significance tests. A one-way ANOVA showed a

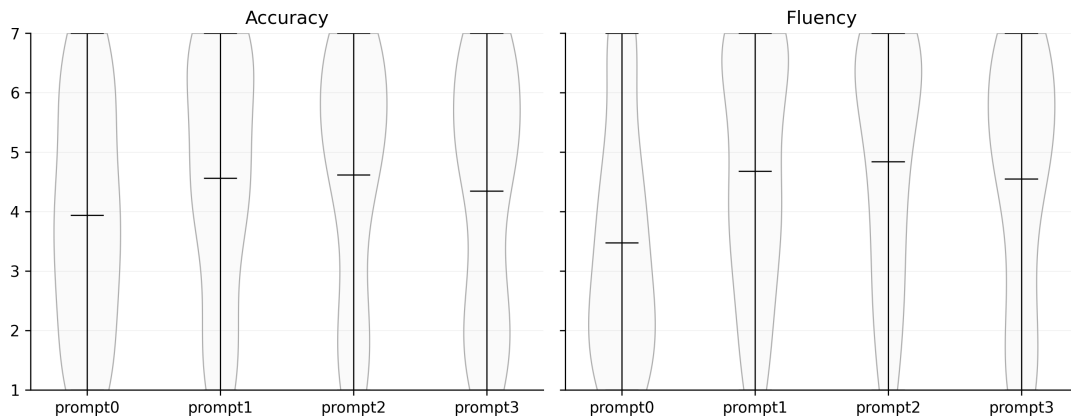


Figure 5: Accuracy and Fluency score distributions per prompt. Horizontal lines indicate median values.

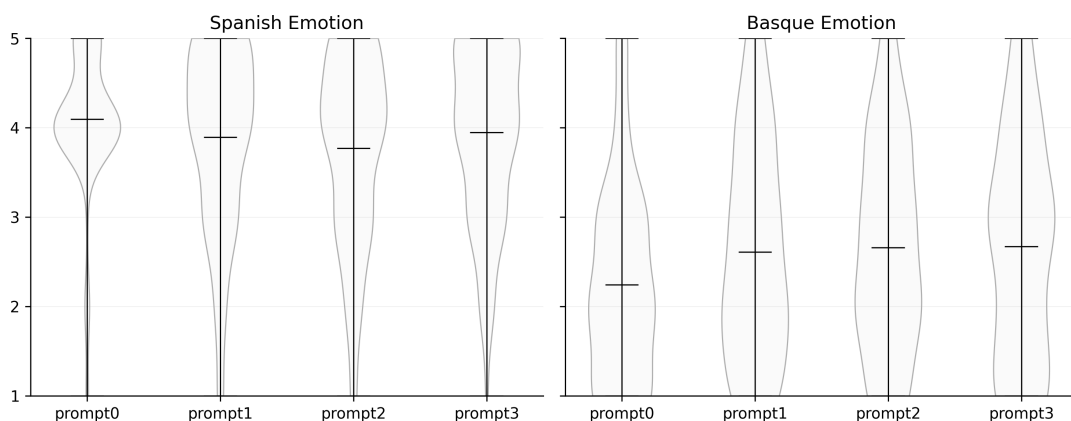


Figure 6: Spanish and Basque emotion intensity score distributions per prompt. Horizontal lines indicate median values.

significant effect of prompting strategy ($F = 4.69$, $p = .001$), which was confirmed by a Kruskal–Wallis test ($H = 18.45$, $p = .001$). Post-hoc Tukey HSD comparisons revealed that the baseline prompt (Prompt 0) differs significantly from all enhanced prompts (Prompt 1–Prompt 3; all $p < .05$), while no significant differences were found among the enhanced prompts themselves (all $p > .80$).

6 Conclusions

We studied the MT of emotion-rich tweets from Spanish into Basque, focusing on translation quality and preservation of emotion intensity. Overall, accuracy and fluency scores lie above the scale midpoint, indicating reasonably good performance, though with substantial room for improvement; for high intensity tweets, however, overall quality falls below mid-range levels, highlighting the difficulty of this subset. We observe no significant quality differences between a traditional NMT system and LLM-based models in baseline settings without

task-specific instructions. Across systems, emotion intensity is systematically attenuated in translation, with an average loss of about one point on a 1–5 scale that increases with source emotion intensity. Contrary to previous work, quality attributes seem to correlate with intensity shift, possibly reflecting differences in overall system quality. Finally, basic LLM prompting strategies—contextual instructions, few-shot examples, and increased temperature—yield only modest reductions in intensity loss for highly emotional tweets, leaving intensity preservation an open challenge in Spanish–Basque MT.

Several limitations should be acknowledged. The dataset is small, limiting generalizability, and the evaluation relies on crowd-based judgments from participants with mixed language backgrounds. Although prior work shows that evaluators with linguistic or translation training tend to be stricter and expertise was considered during data cleaning, its effect was not systematically

tested. Moreover, most translations received only a single evaluation, and judgments of quality and emotion intensity are inherently subjective, adding variability. Future work would benefit from more controlled designs that balance emotion intensity and cover different emotion types, and from examining how the development and use of informal Basque—both in MT training data and in speaker expectations—relate to translation performance and evaluation. Despite these limitations, our results point to emotion intensity preservation as a meaningful challenge in Spanish–Basque tweet MT.

Acknowledgments

This work was partially supported by the MOLVI project (PID2024-157855OB-C32), funded by MICIU/AEI/10.13039/501100011033 and FEDER, EU, and by the project Desarrollo de Modelos ALIA, Resol. SEDIA 19.08.2024, within the framework of the National Language Technologies Plan (ENIA 2024), funded by MTDFP, PRTR, and the European Union–NextGenerationEU. The author thanks the volunteer participants for their contributions to the data creation and crowd-based evaluation.

References

- Inaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluís Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la tarea compartida tweet-norm 2013: Normalización léxica de tuits en español. In *Tweet-Norm@ SEPLN*, pages 1–9.
- Javier J Amores, David Blanco-Herrero, Patricia Sánchez-Holgado, and Maximiliano Frías-Vázquez. 2021. Detectando el odio ideológico en twitter. desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos. info*, (49):98–124.
- Nora Aranberri. 2024. [Analysis of the annotations from a crowd MT evaluation initiative: Case study for the Spanish-Basque pair](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 548–559, Sheffield, UK. European Association for Machine Translation (EAMT).
- Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilaraza, and Kepa Sarasola. 2017. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 51(4):1053–1084.
- Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. [Machine translation in the era of large language models: a survey of historical and emerging problems](#). *Information*, 16(9).
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.
- Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1644–1650.
- Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in twitter. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 56–64. Association for Computational Linguistics.
- Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. [MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Eleftheria Briakou and Marine Carpuat. 2021. A case for evaluating emotion in machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 2855–2869.
- Laura Camargo Fernández. 2021. El nuevo orden discursivo de la extrema derecha española: de la deshumanización a los bulos en un corpus de tuits de vox sobre la inmigración. *Culture, Language and Representation, 2021*, vol. 26, p. 63-82.
- Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use chatgpt](#). Working Paper 34255, National Bureau of Economic Research.
- Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369. Association for Computational Linguistics.
- Agurtzane Elordui. 2025. Heteroglossic management in instagram: Emerging ideological dynamics among basque youth. *Journal of Linguistic Anthropology*, 35(1):1–23.
- Agurtzane Elordui, Jokin Aiestaran, Garbiñe Bereziartua, Irantzu Epelde, Orreaga Ibarra, Oroitz Jauregi, Libe Mimenza, Beñat Muguruza, and Ane Odria. 2020. Gaztesare corpus and data-base.

- Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, and 1 others. 2018. Mt for under-resourced languages: The case of basque. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3761–3766.
- Oded Feldman, Dan Ariely, and Erez Shmueli. 2018. Characterizing the linguistic style of conversational tweets. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 100–109. AAAI.
- Karin Fukuda and Qun Jin. 2022. Analyzing change on emotion scores of tweets before and after machine translation. In *Social Computing and Social Media: Design, User Experience and Impact*, pages 294–308, Cham. Springer International Publishing.
- Maite Heredia, Jeremy Barnes, and Aitor Soroa. 2025. [EuskañoldS: A naturally sourced corpus for Basque-Spanish code-switching](#). In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–5, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Courtenay Honeycutt and Susan C. Herring. 2009. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6282–6293.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- C Pérez, A Rebollar, and M Pérez. 2019. Construcción de un corpus lingüístico a partir de tweets tomando como base un tema específico. *Jornada de Ciencia y Tecnología Aplicada Tecnológico Nacional de México/CENIDET*, 2:15–19.
- Maja Popovic, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. 2024. [Effects of different types of noise in user-generated reviews on human and machine translations including ChatGPT](#). In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 17–30, San
- Ġiljan, Malta. Association for Computational Linguistics.
- Shenbin Qian, Constantin Orasan, Felix Do Carmo, Qiuliang Li, and Diptesh Kanojia. 2023. [Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland. European Association for Machine Translation.
- Eneko Romero, Haritz González, and Arantza Díaz de Ilarraza. 2022. The basque parliament corpus: Construction and applications for low-resource mt. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, pages 1125–1133.
- Hadeel Saadany, Constantin Orasan, Rocio Caro Quintana, Felix Do Carmo, and Leonardo Zilio. 2023. [Analysing mistranslation of emotions in multilingual tweets by online MT tools](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 275–284, Tampere, Finland. European Association for Machine Translation.
- Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. [Instructing large language models for low-resource languages: A systematic study for Basque](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29124–29148, Suzhou, China. Association for Computational Linguistics.
- Michele Zappavigna. 2012. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Continuum, London.

A Prompt wording

- **Prompt 0:**

Itzuli hurrengo tuitak euskarara.

- **Prompt 1:**

Itzultzaile profesionala zara. Eskatu dizute sare sozial bateko erreakzio txioak itzultzeko gaztelaniatik euskarara. Itzulpenek gaztelaniako erreakzio txioen esanahia islatu behar dute eta hizkuntza jatorra erabili. Txioek hizkera informala erabili ohi dute, ironia eta sarkasmoa ere bai, intentsitate emozional handikoak ere izan ohi dira, eta erreferentzia kulturalak izaten dituzte. Horregatik, garrantzitsua da itzulpenek gaztelaniako txioen estiloa mantentzea.

Hemen duzu itzuli beharreko erreakzio txioen zerrenda. Erreakzio-txio bakoitzaren aurretik hasierako txioa ageri da testuinguru gisa erabiltzeko. Itzuli erreakzio txioa bakarrik. Itzulpenak bakarrik erakutsi.

- **Prompt 2:** (Promt 1 + examples)

Hona hemen erreakzio txioen adibideak:

- Buah zelako temazoa txo
- pelikula bateko gidoia izan liteke baina esukal herriko jaixetan danetarik gerta liteke, aupa gaztediiiiiii!!
- Pitxartxarreraz dakidan apurraren arabera, "Ospa hemendik, zomorroide, txitatzen ari gaituk eta!"
- Tibet? etxe ondoan bertan! ai eneeeeeee! ez joan hain urrun ...
- Ez da Simón edo Agirrezabala... baia hau be ez da atezain txarra gero!

A Position Paper on *Toxic Reasoning*: Grounding Categories of Toxic Language in Implications and Attitudes

Stefan F. Schouten, Iliia Markov, Piek Vossen

Vrije Universiteit Amsterdam

{s.f.schouten, i.markov, p.t.j.m.vossen}@vu.nl

Abstract

Automatic detection of toxic language has the potential to considerably improve engagement with online spaces. Previous work has characterized toxic language detection as a classification problem, often using fine-grained classes for increased explainability. In this position paper, we argue for a new way of operationalizing categories of toxic language. Our approach focuses on what is expressed or implied, and breaks down implications based on two traits: (i) the core content of what was expressed, and (ii) relevant stakeholders' attitudes towards that content. We argue for an approach, which we call *toxic reasoning*, where such distinctions are made explicit. We point out the benefits of such an approach, and develop a toxic reasoning *schema*, which can explain categories of toxic language from diverse sources. We demonstrate this by mapping the classes of existing toxic language datasets to the schema. Toxic reasoning promises to provide improved understanding of implicit toxicity while increasing explainability.

Warning: contains examples of toxic language.

1 Introduction

Language that communicates hateful, derogatory, or offensive ideas (hereafter, *toxic language*) is generally undesirable, and needs to be detected and addressed to improve online spaces. Some toxicity is expressed explicitly, making it easier to identify. More challenging is toxicity that is conveyed implicitly. What counts as an instance of implicit toxicity can be subjective, for example, because different people can read the same text as implying different things. Additionally, some implications intended by the author might only be obvious with the appropriate context. Cultural, situational and discourse context may all be necessary to properly identify the plausible toxic implications of a text.

Widespread use of social media has greatly increased the need for automating the detection of

toxic language. Such language can harm individuals and communities, perpetuate stereotypes, and incite violence. Detecting and limiting such language is crucial to the usability and inclusivity of online spaces. A critical aspect of toxic language detection is *explainability*. With machine learning systems' involvement in content moderation, understanding their decisions is of great importance. Explainability not only helps to improve trust and transparency but also ensures that these systems can be audited and improved. Without explanations, both users and moderators may find it challenging to understand why content is flagged, possibly leading to misjudgments and lack of accountability (Nguyen et al., 2023).

In this position paper, we analyze current popular approaches to explainable toxic language detection, identifying aspects that can be improved. To address these concerns, we propose *toxic reasoning*, which requires giving the *reasons* why text is toxic. We argue that categories of toxic language should be defined in terms of what is being communicated, and create a toxic reasoning *schema* to help operationalize this task. This schema includes two main traits: (i) the core content of what is implied, and (ii) the attitudes of relevant stakeholders towards that content. For example, a statement such as “*all jews should be evicted from white countries*”¹ can be broken down into (i) a proposition ‘all jews are evicted from white countries’, and (ii) the author’s preference (or desire) towards this proposition. We believe that, through the toxic reasoning schema, we can explain most toxic narratives by characterizing what is conveyed, while still allowing a mapping to existing categories of toxic language.

The remainder of the paper is organized as follows. In [Section 2](#), we give a brief overview of the toxic language detection task. This is followed by an analysis of the definitions used for categories of

¹Taken from Implicit Hate Corpus (ElSherief et al., 2021).

toxic language in the literature, in [Section 3](#). [Section 4](#) investigates the use of toxic spans, and identifies difficulties that arise when spans are used for implicit toxic language. Then, in [Section 5](#), we give a detailed description of toxic reasoning and our schema, explaining how they overcome the limitations identified in the previous two sections. Next, we explain how existing toxic language categories can be explained by our schema, by providing a mapping between the two ([Section 6](#)).

2 Toxic language detection

Much work has been done on the detection of *toxic language*. We use the term toxic language as a broad umbrella term that includes hate speech ([Jahan and Oussalah, 2023](#); [Fortuna and Nunes, 2018](#)), abusive language ([Alrashidi et al., 2022](#); [Vidgen and Derczynski, 2020](#)), and offensive language ([Pradhan et al., 2020](#)). State-of-the-art results on toxic language detection tasks are usually achieved by fine-tuning language models on annotated datasets. A common distinction differentiates between explicit and implicit toxic language ([Waseem et al., 2017](#)): “explicit abusive language is that which is unambiguous in its potential to be abusive, for example language that contains racial or homophobic slurs”, whereas “implicit abusive language is that which does not immediately imply or denote abuse”.

Much effort has gone into making the detection of toxic language *explainable*. Explainability has so far mostly come in two forms. First, the inclusion of fine-grained annotation and classification schemes, which aim to better specify the exact nature of the toxicity. Multiple datasets have been annotated according to such (multi-level) taxonomies of toxic language (e.g. [Vidgen et al., 2021](#); [EISherief et al., 2021](#); [Kirk et al., 2023](#)). And second, the inclusion of toxic spans (or rationales), which aim to highlight which part of the message is responsible for its toxicity ([Pavlopoulos et al., 2021](#); [Mathew et al., 2021](#)).

3 How to define toxic language?

Annotation and classification schemes, whether coarse-grained or fine-grained, include many kinds of definitions for their categories of toxic language. Definitions are important, because implicitly or explicitly they make their way to the annotation of data (e.g. in the annotator guidelines). Then, when machine learning is used, there is an implicit

expectation for the algorithm to recover whatever decision criteria were used by the annotators.

Previous work has already analyzed definitions used for toxic language ([Fortuna et al., 2020](#); [Khurana et al., 2022](#)). These studies explored what aspects are and should be involved in the definitions of toxic language categories. Our analysis builds on their insights.

3.1 Borrowing definitions

Some definitions compare or contrast to other descriptors of (toxic) language. For example, [Basile et al. \(2019\)](#) mention “... HS [Hate Speech] against immigrants may include: 1. insults, threats, denigrating or hateful expressions ...”, clarifying what is meant by hate speech to those who already have some understanding of the related terms.

Concern 1 (Dependency). Appealing to other descriptive terms benefits explainability only to a limited extent, because the category becomes dependent on them. And, the other terms may themselves have multiple valid definitions. This may also lead to increased subjectivity in annotations, as different annotators may understand those terms differently.

3.2 Defining factors

We observe that many definitions, given a potentially toxic message, tend to focus on one or more of the following factors:

- CAUSE – *What caused the author to write the message?* – Commonly, the author’s intent, with a malicious intent often being required for toxic messages. Other causes might include unconscious biases, lapses in judgment, or ignorance.
- EFFECT – *What are the effects of the message?* – These can include harmful effects on the reader or on groups that are targeted by the message.
- FORM – *What (kinds of) words are used in the message?* – The presence of profanity, like slurs.
- MEANING – *What is conveyed by the message?* – Anything plainly asserted, implied or otherwise insinuated by a message.

To understand the way these factors are used, we will review some example definitions.

One of the ways a message is considered hate speech under the definition given by [Waseem and Hovy \(2016\)](#), is if it ‘seeks to silence a minority’. In this definition, the author’s reason for writing the toxic message is the deciding factor. In this case, the author would have the goal of silencing

minorities, and that goal can be identified as what caused the author to be toxic (CAUSE).

The Perspective API² defines its ‘toxicity’ and ‘severe toxicity’ categories by requiring that a message is (very) “likely to make people leave a discussion”. Thus, it defines the category based on an effect (EFFECT) of the toxicity. The ‘profanity’ category, on the other hand, applies when toxicity is conveyed using swear or curse words (FORM). Finally, the ‘threat’ category requires that the toxic utterance “describes an intention to inflict pain, injury, or violence”. Thus, to qualify, the message must express something like: “I intend to harm person X”. As such, the ‘threat’ category specifies what the message must convey (MEANING).

3.3 Cause and Effect

Fortuna et al. (2020) had the following to say about CAUSE and EFFECT: “*‘abusive’ has been defined based on the speakers’ intention to harm, which cannot always be determined by just looking at the content. Furthermore, definitions also make assumptions on the effect of the messages on the reader, which, obviously, depends entirely on the personality of the reader.*” We agree there are reasons to be skeptical about definitions that are based on CAUSE and EFFECT. Specifically, we present a number of concerns focusing on why CAUSE and EFFECT are harder to operationalize.

If a category only applies when it has (or is likely to have) specific effects, the classifier needs to infer such effects, which requires world modeling. To decide if a message is “likely to make people leave a discussion”, we need to have a world model that includes a ‘reader’. And, this world model should be expressive enough to predict what effect reading the message will have on them. Deciding if the author of a message ‘seeks to silence a minority’ is even more challenging. For this, we need a world model that includes the author’s model of the world. The classifier needs some kind of theory of mind (or at least it needs to behave like it has one).

Concern 2 (Feasibility). World modeling and theory of mind are both advanced capabilities, and it is still a matter of debate whether even the most capable Large Language Models (LLMs) possess these capabilities to a significant degree (Ding et al., 2025; Yildirim and Paul, 2024). Thus, it is unclear

²The Perspective API hosts machine-learned models to detect toxic language. It was created by Jigsaw and Google’s Counter Abuse Technology team. See perspectiveapi.com.

whether expecting classifiers to infer CAUSE and EFFECT is feasible.

Although, perhaps what is really meant are *perceived* intents (or causes), and *representative* effects. Maybe models are only expected to remember statistical patterns of which types of messages are associated with which causes or effects (and corresponding toxic categories), rather than truly reasoning about these matters.

Concern 3 (Failure to explain). Assume we have a language model fine-tuned to detect categories which are based on (perceived) CAUSE and (representative) EFFECT. The goal of introducing the fine-grained categories was to improve explainability. But, this model still fails to be truly explainable, since we still cannot account for why the classifier inferred the prerequisite CAUSE or EFFECT. What has been inferred and why? What statistical pattern is being exploited?

Finally, one more reason for concern arises when focusing on causes or effects.

Concern 4 (Increased Subjectivity). Generally, the only way for a message to have an effect on a reader, is through either the form or (perceived) meaning. Because of that, inferring the intent of the author often requires first inferring the (possible) meaning(s) of the message. Therefore, any subjectivity in the inference of cause and effect is *in addition* to subjectivity inherent in the inference of meaning.

4 Which spans are toxic?

Toxic spans detection (Pavlopoulos et al., 2021) or rationale extraction (Mathew et al., 2021) is a task where a model is trained to identify which parts of a message are responsible for its toxicity. By pointing moderators towards the relevant part of a message, their work can be made both easier and faster. For explicit toxicity, the identification of spans is generally quite straightforward, but things become more difficult the more implicit the toxicity is. To see this, we look at some examples from Pavlopoulos et al. (2021), with the toxic spans being underlined:

▮ Youre an idiot.

The offensive word ‘idiot’ is marked, which corresponds to what was asked of annotators, namely to mark spans that are responsible the toxicity. However, in other examples this is not so clear.

you are a paid liberal hack of stupidity shore living the liberal lie every day

Say the marked span was removed, would calling someone a ‘paid liberal hack’ not still be sufficient to say the comment is toxic? The following example was annotated with no spans, indicating that the entire message is toxic.

Fits the despicable piece of trash like a glove.

But, if ‘the despicable piece of trash’ was replaced by ‘him’, the post no longer seems toxic. So should the toxic span not be limited to that description?

The reason why this task is difficult, is because of the different ways a message can be toxic, and how that relates to the responsibility of any particular span. Some toxicity derives just from the use of a particular word (we refer to this as *form*). If the claims, implications or insinuations being made are innocent, but the choice of words with which they are conveyed are not, then those words alone are responsible for the toxicity. Other toxicity is conveyed without bad words, but still ‘in plain sight’. For example, an assertion made in a simple declarative sentence, like the sentence “all jews should be evicted from white countries” from the introduction (we refer to this as *standing meaning*). For those cases, generally the entire sentence is responsible for the toxicity. Almost any individual word could change what is being asserted into something non-toxic. And finally, for the hardest cases, the toxicity of statements can depend on conversational context, as seen here in an example from Yu et al. (2022):

User 1: That chick needs a high-five in the face with a chair. Damn her for making us look bad!

User 2: A brick is more effective.

Here the toxicity of the second user can only be understood when read as it was originally written, as a reply to the first message (making it an instance of *utterance meaning*). Without the first message’s claim that “That chick needs ... in the face” in the first message, it is unclear what the brick is more effective *for*. So are both messages part of a single toxic span? What if User 1’s message had not been toxic, for example:

User 1 (new): Haha, that kind of grin just calls for a pie to the face ;-)

User 2: A brick is more effective.

The dependency on User 1’s message is still there, but now the toxicity is added by User 2.

Concern 5 (Vague spans). Spans that only indicate if tokens bear any kind of responsibility towards the toxicity are too vague to be useful for context-dependent and implicit toxic language.

5 Toxic Reasoning

We define toxic reasoning as toxic language detection that further requires making explicit the reasons for why something should be considered toxic. The reasons that are given should relate to the conditions under which a category is supposed to apply as specified in its definition. So if a definition only applies if the author intended to harm someone, toxic reasoning would require specifying who is harmed and how. This provides a level of explainability beyond fine-grained categories.

5.1 Approach

Because of concerns 1 through 4 that were raised in Section 3, our approach focuses on characterizing toxicity in terms of what is conveyed. Summarizing, we cite the following benefits:

1. Avoids categories of language being defined while appealing to other categories of language.
2. Avoids the hard problems of world modeling, and theory of mind whenever possible.
3. More explainable, since models are not made to learn statistical patterns about what types of messages have toxic causes or effects.
4. Less subjective, since it avoids the subjectivity in determining what causes led up to, or effects are downstream from, a message that conveys a given implication.
5. More meaningful span annotations that refer to specific aspects of what is conveyed.

We focus on the creation of a toxic reasoning schema, intended primarily to cover implicit toxicity, since that is where progress is needed the most (Ocampo et al., 2023). Like ElSherief et al. (2021), who annotate a dataset with toxic implications, our schema also considers the *implications* of a message. We deviate in the way the implications are characterized. ElSherief et al. annotate fine-grained categories, and then add free-text descriptions of the implications. We advocate for first identifying what implication(s) are present in a (potentially) toxic message, then characterizing those implications using (categories defined in terms of) our *toxic reasoning schema*.

Looking at the implicit hate speech categories defined by ElSherief et al. (2021), we notice that

Examples		① ③ ④ Latent Hatred	⑥ EDOS	② ⑤ Created by authors for illustrative purposes.
① "globalization = white genocide [...]" → Genocided(white people)	④ "you just kill people [...]" → Killer('you')			
② "I'll hire a man for this job" → Hired(man)	⑤ "you Japanese are just smarter" → Intelligent(japanese)			
③ "the white race [...] invented [...] civilization" → Inventors(white race)	⑥ "stop being such a little b*tch" → Timid(women)			

Content	Polarity	
	Negative	Positive
Category		
...Situational	① (genocide)	② (employment)
...Qualitative	⑥ (timidity)	⑤ (intelligence)
...Behavioral	④ (killing)	③ (inventing)

Temporality	past	present	future
		③ ① ④ ⑤ ⑥	②

Specificity	personal	group-member	group-whole
	④	⑤	① ② ③ ⑥

Subject Role	author	interlocutor	other
	① ③	④ ⑤	② ⑥

Other Role [†]	Group Type [‡]

Attitudes	author	ordinary person	expert
Belief 'It is ...'	① TRUE ③ ⑥ FALSE	④ TRUE ① FALSE ⑤ ⑥ ③	TRUE ④ FALSE ① ⑤ ⑥
Desire 'It should be ...'	② TRUE ① FALSE	TRUE ① FALSE	
Intention 'I [will/did] make it ...'	② TRUE FALSE		

Inferential Origin	form	standing meaning	utterance meaning
	⑥	② ③ ④ ⑤	①

† does not apply to selected examples ‡ omitted to conserve space

Figure 1: Application of our Toxic Reasoning Schema, showing how six examples would be characterized. The full text and source for all examples can be found in [Appendix B](#).

many categories are already defined in terms of what is conveyed. We consider two main aspects of what is implied: (1) a proposition predicating something of a subject, and (2) the author (implicitly) reporting an attitude (of belief, desire, or intent) towards this proposition. The second aspect is most relevant when contrasted to the attitude we expect either the general public or experts³ to have toward that same proposition. Given this observation, we hypothesize that toxic reasoning should be approached by including propositional content and attitudes in the schema independently.

5.2 Toxic Reasoning Schema

Here we describe the traits that make up our toxic reasoning schema. In [Figure 1](#), we give examples that cover all the traits of our schema. In [Section 6](#), we show how the schema maps onto existing categories of toxic language, which also motivates why each trait was included.

The schema’s first trait categorizes what is at the core of a message’s implication: the message **content** characterizes the proposition that is central to what the text is conveying. We identify three high-level categories of content.

- *Situation*(subject):
a situation (e.g. environment, circumstance, condition, etc.) applies to the subject.

³By expert, we mean a person with the expertise to assess the truth of the relevant proposition.

- *Quality*(subject):
the subject possesses a given inherent quality, or has a certain nature.
- *Behavior*(subject):
the subject behaves in a particular way.

All can occur with a *Negative* or a *Positive polarity*. *Situation* includes statements where a subject is affected by something external. For example, the text from the introduction that mentions the eviction of a group of people would be a *NegativeSituation*. The second category covers a subject’s inherent qualities or nature. *NegativeQuality* statements could include those describing the subject as ‘sub-human’ or ‘vermin’, whereas a *PositiveQuality* might mention intelligence. Finally, the third category is about a subject’s behavior, *Negative* could be stealing, conspiring, or terrorism, whereas *Positive* could include inventing.

Each of the categories can also apply in a comparative manner. When the statement does not make an absolute statement about the subject, but rather a relative one, where the *Situation*, *Quality*, or *Behavior* of the subject is compared (positively or negatively) to the other. We denote this using a two-place predicate, like *NegativeSituation*(subject, other).

The subject (and the other) can be individuals, or groups. We differentiate between three degrees of **specificity**. If the toxicity is entirely

personal, then no group is mentioned, described or implied at all. Alternatively, the toxicity can be aimed at a *group-member*, where a person is attacked because of the affiliation or relation to a group. And finally, toxicity can be directed at a group as a whole (*group-whole*) without targeting any particular person. This trait captures what previous work calls either directed or generalized hate (Waseem et al., 2017). For our schema, a value of *group-member* indicates that the toxicity is both directed and generalized.

When the specificity of the subject or other involve a group, we also include a **group type** trait for the subject and other. Toxicity can be specified as targeting based on a specific characteristic, like *race*, *nationality*, *religion*, etc. Or, it might be more specifically targeting an instance of that characteristic, like how in the sexism-focused EDOS dataset (Kirk et al., 2023) all categories are defined as being directed at *gender:women*.

The subject and other (if present) fulfill roles with respect to the conversation. These are covered by the **subject role** and **other role** traits. For this trait, the possible values indicate that the subject (or other) of the implication is: the *author*, indicating that the author is talking about (or comparing to) themselves or their in-group (the group the author belongs to); an *interlocutor*, indicating that the author is talking about (or comparing to) another participant in the conversation or a group they belong to; or, *outside*, meaning the author is talking about (or comparing to) a person or group outside the conversation.

The next trait is the **temporality** which specifies the point in time that the implication is meant to apply: the *past*, *present* or *future*.

While we focus on the implications of a message, we recognize that implications can arise from different aspects of a message. Thus, we include the **inferential origin**, which distinguishes between *form*, *standing meaning*, and *utterance meaning*. In Figure 1, asking someone to “stop being such a little bitch”⁴ is taken to imply something negative about women. The message uses a gendered term in a way that implies someone is being scared, timid, or cowardly. By using this term to have that meaning, it implies that women (are more likely to) possess this quality of timidity or cowardice. In this case, the inferential origin is the *form* of the message, rather than its meaning, since the au-

thor may not have meant to imply this, but did so anyway through their choice of words.

Finally, our schema includes the following stakeholder **attitudes**, which describe the attitudes that relevant stakeholders have towards the implication: (i) *author-belief*, does the author believe the implication? (ii) *author-desire*, does the author desire or prefer the implication to be true? (iii) *author-intent*, does the author represent themselves as committed to or accountable for the truth of the implication? For example, when bragging about a past event, or when expressing an intent to make a future event happen. (iv) *typical-belief*, would a typical person believe the implication? (v) *typical-desire*, would a typical person desire or prefer the implication to be true? (vi) *expert-belief*, would an expert believe the implication? Note that what matters here are the attitudes as inferred based on the message and its context, which are not necessarily the actual attitudes held by the stakeholder.

We also considered including the subject group as a stakeholder. However, we are not aware of any categories of toxic language for which the attitude held by the subject group toward the implication determines whether the category applies or not. Therefore, we do not include subject attitudes here.

5.3 Benefits

Better spans. We envision the schema being used in combination with span detection. Rather than identifying spans that bear any kind of responsibility for the toxicity, we envision spans being related to specific traits of our schema. Specifically, different spans can be used to mark different traits in the schema. This includes spans corresponding to the subject of the implication, similar to previous work annotating spans for targets (Barbarestani et al., 2022; Jafari et al., 2024). Returning to the example from Section 4:

User 1: That chick needs a high-five in the face with a chair. Damn her for making us look bad!

The following span annotations could be provided. For content and polarity ‘high-five in the face with a chair’ indicate a *NegativeSituation*. And, for the subject ‘That chick’ indicates a subject role of *outside*, a specificity of *group-member*, and a subject group of *gender:woman*. The explanatory value of spans is improved by highlighting specific aspects of toxic implications.

⁴Taken from Kirk et al. (2023).

Better insight. Categories of toxic language usually have multiple conditions. When an annotator or model judges some text to contain a specific kind of toxic language, they implicitly assert that a set of conditions apply which are sufficient to reach that conclusion. In our schema, these conditions are made explicit, allowing for better insight into models. For example, if a model tends to confuse two classes, we can identify which of the underlying traits are at the root of the problem.

Controlled subjectivity. Röttger et al. (2022) describe two contrasting paradigms for data annotation. The descriptive paradigm encourages annotator subjectivity, while the prescriptive paradigm discourages it. Our viewpoint is that what is useful, is trying to ‘prescribe away’ the subjectivity that arises from disagreement on the (definitions of) categories themselves. What is much more difficult to prescribe than definitions, are world views or belief systems. Some subjectivity is dependent on an annotator’s understanding of the world rather than anything directly related to the annotation. These kinds of subjectivity are generally not (practically) reducible, but the schema does provide us with new options. For example, asking annotators what they think an expert would believe about a given implication, is essentially asking them to give their best-faith impression of what is true (descriptive). Alternatively, one could ask actual experts if they believe certain implications are correct (prescriptive). So in this way, the toxic reasoning schema can facilitate both paradigms.

5.4 Counterarguments

The schema is too complex to annotate. We do not advocate for the community to annotate all of its datasets according to this schema directly. Annotations for the whole schema on a broad dataset will be necessary for validation (to ensure that there are no significant forms of toxicity that the schema cannot capture, for example). And for that reason, we intend to publish such a dataset in the future.

What we advocate the community do differently, is to ensure its dataset category definitions map onto the schema, and that annotators for those datasets adhere to those definitions. The latter can be ensured, for example, by asking (a subset of) annotators validation questions, such as “You chose ‘threat’, this means author intends to harm an individual or group, is that correct?”. This will ensure the schema is adhered to, without making every

annotation campaign considerably more complex and labor intensive. Different datasets will have categories defined based on different conditions. For example, EDOS (Kirk et al., 2023) also has a threat category, but because it is defined there as a subcategory of sexism, rather than toxic language in general, the validation question would include the additional constraint on the *group type* trait: “You chose ‘threat’, this means author intends to harm a woman or group of women, is that correct?”

Not all toxicity is reducible to what is conveyed.

We acknowledge that such categories of toxic language exist. Certainly, the most explicit forms of toxic language (e.g. slurs or swear words) while perhaps theoretically reducible to propositional content, are most easily categorized directly by whether those kinds of words are indeed used in a message. And, insofar as causes and consequences are necessary to define categories, we advocate for an approach where part of the reasoning work is essentially ‘front-loaded’. This approach would require first identifying what kinds of implications tend to be present in messages that are produced by authors with malicious intents, and similarly for consequences. This would still be better than leaving this process up to the classifier, as any such patterns it identifies will likely remain unknown. Uncovering the patterns explicitly and arguing for why it is acceptable to use them allows for a far greater degree of transparency.

6 Mapping categories to the schema

To demonstrate how the schema captures common categories used in existing datasets, we create a mapping to the classes of the IHC (ElSherief et al., 2021). This is not intended to be a definitive mapping, and ideally mappings are created by the authors of taxonomy or classification scheme as they are developed. The constructed mapping can be found in Table 1. Throughout, when we say that a stakeholder should (not) believe, desire, or intend for the proposition to be true, this is reflected in the mapping by including a constraint on the relevant attitude in the mapping. An additional mapping, to the EDOS categories can be found in Appendix A.

In the mappings, the content category and polarity are presented as predicates (e.g. NegativeSituation). When there is only a subject, the predicate has one argument. If there is also an other, the predicate has two arguments.

	Threat	Incitement		Griev.	Inferiority			Misinfo.		
		*	Flaunt.		*	Dehuman.	Tox.	*	Stereotype	
Content	Something(target)	✓	✓
	SomethingNegative(·)
	NegativeSituation(·)
	Harmed(target)	✓	✓
	Harmed(in_group)	.	.	.	✓
	NegativeQuality(·)
	NotHuman(target)	✓
	Vermin(target)	✓	.	.	.
	Inferior(target, in_group)	✓
	SomethingPositive(·)
	PositiveQuality(·)
	Power(in_group)	.	.	✓
	Unity(in_group)	.	.	✓
	Superior(in_group, target)	✓
temporality	future	future
specificity	.	.	grp-*	grp-*	grp-*	grp-*	grp-*	.	grp-*	
Attitudes	author_belief	+	.	+	+	+	+	+	+	.
	author_desire	.	+	.	-
	author_intent	+	-	.	-
	typical_belief	.	.	.	-
	typical_desire	.	.	.	-
	expert_belief	.	.	+	-	.	.	.	-	-
	stereotype	+

Table 1: Mapping to our toxic-reasoning schema, with each column representing an IHC (sub-)class, it contains: (1) a check mark for the content (rows) that could be responsible for that class; (2) the values for the temporality, specificity that are required by that class (or ‘.’ for no requirements); and (3) constraints on the attitudinal attributes for that class (with ‘+’ indicating the attitude is necessary, ‘-’ indicating the attitude is prohibited, and ‘.’ indicating no constraints). The value grp-* is an abbreviation, short for either *group-member* or *group-whole*.

Threat/Intimidation. This class includes texts that “convey a speaker commitment to a target’s pain, injury, damage, loss, or violation of rights” (ElSherief et al., 2021). It specifically also includes non-violent threats about the “implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation”. Inherent in a threat (as defined here) is a *future*-oriented application of a *NegativeSituation*. Furthermore, the author should: believe what they are threatening (not an idle threat); and, seem personally committed to (intend to carry out) the threat. In short, the author believes, and intends for the target to be subjected to a negative situation.

(White) Grievance. When a majority group presents themselves as either the real victims (or presents a minority group as actually having privilege) this is classified as (White) Grievance. For this category the author implies that a *NegativeSituation* applies to their in-group. Additionally, the author must not prefer that the implication is true (not self-hating). To qualify as hateful the grievance must also not have any legitimacy, in other words: a typical person must not believe or prefer for it to

be true; and experts should not believe the implication either.

Inferiority. This category maps on to our schema fairly straightforwardly. Inferiority language includes both: 1) comparisons of two groups, which we model by including *Inferior*(subject, other) as the two-place (relative) version of *SomethingNegative*(subject); or 2) by implying that a target group either lacks a *PositiveQuality* or possesses a *NegativeQuality*. Regarding 2), ElSherief et al. (2021) specifically mention dehumanization and toxification. These two can be modeled by including *NotHuman* and *Vermin* as subcategories of *NegativeQuality*.

Incitement to Violence. The definition given by ElSherief et al. (2021) for incitement is: “flaunting in-group unity and power or elevating known hate groups and ideologies”. We model this by introducing *Power* and *Unity* as properties as specific kinds of *PositiveQuality*, which are asserted about the author’s in-group. The author must *believe* what they are saying, and also find it desirable (they do not regret their in-groups’s power and unity). Experts

should also believe that the author’s in-group has power and/or unity, or it would not be flaunting.

Going beyond the definition given by ElSherief et al. (2021), we also include under incitement any cases that would have been classified as ‘Threat/Intimidation’, except for a lack of author intent. Take the example from the introduction: “all jews are evicted from white countries”. The author of that message is talking about what should happen, and it is not obvious that this person is in a position to enact or enable any eviction of Jewish people, making any intentions unlikely. But they are still clearly expressing a desire for eviction, thus reasonably making it an instance of incitement.

Stereotypes & Misinformation. This category is arguably the broadest category in the IHC. In our schema, misinformation comes down to the author stating something (anything) of a subject that an expert would disagree with. Thus, we model it by including a *Something* category, which unifies all other content categories, and by requiring a lack of expert belief.

Stereotypes are a specific kind of misinformation where some (often negative) characteristics are associated with a group. We model this by introducing an additional attitude for it, thereby characterizing stereotypes as propositions towards which society has a specific kind of attitude.

7 Conclusion

We have proposed toxic reasoning, and our toxic reasoning schema, which separates a message’s main propositional content from the attitudes towards those propositions. After identifying the difficulties associated with defining (categories of) toxicity based on causes and effects, we have defended our choice of focusing on what is conveyed by messages instead.

We have argued that toxic reasoning improves explainability by: (1) enabling explanations that appeal to the conditions on which the predictions are based; and (2) allowing for span annotations that focus on specific aspects of what is implied.

To demonstrate how our toxic reasoning schema is used to ground toxic categories in toxic implications, we have presented a mapping to the IHC and to EDOS. We foresee mappings to our toxic reasoning schema being made for many taxonomies focusing on various kinds of toxic language. We expect that the schema will be useful for rigorously defining categories of toxic language.

In future work, we hope to see datasets annotations that make use of the schema for increased consistency, and decreased subjectivity. With such annotated data, we can also hope for the creation of better, more robust toxic language detection models. Models could potentially be trained to predict the traits of the schema directly, using the mappings to produce predictions for existing taxonomies. If datasets were to be annotated for (parts of) the schema, the traits could also be used as an auxiliary objective during training, or as probing data to uncover if models struggle to represent any traits.

8 Limitations

By focusing on what a message conveys, we do not necessarily avoid the problematic aspects of CAUSE and EFFECT.

There are categories where even if we define it in terms of what is being implied, we still run into causes or effects. Take for example, the case of ‘mansplaining’, which is known in EDOS as “*Animosity – 3.4 Condescending explanations or unwelcome advice*”. We argue that at the core of this type of toxicity lies an implication that women are ignorant, unintelligent, or otherwise in need of having things explained to them. However, crucial to the category is the fact that this is not asserted directly. Instead, the sexist implication originates from the (perceived) motivation behind the explanation. The inferential origin is the CAUSE of the toxicity. This is why in Section 5.1 we say “avoids the hard problems of world modeling, and theory of mind *whenever possible*.” Sometimes, categories appeal to causes or effects by definition.

Acknowledgments

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

References

- Bedour Alrashidi, Amani Jamal, Imtiaz Khan, and Ali Alkhatlan. 2022. *A review on abusive content automatic detection: approaches, challenges and opportunities*. *PeerJ Computer Science*, 8:e1142. Publisher: PeerJ Inc.
- Baran Barbarestani, Isa Maks, and Piek Vossen. 2022. *Annotating Targets of Toxic Language at the Span Level*. In *Proceedings of the Third Workshop on*

- Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 43–51, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. 2025. [Understanding World or Predicting Future? A Comprehensive Survey of World Models](#). *ACM Comput. Surv.*, 58(3):57:1–57:38.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent Hatred: A Benchmark for Understanding Implicit Hate Speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Paula Fortuna and Sérgio Nunes. 2018. [A Survey on Automatic Detection of Hate Speech in Text](#). *ACM Computing Surveys*, 51(4):1–30.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. [Target Span Detection for Implicit Harmful Content](#). In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '24*, pages 117–122, New York, NY, USA. Association for Computing Machinery.
- Md Saroar Jahan and Mourad Oussalah. 2023. [A systematic review of hate speech automatic detection using natural language processing](#). *Neurocomputing*, 546:126232.
- Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. [Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. [SemEval-2023 Task 10: Explainable Detection of Online Sexism](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. [HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875. Number: 17.
- Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. [Towards Conceptualization of “Fair Explanation”: Disparate Impacts of anti-Asian Hate Speech Explanations on Content Moderators](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717, Singapore. Association for Computational Linguistics.
- Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. [An In-depth Analysis of Implicit and Subtle Hate Speech Messages](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.
- John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. [SemEval-2021 Task 5: Toxic Spans Detection](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.
- Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. [A Review on Offensive Language Detection](#). In Mohan L. Kolhe, Shailesh Tiwari, Munesh C. Trivedi, and Krishn K. Mishra, editors, *Advances in Data and Information Sciences*, volume 94, pages 433–439. Springer Singapore, Singapore. Series Title: Lecture Notes in Networks and Systems.
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Bertie Vidgen and Leon Derczynski. 2020. [Directions in abusive language training data, a systematic review: Garbage in, garbage out](#). *PLOS ONE*, 15(12):e0243300. Publisher: Public Library of Science.
- Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. [Introducing CAD: the Contextual Abuse Dataset](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. [Understanding Abuse: A Typology of Abusive Language Detection Subtasks](#). In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. [Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ilker Yildirim and L.A. Paul. 2024. [From task structures to world models: what do LLMs know?](#) *Trends in Cognitive Sciences*, 28(5):404–415.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. [Hate Speech and Counter Speech Detection: Conversational Context Does Matter](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

A EDOS Mapping

The mapping to the classes of EDOS can be found in [Table 2](#).

B Full examples from Figure 1

Each of the examples in [Figure 1](#), by number.

1. “globalization = white genocide. usa next on the list! never in america. keep your guns!”
Source: IHC
Line nr.: 2,250
2. “I’ll hire a man for this job”
Source: Created for illustrative purposes.
3. “agreed. the white race has built and invented every aspect of modern civilisation.”
Source: IHC
Line nr.: 3,932
4. “you just kill people and mosquitoes during your marches”
Source: IHC
Line nr.: 18,412
5. “you Japanese are just smarter”
Source: Created for illustrative purposes.
6. “Stop being such a little bitch”
Source: EDOS ([Kirk et al., 2023](#)) – Table 1
Identifier: ‘sexism2022_english-13065’

	‘Threats, [...] and Incitement’		‘Derogation’			‘Animosity’				‘Prejudiced Discussion’		
	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	4.1	4.2	
Implications	Something(·)	
	SomethingNegative(target)	✓	
	NegativeSituation(·)	
	Harmed(target)	✓	✓	
	Mistreatment(target)	✓	
	SystemicDiscrimination(target)	✓
	NegativeQuality(target)	.	.	✓	✓	✓	.	.
	NotHuman(target)	✓
	SexualObject(target)	✓
	NegativeBehavior(target)	.	.	✓
	AuthorDislike(target)	.	.	.	✓
	Inferior(target, other)	✓
temporality	future	future	
specificity	.	.	grp-*	grp-*	grp-*	grp-w	grp-*	grp-*	grp-*	grp-*	grp-w	
origin	*-mn	*-mn	*-mn	.	*-mn	form	*-mn	u-mn	cause	*-mn	*-mn	
Attitudes	author_belief	+	.	+	+	+	-	+	.	.	.	
	author_desire	.	+	+	+	
	author_intent	+	-	
	typical_belief	
	typical_desire	
	expert_belief	.	.	-	.	.	.	-	.	.	.	
	stereotype	+	.	.	.	

Table 2: Mapping to our toxic-reasoning schema for EDOS. Each column represents an EDOS (sub-)class, it contains: (1) check marks for the implications that could be involved in a sample of that class; (2) the values for the temporality, specificity that are required by that class (or ‘.’ for no requirements); and (3) constraints on the possible values for the attitudinal attributes for that class (with ‘+’ indicating the attitude is necessary, ‘-’ indicating the attitude is prohibited, and ‘.’ indicating no constraints). For EDOS, the group type for all classes would be constrained to *gender:women*. The values *grp-**, *grp-w*, **-mn*, and *u-mn* are abbreviations, with: *grp-w* being short for *group-whole*, *grp-** allowing for either *group-member* or *group-whole*; *u-mn* standing for *utterance-meaning*; and **-mn* allowing for either *utterance-meaning* or *standing meaning*.

Is Sentiment Banana-Shaped? Exploring the Geometry and Portability of Sentiment Concept Vectors

Laurits Lyngbaek*, Pascale Feldkamp*, Yuri Bizzoni,
Kristoffer L. Nielbo, Kenneth Enevoldsen.

Aarhus University, Aarhus, Denmark

*Shared First Authorship

Abstract

Use cases of sentiment analysis in the humanities often require contextualized, continuous scores. Concept Vector Projections (CVP) offer a recent solution: by modeling sentiment as a direction in embedding space, they produce continuous, multilingual scores that align closely with human judgments. Yet the method’s portability across domains and underlying assumptions remain underexplored. We evaluate CVP across genres, historical periods, languages, and affective dimensions, finding that concept vectors trained on one corpus transfer well to others with minimal performance loss. To understand the patterns of generalization, we further examine the linearity assumption underlying CVP. Our findings suggest that while CVP is a portable approach that effectively captures generalizable patterns, its linearity assumption is approximate, pointing to potential for further development. Code available at: github.com/lauritswl/representation-transfer

1 Introduction and Related Works

Sentiment Analysis approaches to data in the Humanities often need continuous sentiment scores to develop meaningful models of texts, for tasks such as tracing the “sentiment arc” of a story (Jockers, 2014; Reagan et al., 2016; Zehe et al., 2016; Bizzoni et al., 2023), gauging sentiment fluctuations in news (Daudert, 2021) or modeling changes in online discourse (Xie and He, 2025), but existing tools struggle to capture the necessary nuances effectively. Many dictionary-based methods are continuous, but struggle with extended context, whereas Transformer models produce binary or ternary outputs that only approximate continuous sentiments through post-hoc adjustments (Bizzoni and Feldkamp, 2023; Lyngbaek et al., 2025).

A recent alternative (Lyngbaek et al., 2025) uses a projection-based method in a homogeneous semantic space to generate continuous sentiment

scores that align with human judgments and match or surpass Transformer-based methods on literary data, while producing smoother distributions. This approach, called Concept Vector Projection (CVP), rests on the “linear representation hypothesis” (Park et al., 2024): the idea that semantic concepts, such as sentiments, can be represented linearly in embedding space (Wehner et al., 2025; Vu and Parker, 2016; Li et al., 2021; Zhao et al., 2024). Under this idea, a given semantic concept corresponds to a *direction* in the embedding space, so that moving further along this direction increases its intensity (see Figure 1).

While studies have validated this idea at various levels of abstraction (Lyngbaek et al., 2025; Wehner et al., 2025; Vu and Parker, 2016; Li et al., 2021; Zhao et al., 2024), its *portability* across different data domains and semantic dimensions remains underexplored. Literary texts, blogs, newspapers, and social media differ in style and affective expression (Feldkamp et al., 2024; Vishnubhotla et al., 2024),¹ and language or period differences can complicate the transfer. The general trend in sentiment analysis has been to assume non-portability and train or fine-tune specialized models for specific domains, languages and historical variants (Allaith et al., 2023; Schmidt and Burghardt, 2018) – models that are then difficult to adapt for other use cases.

In this work, we test the CVP’s portability across three datasets, spanning genres (social media to letters), periods (1798-2013), and languages (English and Danish). After presenting the Data (section 2) and Methods (section 3), we test the CVP through several cross-dataset experiments (subsection 4.1) to assess whether the resulting scores retain their alignment with human judgments. We also explore

¹How much domains differ varies. For example, if using a model fine-tuned on Twitter posts, poetry shows the weakest correlation with human ratings, prose falls in the middle, and Facebook posts show the strongest correlation (Feldkamp et al., 2024).

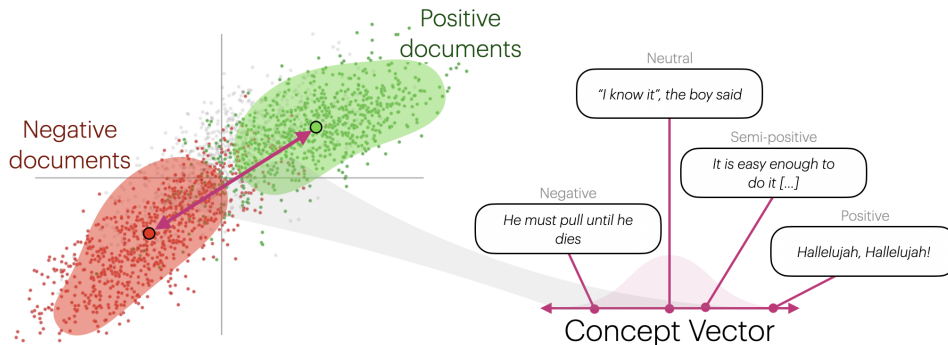


Figure 1: A visualization of how the Concept Vector Projection is constructed. It shows how to construct a positive-negative concept vector to predict sentiment in an unlabeled corpus in a continuous space.

the portability of the CVP beyond valence – to related affective dimensions, such as arousal and dominance (subsection 4.2), and consider whether imperfect linearity in the projections might be the cause of some of the method’s inaccuracies (subsection 4.3).

2 Datasets

To represent diversity across the literary and non-literary domains, we select three datasets that span different genres, domains, languages, and historical periods, all using a human-annotated continuous scale.

Emobank (Buechel and Hahn, 2017) contains sentences from the MASC dataset annotated according to the Valence-Arousal-Dominance (VAD) scheme (Mehrabian and Russell, 1974). The dataset includes: Letters, Blog, Newspaper, Essays, Fiction, and Travel guides.

Facebook (Preotiuc-Pietro et al., 2016) consists of status updates collected by (Kosinski et al., 2013) and annotated for valence and arousal.

Fiction4 (Feldkamp et al., 2024) comprises literary texts spanning four genres and two languages (English/Danish) from the 19th and 20th centuries. It consists of three main authors – Sylvia Plath (poetry), Ernest Hemingway (prose), and H.C. Andersen (fairytales) – and hymns from Danish official church hymnbooks (published 1798-1873). Two or more human annotators scored each sentence (/line, for poetry) for valence.²

²Although lower than Facebook posts, IRR for Fiction4 ($\alpha=0.67$) is high for continuous annotations of literary texts. Humans rarely reach $\alpha > 0.80$ for polarity tagging on *non-literary* texts (Wilson et al., 2005) and achieve lower IRR for continuous scales on literary texts (Batanović et al., 2020; Rebora et al., 2023).

Dataset	Period	Sentences	Kripp. α (Scale)
EmoBank	1990–2008	10,062	
Valence			.34 (1-5)
Arousal			.25 (1-5)
Dominance			.22 (1-5)
Facebook	2012–2013	2,895	
Valence			.72 (1-9)
Arousal			.82 (1-9)
Fiction4	1798–1965	6,300	
Valence			.67 (0-10)

Table 1: Summary of annotated corpora. We report sentence counts, average length, and inter-rater agreement (α). The total number of sentences considered is $n = 19,257$. Full breakdown of subgenres (in Fiction4 and Emobank) and number of annotators in Appendix D.

3 Methods

To construct the concept vector, we follow the approach introduced by (Lyngbaek et al., 2025), where a pre-trained sentence-embedding model \mathbf{M} embeds a set of source_{negative} and target_{positive} exemplar sentences. We compute the mean embeddings of source_{negative} and target_{positive} examples and define the concept vector $\hat{\mathbf{v}}$ as the unit vector of the difference between mean embeddings. The assumption is that this averaging will reduce non-sentiment information to Gaussian noise with a mean of zero, leaving the sentimental signal behind (Kim et al., 2018; Zhao et al., 2024). With this method, we score a sentence s by projecting its embedding onto the concept vector $\hat{\mathbf{v}}$ via the dot product $\mathbf{M}(s) \cdot \hat{\mathbf{v}}$, yielding a continuous sentiment score. We normalize the scores using a z-score normalization. We define the details for the CVP algorithm in Appendix B.³ To define source_{negative} and target_{positive}, we set sentiment thresholds relative to each corpus’ valence distribution. Sentences

³Implementation available at [representation-transfer](#)

at least one standard deviation above the mean are positive, sentences below by one standard deviation are negative, and the rest are neutral (for the formalization, see [Appendix A](#)). We estimate *concept vectors* from these positive–negative contrasts, capturing the extremes rather than absolute ratings. This approach yields comparable sentiment contrasts across datasets with different scales and distributions. For testing linearity, we created three concept vectors: positive–negative, negative–neutral, and neutral–positive.

3.1 Model

To allow for comparability with previous works ([Lyngbaek et al., 2025](#)), we use the embedding model paraphrase-multilingual-mpnet⁴ ([Reimers and Gurevych, 2019](#)), a 278M-parameter model based on a mean-pooled BERT architecture optimized for sentence similarity via Siamese and Triplet networks. This model is notable for its multilingual capabilities, previous performance ([Lyngbaek et al., 2025](#)), and excellent size-to-performance ratio.⁵

4 Results

4.1 Portability

Our results show that the projection method is robust: continuous valence scores remain well-aligned to human scores across all three datasets and their constituent subgenres ([Table 2](#)), suggesting that the approach captures generalizable sentiment patterns beyond the idiosyncrasies of literary, journalistic, or social media language. It highlights the portability of continuous sentiment scoring across genres, which can be crucial for research spanning multiple text types or for investigating historical and contemporary corpora side by side.

Dataset	Correlation, when trained on:		
	Fiction4	Emobank	Facebook
Fiction4	0.66	0.65	0.64
Emobank	0.67	0.70	0.66
Facebook	0.66	0.66	0.68

Table 2: Spearman correlations between human and projected valence scores across corpora. Values indicate correlations when trained on the indicated corpus (columns) and tested on itself or another corpus (rows).

⁴[sentence-transformers/paraphrase-multilingual-mpnet-base-v2](#)

⁵A larger model may increase model correlation with human scores at the expense of computation budget.

	Emobank	Facebook	Fiction4
Valence	.71±.02(.70)	.70±.02(.68)	.66±.02(.66)
Arousal	.36±.02(.42)	.65±.02(.67)	
Dominance	.35±.01(.37)		

Table 3: Cross-validation of Spearman correlations between CVP scores and human scores for valence, arousal, and dominance per corpus. The scores are the mean correlation obtained from a five fold analysis, with a standard deviation notated by ±. The score parenthesis indicates the Spearman correlation obtained when no split was conducted. Only Emobank has human scores of all V-A-D labels.

4.2 Beyond Valence

To test the CVP’s ability to generalize beyond valence – which refers to the positivity/negativity spectrum – we tested the approach on semantic properties associated with valence in sentiment analysis: arousal and dominance. Arousal refers to the intensity of the concept conveyed by a given word (*ecstatic* and *serene* are both positive, but the first word elicits a higher arousal); dominance refers to the amount of control associated with a term (*angry* and *helpless* are both negative, but the first word has more dominance). We find that CVP generalizes well for these subtler concepts ([Table 3](#)) with similarly continuous distributions (see [Appendix F](#)), without reaching the performance achieved on valence.

4.3 Linearity assumption

CVP treats sentiment as linear in embedding space: negative and positive extremes form the main axis, with neutral sentences in the middle. We create similar vectors with negative–neutral and neutral–positive extremes, and examine the cosine similarity between all three vectors ([Figure 3](#)). Across corpora, the negative–positive axis aligns most strongly with the other two axes, consistent with a geometrical interpretation that neutral texts are located close to the axis, but vary along an undefined semantic dimension, so that the three vectors will form a triangle outlining the centroids of the three classes. We explore the geometry of Fiction4’s valence space by creating a two-dimensional basis that visualizes the data. We define our first dimension as the negative-positive vector, \mathbf{v}_{np} . The second semantic dimension we define as the neutral-component, \mathbf{v}_{nc} . The neutral-component vector captures the remaining semantic information encoded in the neutral centroid.

To define the neutral component, we treat the

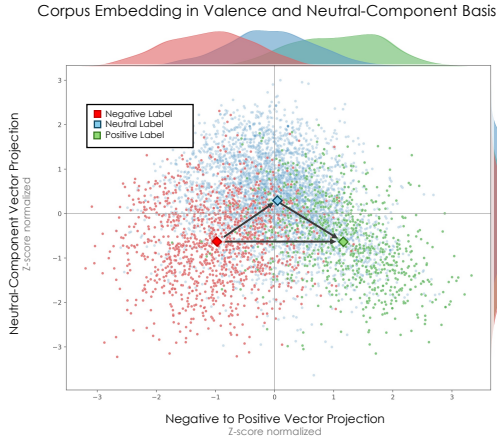


Figure 2: Scatterplot visualizing the Fiction4 embeddings projected onto the Fiction4 Pos-Neg Sentiment vector and the corresponding Neutral-Component. The Marginal plots are Kernel Density Estimations of the label distributions. All dimensions are Z-score normalized to make the projections interpretable.

problem geometrically in an affine space. Because an affine space has no natural origin, projections cannot be applied directly to centroids. Instead, we work with *difference vectors*, which encode relative positions between centroids. Let \mathbf{v}_{np} denote the vector from the negative to the positive centroid and define the corresponding unit direction

$$\hat{\mathbf{v}}_{np} = \frac{\mathbf{v}_{np}}{\|\mathbf{v}_{np}\|}.$$

We also define the vector from the negative to the neutral centroid as \mathbf{v}_{nn} . Projecting \mathbf{v}_{nn} onto $\hat{\mathbf{v}}_{np}$ gives the scalar projection

$$k = \mathbf{v}_{nn} \cdot \hat{\mathbf{v}}_{np}.$$

This scalar specifies the relative position of the neutral centroid along the negative–positive axis, yielding the projected component

$$k \hat{\mathbf{v}}_{np}.$$

Finally, we construct the *neutral component vector* by removing this projected information:

$$\mathbf{v}_{nc} = \mathbf{v}_{nn} - k \hat{\mathbf{v}}_{np}.$$

This residual vector represents the component of the neutral direction that is orthogonal to the sentiment axis. We use this constructed basis to visualize the geometric structure of sentiment embeddings, as seen in Figure 2. This result aligns with the high-dimensional cosine-similarity observed in Figure 3.

	neg-pos	0.83	0.80	0.86	0.76	0.72	0.76	0.56	0.67
Fiction4	neg-pos	1.00	0.83	0.80	0.86	0.76	0.72	0.76	0.56
	neut-pos	0.83	1.00	0.34	0.65	0.77	0.34	0.61	0.60
	neg-neut	0.80	0.34	1.00	0.77	0.47	0.86	0.64	0.31
	neg-pos	0.86	0.65	0.77	1.00	0.87	0.86	0.78	0.55
Emobank	neg-pos	0.76	0.77	0.47	0.87	1.00	0.49	0.76	0.72
	neut-pos	0.76	1.00	0.49	0.87	1.00	0.49	0.76	0.72
	neg-neut	0.72	0.49	1.00	0.86	0.49	1.00	0.59	0.23
	neg-pos	0.76	0.61	0.64	0.78	0.76	0.59	1.00	0.84
Facebook	neg-pos	0.56	0.60	0.31	0.55	0.72	0.23	0.84	1.00
	neut-pos	0.56	1.00	0.29	0.55	0.72	0.23	0.84	1.00
	neg-neut	0.67	0.29	1.00	0.67	0.72	0.76	0.77	0.29
	neg-pos	0.67	0.36	0.76	0.72	0.48	0.76	0.77	0.29
	neut-pos	0.36	1.00	0.76	0.48	0.76	0.77	0.29	1.00
	neg-neut	0.76	0.76	0.76	1.00	0.76	0.76	0.76	0.76

Figure 3: Cosine similarity between Concept Vectors for each corpus (values in each cell). Internal correlations among neg-pos, neut-pos, and neg-neut pairs are strong, with neut-pos and neg-neut closer to neg-pos, reflecting a centrality of the negative–positive axis across corpora.

We see that our embeddings tend to be linear, but that neutral embeddings encode spurious information that remains unaccounted for in the sentiment direction. This property gives the centroids a triangular shape – and structures the Fiction4 embeddings as a banana-shaped manifold.

5 Discussion & conclusions

We find that Concept Vector Projections transfer well across genres, periods, and languages — a vector derived from a corpus including 19th-century Danish hymns predicts sentiment in contemporary Facebook posts nearly as well as one trained on in-domain data. This portability suggests CVP captures generalizable properties of how sentiment is encoded in embedding space, rather than domain-specific patterns. The approach also extends to arousal and dominance, though with reduced performance, consistent with valence being the most reliable and consistent of the three affective dimensions (Warriner et al., 2013). For researchers working with historical or low-resource corpora, this means domain-specific training data may not be necessary – as suggested in Lyngbaek et al. (2025).

Our geometric analysis shows that the linearity assumption is approximate: neutral sentences do not lie exactly on the positive–negative axis but form a continuous, banana-shaped curve. This suggests neutrality encodes semantic content beyond the absence of valence – a property that future methods might exploit.

Limitations

While the goal of this work, is not to explore how to optimize the performance of the CVP, but examine its implications, we only examine one model to ensure comparability with [Lyngbaek et al. \(2025\)](#). Further analysis should explore alternative models as indicate evidence suggests that newer models like EmbeddingGemma ([Vera et al., 2025](#)) might surpass the one currently used.

To examine the CVP ability to generalize to the related concepts arousal and dominance, we utilize the dataset itself as the source dataset for deriving the concept vector. This likely leads to a modest overestimation of the correlation as we see for valence in [Table 2](#) and arousal in [Appendix G](#).

Our cross-lingual evaluation, while leveraging a multilingual embedding model, is restricted to Danish and English. These languages, though differing in resource availability, belong to the same language family; generalization to typologically distinct languages remains untested.

Acknowledgments

This work was partially supported by the Danish National Research Foundation (DNRF193) through TEXT: Center for Contemporary Cultures of Text, Aarhus University.

References

- Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershovich. 2023. [Sentiment Classification of Historical Danish and Norwegian Literary Texts](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.
- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. [XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.
- Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. [A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts](#). *PLoS ONE*, 15(11).
- Yuri Bizzoni and Pascale Feldkamp. 2023. [Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study](#). In *Proceedings of the 3rd International Workshop on Natural Language Processing for Digital Humanities*, pages 219–226, Tokyo, Japan. Association for Computational Linguistics.
- Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. [Sentimental matters - predicting literary quality by sentiment analysis and stylometric features](#). In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.
- Sven Buechel and Udo Hahn. 2017. [EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Tobias Daudert. 2021. [Exploiting textual and relationship information for fine-grained financial sentiment analysis](#). *Knowledge-Based Systems*, 230:107389.
- Pascale Feldkamp, Ea Overgaard Lindhardt, Kristoffer L. Nielbo, and Yuri Bizzoni. 2024. [Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond](#). In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 681–706.
- Matthew Jockers. 2014. [A Novel Method for Detecting Plot](#). Matthew L. Jockers Blog.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018.

- Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.
- Michal Kosinski, David Stillwell, and Thore Graepel. 2013. [Private traits and attributes are predictable from digital records of human behavior](#). *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.
- Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. [Implicit representations of meaning in neural language models](#). 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 1813–1827, Online. Association for Computational Linguistics.
- Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer L. Nielbo, and Kenneth Enevoldsen. 2025. [Continuous Sentiment Scores for Literary and Multilingual Contexts](#). In *Anthology of Computers and the Humanities*, volume 3, pages 480–497, Luxembourg. ACH.
- Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology*. the MIT Press.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The Linear Representation Hypothesis and the Geometry of Large Language Models](#). *arXiv preprint*. ArXiv:2311.03658 [cs].
- Daniel Preoticiu-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. [Modelling valence and arousal in Facebook posts](#). In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.
- Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. [The Emotional Arcs of Stories Are Dominated by Six Basic Shapes](#). *EPJ Data Science*, 5(1):1–12.
- Simone Rebora, Marina Lehmann, Anne Heumann, Wei Ding, and Gerhard Lauer. 2023. [Comparing ChatGPT to human raters and sentiment analysis tools for german children’s literature](#). In *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023*, volume 3558 of *CEUR Workshop Proceedings*, pages 333–343. CEUR-WS.org.
- 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.
- Thomas Schmidt and Manuel Burghardt. 2018. [An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing](#). In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftexhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. [Embeddinggemma: Powerful and lightweight text representations](#). *Preprint*, arXiv:2509.20354.
- Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif Mohammad. 2024. [The emotion dynamics of literary novels](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2557–2574, Bangkok, Thailand. Association for Computational Linguistics.
- Thuy Vu and D. Stott Parker. 2016. [k-embeddings: Learning conceptual embeddings for words using context](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1262–1267, San Diego, California. Association for Computational Linguistics.
- Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. [Norms of valence, arousal, and dominance for 13,915 English lemmas](#). *Behavior Research Methods*, 45(4):1191–1207.
- Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. [Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models](#). *arXiv preprint*. ArXiv:2502.19649 [cs].
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. [Recognizing contextual polarity in phrase-level Sentiment Analysis](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Haihua Xie and Miao He. 2025. [Tracking Fine-Grained Public Opinions: Two Datasets from Online Discourse on Trending Topics](#). *Mathematics*, 13(21):3433.
- Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. [Prediction of Happy Endings in German Novels Based on Sentiment Information](#). In *Interactions between Data Mining and Natural Language Processing*, pages 9–16, Riva del Garda.
- Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2024. [Beyond single concept](#)

vector: Modeling concept subspace in llms with gaussian distribution. *arXiv preprint arXiv:2410.00153*.

A Vector polarity selection procedure

For each text unit i in a corpus, we compute its valence score v_i . Let μ and σ denote the mean and standard deviation of valence scores across the corpus. Polarity labels are assigned as follows:

$$\text{label}_i = \begin{cases} \text{positive,} & \text{if } v_i \geq \mu + \sigma \\ \text{negative,} & \text{if } v_i \leq \mu - \sigma \\ \text{neutral,} & \text{otherwise} \end{cases}$$

This scheme assigns a label based on deviation from the corpus mean by one standard deviation.

B CVP algorithm

The following algorithm formally describes the procedure for defining and applying a concept vector by using labeled sentence embeddings.

Algorithm 1 Concept Vector Projection

Input:
 \mathbf{M} = Language Model
 \mathbf{S} = A set of sentences s_i , labeled via mean \pm SD thresholds for valence: $s_i \in \{\text{positive}^+, \text{negative}^-, \text{neutral}^\emptyset, \text{unknown}^?\}$

Output:
 $\hat{\mathbf{v}}$ = Concept vector
 $\text{score}(s_i)$ = projection scores for unknown sentences

Computation:

- 1: Embed all sentences: $\mathbf{e}_i = \mathbf{M}(s_i)$
- 2: $P^+ \leftarrow \{\mathbf{e}_i \mid s_i = \text{positive}\}$
- 3: $N^- \leftarrow \{\mathbf{e}_i \mid s_i = \text{negative}\}$
- 4: Compute means: $\mu_{S^+} = \text{mean}(P^+)$, $\mu_{S^-} = \text{mean}(N^-)$
- 5: Compute concept vector: $\vec{\mathbf{v}} = \mu_{S^+} - \mu_{S^-}$
- 6: Normalize: $\hat{\mathbf{v}} = \frac{\vec{\mathbf{v}}}{\|\vec{\mathbf{v}}\|}$
- 7: **for each** embedding \mathbf{e}_i **do**
- 8: $\text{score}(s_i) = \mathbf{e}_i \cdot \hat{\mathbf{v}}$
- 9: **end for**
- 10: Standardize scores: $\frac{\text{score}(s_i) - \text{mean}(\text{score}(s_i))}{\text{std}(\text{score}(s_i))}$ // Embedding projection

C Performance baseline

To contextualize the correlations between the CVP and human scores, we also include the correlations between a transformer-based model and human scores. We choose the best-performing model in Lyngbaek et al. (2025), the multilingual cardiffnlp/xlm-roberta-base-sentiment-multilingual (here abbreviated xlm-R-b)⁶, which is an xlm-roberta model finetuned for sentiment on Twitter data (Barbieri et al., 2022). The model’s output was transformed using its confidence scores, consistent with the approach in Lyngbaek et al. (2025) and Bizzoni and Feldkamp (2023). We do not compare to continuous dictionary-based sentiments approaches like VADER, but for a comparison against these methods, we refer to Lyngbaek et al. (2025).

Note that while the **xlm-R-b** model performs better than the Concept Vector Projection on Facebook data in terms of Spearman’s ρ (see Table 2), the distributions of these scores remain pseudo-trinary (see the Figure 4), unlike the distribution of the Concept Vector Projection’s scores (Figure 5).

⁶<https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual>

Dataset	Subcategory	xlm-R-b	Correlation, when trained on:		
			Fiction4	Emobank	Facebook
Fiction4	<i>overall</i>	0.60	0.66	<u>0.65</u>	0.64
	<i>fairytales</i>	0.62	0.67	<u>0.64</u>	0.61
	<i>hymns</i>	0.59	0.67	<u>0.66</u>	0.62
	<i>poetry</i>	0.57	0.72	<u>0.71</u>	0.68
	<i>prose</i>	0.61	0.64	<u>0.62</u>	<u>0.63</u>
Emob.	<i>overall</i>	0.65	0.67	0.70	0.66
	<i>SemEval</i>	0.64	<u>0.66</u>	0.71	0.65
	<i>blog</i>	0.65	0.64	<u>0.68</u>	0.69
	<i>essays</i>	0.58	<u>0.59</u>	0.63	0.55
	<i>fiction</i>	0.56	<u>0.67</u>	0.69	<u>0.67</u>
	<i>letters</i>	0.68	<u>0.68</u>	0.71	0.66
	<i>newspaper</i>	<u>0.65</u>	<u>0.67</u>	0.69	0.65
	<i>travel-guides</i>	0.49	<u>0.56</u>	<u>0.58</u>	0.59
FB	<i>overall</i>	0.74	<u>0.66</u>	<u>0.66</u>	0.68

Table 4: Correlations with human and projected valence scores across corpora. Values indicate correlations when trained on the indicated corpus (columns) and tested on the datasets overall and across subgenres (rows). Correlation of the transformer-based model and human score is indicated in column **xlm-R-b**.

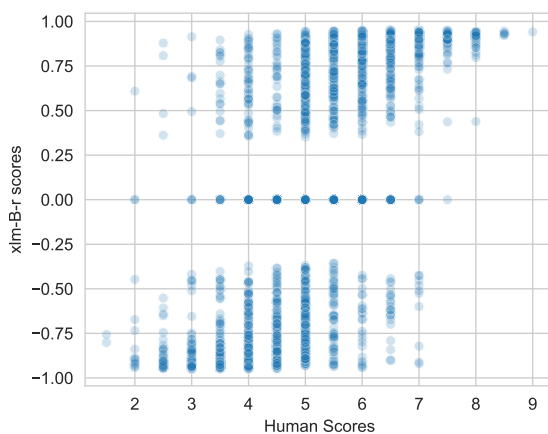


Figure 4: Correlation (Spearman’s ρ) between transformer model (xlm-R-b) and human scores in the Facebook dataset.

D Data details

A note on *EmoBank* categories: *Essays* include, i.e., “A Brief History of Steel in Northeastern Ohio”. *Fiction* comprises prose pieces, i.e., Richard Harding’s “A Wasted Day” or the SciFi story “Captured Moments”. *Newspapers* contain reports and longer reportages. *Travel Guides* include both local histories and reflective pieces (e.g., “Dublin and the Dubliners”).⁷

⁷See the full MASC corpus at: <https://anc.org/data/masc/corpus/browse-masc-data/>

Dataset	Period	N annotations	N words	\bar{x} words/sentence	N annotators	Krippendorff's α		
						V	A	D
→ <i>Facebook</i>	2012-2013	2,895	46,868	16.19	2	.72	.82	-
↓ <i>EmoBank</i>	1990-2008	10,062	151,259	15.03	10	.34	.25	.22
Letters		1,413	21,639	15.31	10	.35	.25	.25
Blog		1,336	20,874	15.62	10	.32	.22	.18
Newspaper		1,314	25,992	19.78	10	.30	.22	.22
Essays		1,135	26,349	23.21	10	.33	.21	.21
Fiction		2,753	31,491	11.44	10	.35	.22	.22
Travel-guides		919	17,154	18.67	10	.28	.23	.23
SemEval		1,192	7,760	6.51	10	.37	.20	.20
↓ <i>Fiction4</i>	1798-1965	6,300	73,250	11.6	≥ 2	.67	-	-
🇩🇰 Hymns	1798-1873	2,026	12,798	6.3	2	.72	-	-
🇩🇰 Fairy tales	1837-1847	772	18,597	24.1	3	.69	-	-
Prose	1952	1,923	30,279	15.7	2	.63	-	-
Poetry	1965	1,579	11,576	7.3	3	.59	-	-

Table 5: Datasets with valence annotation. Valence was annotated on a sentence basis, so ‘N annotations’ indicates the number of sentences. ‘N annotators’ indicates the number of annotators reported per sentence. IRR per dataset and category is shown in α . Since *EmoBank* lacks unique annotator IDs, we cannot correlate individual annotators’ scores. Therefore, we use Krippendorff’s α measures agreement across V-A-D ratings per item in the full dataset and in subcategories. Only Emobank includes the full V-A-D annotation. Note that texts not indicated as Danish (flag) are all in English.

E Train to test dataset correlations

Visualizations of portability between datasets for valence. This figure is a visualization of [Table 2](#).

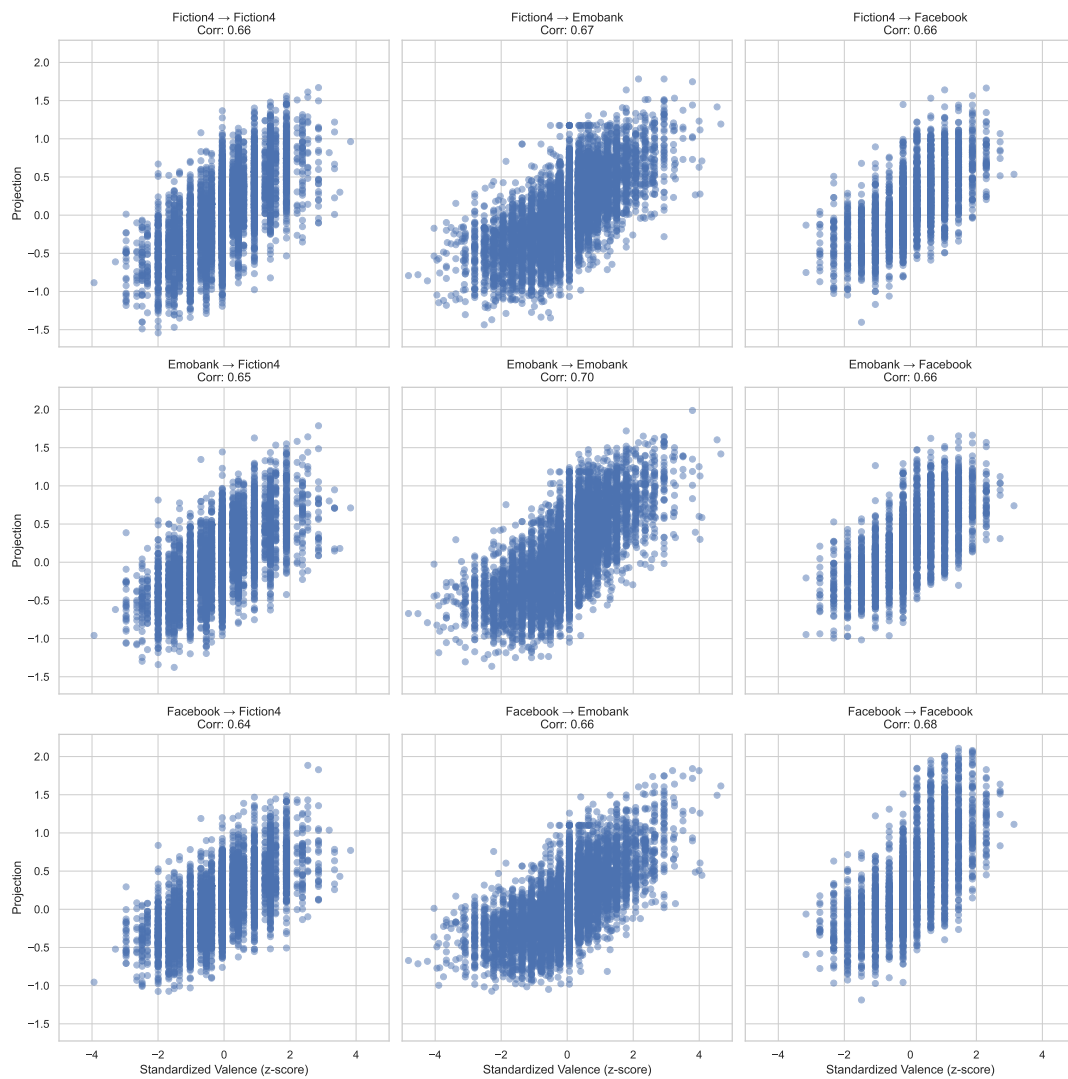


Figure 5: Relation between Concept Vector Projection scores (y-axis) and human scores (x-axis) on standardized valence across datasets. On top of each figure, the training set (on the left of the arrow) and the test set (on the right of the arrow) are shown.

F Beyond valence, visualized correlations

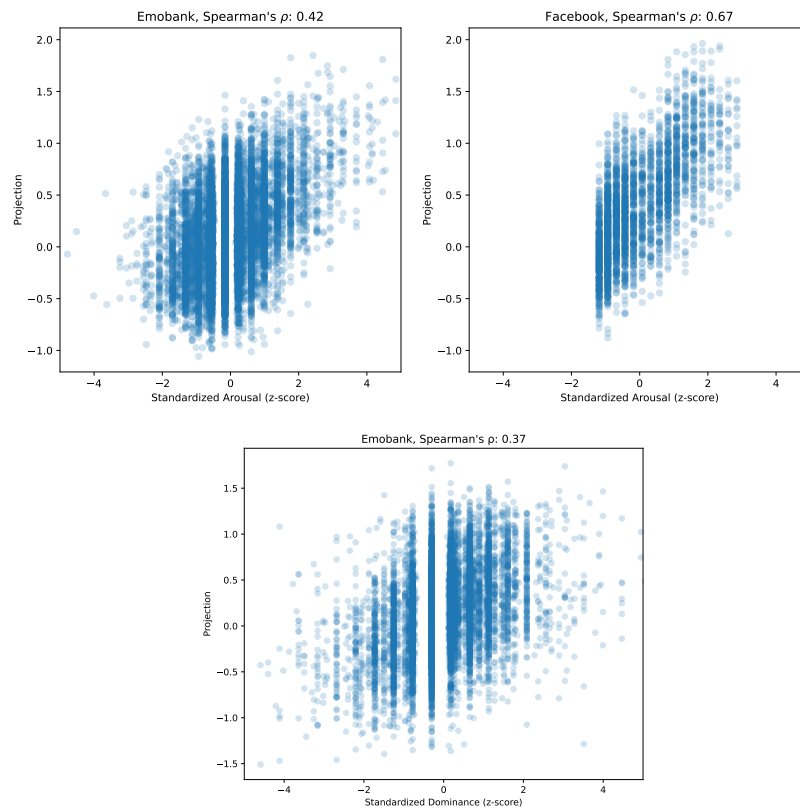


Figure 6: Top: Relation between Concept Vector Projection scores (y-axis) and human scores of standardized **arousal** in the Emobank and Facebook corpora. Bottom: Relation between Concept Vector Projection scores (y-axis) and human scores of standardized **dominance** in the Emobank corpus.

G Portability of Arousal

Visualizations of portability between datasets for arousal. This figure is a visualization of [Table 2](#).

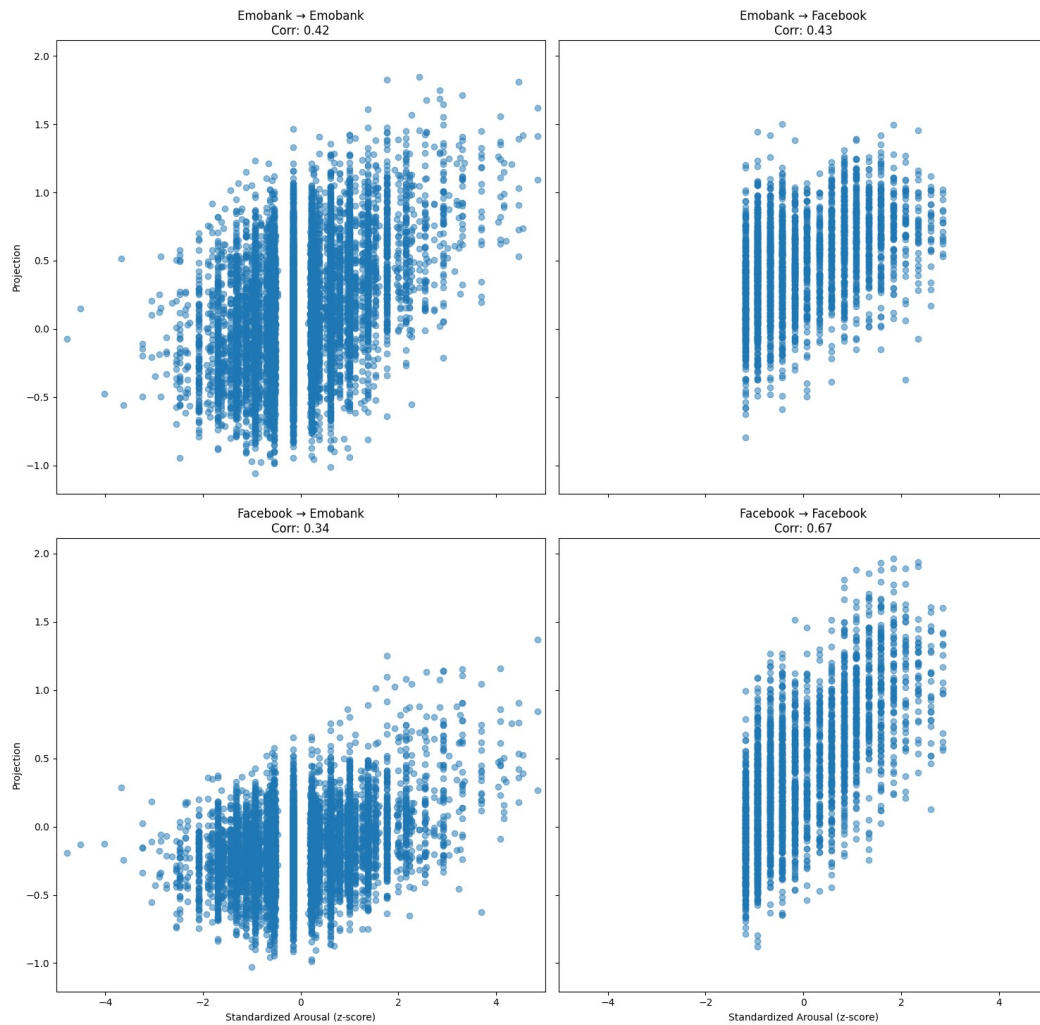


Figure 7: Relation between Concept Vector Projection scores (y-axis) and human scores (x-axis) on standardized arousal across datasets. A title such as Emobank→Facebook should be read as: Correlation between projections of arousal and human arousal ratings, when arousal vector is defined by the Emobank corpus and predictions are tested on the Facebook corpus.

H Downstream differences between Human Annotators and Projection models

As a sanity check on downstream analysis using projection scores instead of human annotators we tested a simple hypothesis. That both high and low valence scores correlate with high arousal. This would imply that arousal only correlates with valence, when we use the absolute value of valence (i.e. distance from the mean). While the slope of our linear regression varies between the two methods, we reach the same conclusion with both models. That there is a positive relation between absolute valence and arousal scores.

Correlation between Valence and Arousal for original ratings and projected ratings:

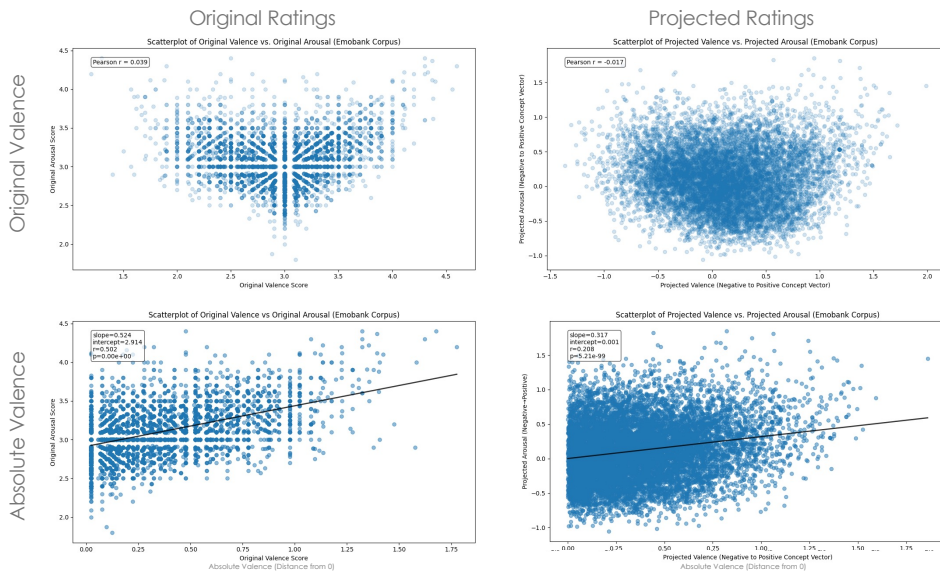


Figure 8: Scatterplots showing the relation between arousal and valence. The two top plots show no correlation between valence and arousal. The two bottom plots use absolute valence instead of valence, and depicts a positive significant relationship between absolute valence and arousal. Left side plots uses human annotations of EmoBank. Right side plots use projected ratings of EmoBank, and using the pos-neg vector defined on EmoBank.

Disentangling Emotion Understanding and Generation in Large Language Models

Sadegh Jafari, Els Lefever and Véronique Hoste

LT3, Language and Translation Technology Team

Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

{sadegh.jafari, els.lefever, veronique.hoste}@ugent.be

Abstract

Large language models (LLMs) have demonstrated strong performance on emotion understanding tasks, yet their ability to faithfully generate emotionally aligned text remains less well understood. We propose a semantic evaluation framework that jointly assesses emotion understanding, emotion generation, and internal consistency, using a VAE-based emotion cost matrix that captures graded semantic similarity between emotion categories. Our framework introduces four complementary metrics that disentangle baseline understanding, human-perceived emotion in generated text, generation quality, and model consistency. Experimental results show that while understanding and consistency scores are highly correlated, emotion generation exhibits substantially weaker correlations with these metrics. These findings motivate the development of specialized evaluation protocols that independently measure emotional understanding and generation, enabling more reliable assessments of LLM emotional intelligence.

1 Introduction

The advancement of LLMs has brought their emotional intelligence to the forefront of AI research, as it is a critical component for effective and meaningful human-computer interaction (Wang et al., 2023; Li et al., 2023; Hu et al., 2025). The ability of these models to comprehend and express emotions is essential for their integration into real-world applications that require social awareness and empathy (Sabour et al., 2024; Ishikawa and Yoshino, 2025). Within this domain, two fundamental yet distinct capabilities are often discussed: **emotion understanding**, the capacity to accurately identify emotions from text, and **emotion generation**, the ability to produce text that conveys a specific, appropriate emotion (Li et al., 2024; Liu et al., 2025). While both are integral to emotional intelligence, a key challenge lies in evaluating them, as current

benchmarks often focus more on recognition than other essential emotional skills (Sabour et al., 2024; Zhang et al., 2025). This may lead to an incomplete assessment of an LLM’s true emotional capabilities.

This paper argues that evaluating the emotion generation capability of LLMs is fundamentally different from evaluating their emotion understanding abilities. To investigate this distinction, we conduct experiments with five distinct LLMs (see Table 1). Our methodology centers on the UniC dataset (Du et al., 2025), which consists of expert-labeled conversational texts annotated with emotions. We adopt a two-phase process involving emotion neutralization followed by emotion re-injection, prompting the models to generate text samples that align with the original emotion labels provided in UniC. To establish a reliable ground truth for the generated texts, all generated samples were annotated by human evaluators.

By analyzing the performance of the five LLMs across emotion understanding (OERS, GERS), emotion generation (EGS), and internal consistency (ECS) metrics (see Section 4), this study aims to demonstrate the divergence between their ability to understand emotion and their ability to generate emotionally aligned text. Our central hypothesis is that high performance on understanding-based metrics does not necessarily translate to high performance on generation-based metrics. Our findings reveal a stark contrast in the relationships between these scores: the average correlation among the understanding (OERS, GERS) and consistency (ECS) metrics is 95%, indicating that they measure similar constructs. However, the average correlation between the Generation Score (EGS) and these other three metrics (OERS, GERS, and ECS) is only 73%. This discrepancy strongly suggests that the other metrics are not good representatives for evaluating an LLM’s emotion generation capability. These findings underscore the need for specialized

evaluation protocols that can independently assess both understanding and generation, thereby providing a more accurate and nuanced picture of an LLM’s emotional intelligence and guiding the development of more genuinely empathetic artificial agents (Li et al., 2024; Liu et al., 2025).

2 Related Works

Emotional Intelligence (EI) is broadly defined as the ability to manage one’s own emotions and to understand the emotions of others (Sabour et al., 2024). As LLMs become more integrated into human-centric applications, imbuing them with EI has become a critical area of research (Wang et al., 2023; Raj, 2024). In the context of LLMs, this dual capability translates into two distinct research strands: emotion understanding (perceiving emotion in text) and emotion generation (expressing emotion through text).

A significant body of research has focused on benchmarking an LLM’s emotion understanding, which corresponds to the ability to "recognize emotions" (Samad, 2014). This is most commonly evaluated as a text classification task using established datasets. The GoEmotions dataset, for instance, is frequently used to test model performance (Demszky et al., 2020). Recent work has used GoEmotions to compare various LLMs, analyzing their accuracy in identifying human emotions from text (Lecourt et al., 2025). However, some researchers argue that such datasets, while useful, are insufficient for a comprehensive evaluation and have proposed new benchmarks like EmoBench, which includes hand-crafted questions designed to require deeper emotional reasoning beyond simple recognition (Sabour et al., 2024). These studies treat classification accuracy on such benchmarks as a primary indicator of an LLM’s emotion understanding capability.

In parallel, research on emotion generation focuses on the other aspect of EI: the ability to "invoke and reason with emotions" (Samad, 2014). For an LLM, this translates into generating text that is both emotionally appropriate and coherent. This has motivated the development of techniques aimed at enhancing the emotional expressiveness of model outputs. For instance, the Emotional Chain-of-Thought (ECoT) method prompts LLMs to reason about emotions before generating a response, thereby improving their performance on generation tasks (Li et al., 2024). Evaluating the quality of

such emotionally generated text presents its own challenges, which have led to the proposal of new metrics. The Emotional Generation Score (EGS) evaluates generated outputs based on psychological theories (Li et al., 2024). This metric is different from the EGS defined in the Equation 3, although the two share a similar name. In contrast, Jafari et al. (2025) introduced an embedding-based automatic evaluation metric for emotional text generation.

While these two components of EI are often studied in isolation, some prior work has acknowledged the need to connect them for a more holistic view of emotional intelligence (Zhao et al., 2024). However, our work is motivated by the hypothesis that strong performance in emotion understanding does not necessarily translate to high performance in emotion generation. We directly investigate this potential divergence by quantitatively comparing these two capabilities across multiple LLMs to empirically demonstrate the degree to which they differ and motivate the need for distinct evaluation metrics for understanding versus generation.

3 Dataset

We use two different datasets: one for evaluating emotional text understanding and another for evaluating emotional text generation. The UniC dataset (Du et al., 2025) is used for the understanding evaluation, while the generation evaluation relies on texts produced by various LLMs that are annotated with emotion labels. The details of each component are provided below.

3.1 Emotion Understanding Dataset

A considerable body of research has investigated the emotion understanding capabilities of LLMs by evaluating them on established emotion classification datasets. Many of these datasets originate from social media platforms or news headlines. For example, the widely used GoEmotions dataset contains 58,000 Reddit comments annotated with 27 fine-grained emotion categories (Demszky et al., 2020; Fitriana and Setiawan, 2025). Similarly, datasets constructed from microblogging platforms have been employed to study emotion classification in short and informal text (Wen and Wan, 2014). Other studies have utilized news headlines, such as the dataset from the SemEval 2007 workshop (Strapparava and Mihalcea, 2007), to explore relationships between lexical semantics and emotion

Model	Developer	Parameters	Context Window(tokens)	License / Source
GPT-4.1 (OpenAI, 2025a)	OpenAI	Proprietary	1M	Closed / Commercial API
GPT-4o-mini (OpenAI, 2025b)	OpenAI	Proprietary	128K	Closed / Commercial API
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	Meta AI	70B	128K	Open (Meta Llama License)
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Meta AI	8B	128K	Open (Meta Llama License)
Mistral-NeMo-12B-Instruct (AI, 2024)	Mistral AI	12B	128K	Apache-2.0 (Open)

Table 1: Technical overview of the LLMs used in this study, including developer, parameter count, context window size, and licensing/source information.

(Danisman and Alpkocak, 2008). Although these datasets are valuable for training and benchmarking models on explicit emotional expressions, they are limited in their ability to capture more nuanced forms of emotional intelligence. Social media posts and headlines tend to be brief, highly contextual, and often include performative or fact-based language rather than implicitly conveyed, naturally expressed personal emotions (Yang et al., 2023). These limitations motivated our adoption of the UniC dataset (Du et al., 2025), which was curated from YouTube monologues (such as book and film reviews). UniC captures multimodal, non-acted, implicit, and naturally occurring emotional expressions, making it well-suited for our study. The dataset comprises 964 video clips derived from a set of source videos and curated through a multi-step pipeline involving keyword search, subtitle filtering, and manual validation. The clips are short (approximately 10 seconds each) and were independently annotated across four modalities: text, audio, silent video, and all modalities combined. Annotations include categorical labels (26 initial categories later clustered into seven emotions) as well as dimensional valence-arousal scores. The seven emotions are disappointment, disgust, confusion, neutral, contentment, joy, and surprise. In this study, we rely solely on the text modality (i.e., transcripts).

3.2 Emotion Generation Dataset

To develop a dataset suitable for evaluating emotion-aware text generation systems, we construct parallel textual variants that differ only in emotional style while preserving semantic meaning. The UniC dataset consist of 964 transcripts drawn from 18 videos, each containing spontaneous emotional cues.

3.2.1 Generate Neutralized and Emotionalized Samples

Each transcript is processed through a two-step prompting pipeline. In the first step, a neutralization prompt (Box A.1) removes explicit emotional

expressions while maintaining the original semantics and sentence structure. This ensures that emotional cues present in the source text do not influence the LLM during emotion generation, enabling a controlled comparison with the original transcript. Second, an emotion reinjection prompt (Box A.2) introduces a target emotion with minimal stylistic deviation from the neutralized version. This *neutralize* \rightarrow *reinject* procedure produces three aligned versions per sample: the original transcript, a neutralized form, and an emotionalized version. Five LLMs (see Table 1) are used for generation, producing a total of $964 \text{ samples} \times 5 \text{ models} = 4,820$ generated texts. If the original text is labeled as neutral, both the neutralization and neutral emotion re-injection stages should ideally leave the text unchanged, resulting in identical original, neutralized, and re-injected outputs.

3.2.2 Sample Selection for Human Annotation

Direct human annotation of all 4,820 generated texts would be prohibitively costly. Therefore, we selected a representative subset using a two-step sampling strategy. First, for each emotion category, transcripts were sampled from both the beginning and the end of each of the 18 videos to ensure content diversity. Second, we prioritized samples with high-quality ground-truth labels. Specifically, from a smaller manually annotated portion of the UniC dataset (61 samples labeled by three annotators), we retained only samples whose majority score exceeded 0.6, corresponding to agreement in at least 2 of the 3 annotations, resulting in 55 reliable samples. Applying both criteria resulted in a final subset of 277 high-quality samples. Across all models, this corresponds to: $277 \text{ samples} \times 5 \text{ models} = 1,385$ generated texts. Thus, for emotional text generation, we annotated a total of 1,385 samples.

The neutralization process is not the main component of the pipeline, so we annotated a representative subset of the 277 samples. Specifically, we selected 12 samples per emotion category from the 277 samples. Given 7 emotion categories, this

Model	Krippendorff’s α	Fleiss’ κ
GPT-4.1	0.2918	0.2903
LLaMA-3.3-70B	0.3023	0.3008
LLaMA-3.1-8B	0.2851	0.2835
Mistral-Nemo-12B	0.3924	0.3911
GPT-4o-mini	0.2268	0.2252

Table 2: Inter-annotator agreement (IAA) scores for emotion neutralization evaluation across five different LLMs.

resulted in: $12 \times 7 = 84$ samples. Across the 5 models, this yields a total of: $84 \times 5 = 420$ samples for neutralization annotation.

3.2.3 Human Annotation of Neutralized Texts

Human evaluation of neutralization was conducted on an 84-sample subset per LLM to assess the effectiveness of emotion neutralization. Annotators were informed that each text was an automatically neutralized version of an originally emotional sentence and were asked to identify the perceived emotion. Selecting *neutral* indicates successful neutralization, whereas choosing any other emotion suggests a failure in the neutralization process. Five annotators independently annotated each sample. To assess the reliability and consistency of the annotations, we computed two robust inter-annotator agreement (IAA) metrics: Krippendorff’s α (Krippendorff, 2018) and Fleiss’ κ (Fleiss, 1971). Unlike simple percentage agreement, both metrics correct for chance agreement and therefore provide a more conservative estimate of annotation reliability. The IAA results for all evaluated models are shown in Table 2. Overall, the agreement scores indicate *fair agreement* across models. Higher values of α and κ reflect greater consistency among annotators in identifying neutralized text, whereas lower values suggest residual emotional signals or ambiguity in the generated outputs.

3.2.4 Human Annotation for Generated Emotional Texts

Human evaluation was performed on the 277-sample subset to assess the perceived emotional accuracy and consistency of the generated outputs. Annotators are asked to select the most appropriate emotion label from the predefined set of seven categories. Five annotators independently annotated each sample. To further validate the consistency of the annotations, we again computed the two inter-annotator agreement metrics Krippendorff’s α and Fleiss’ κ . As shown in Table 3, α and κ values fall

between 0.38 and 0.45, which indicates a moderate level of agreement among annotators. These values are typical in emotion annotation tasks (Du et al., 2025), which inherently involve subjective interpretation. Notably, GPT-4.1 yields the highest agreement scores ($\alpha = 0.4497$, $\kappa = 0.4469$), suggesting that annotators slightly more consistently perceived the emotional cues in its outputs compared to the other models.

Model	Krippendorff’s α	Fleiss’ κ
GPT-4.1	0.4497	0.4469
LLaMA-3.3-70B	0.4261	0.4232
LLaMA-3.1-8B	0.4394	0.4364
Mistral-Nemo-12B	0.4451	0.4422
GPT-4o-mini	0.3841	0.3807

Table 3: Inter-annotator agreement scores for emotion re-injection across models using Krippendorff’s α and Fleiss’ κ . Higher values indicate stronger annotator consensus.

4 Metrics

To evaluate the performance of LLMs in emotion understanding (OERS, GERS), emotion generation (EGS), and internal consistency (ECS), we define four metrics based on a VAE-based emotion cost matrix. Let N denote the total number of samples, o_i the target (ground-truth) emotion for the i -th sample, \hat{o}_i the emotion predicted by the model, and $C(\cdot, \cdot)$ an emotion cost function derived from a VAE-based emotion embedding space. Unlike standard classification metrics such as F1, the VAE-based cost matrix (VCM) explicitly captures semantic similarity between emotion categories (for a more detailed discussion, see Section 4.2).

- **Original Emotion Recognition Score (OERS):** This metric evaluates a model’s ability to recognize emotions in the original data by comparing its predictions with the reference emotion labels.

$$\text{OERS} = \frac{1}{N} \sum_{i=1}^N C \left(o_i^{\text{original}}, \hat{o}_i^{\text{original}} \right) \quad (1)$$

- **Generated Emotion Recognition Score (GERS):** This metric measures how well a model recognizes emotions in generated text as perceived by human annotators, by comparing model predictions with human-labeled

Model	OERS	GERS	EGS	ECS
GPT-4.1	0.4943	0.7389	0.7248	0.7669
LLaMA-3.3-70B	0.4559	0.7406	0.7129	0.7515
LLaMA-3.1-8B	0.2925	0.5452	0.6470	0.5010
Mistral-Nemo-12B	0.3306	0.4964	0.6033	0.4970
GPT-4o-mini	0.3436	0.5818	0.5988	0.5499

Table 4: Macro-averaged F1 scores across four evaluation dimensions: UniC and Prolific Understanding Scores (OERS, GERS), Generation Score (EGS), and Consistency Score (ECS).

emotion annotations.

$$\text{GERS} = \frac{1}{N} \sum_{i=1}^N C(o_i^{\text{generated}}, \hat{o}_i^{\text{generated}}) \quad (2)$$

- **Emotion Generation Score (EGS):** This metric evaluates emotion generation quality by measuring the alignment between the reference emotion label of the source data and the emotion perceived by human annotators in the generated output.

$$\text{EGS} = \frac{1}{N} \sum_{i=1}^N C(o_i^{\text{original}}, o_i^{\text{generated}}) \quad (3)$$

- **Emotion Consistency Score (ECS):** This metric assesses the internal emotional consistency of the model by measuring whether the emotion explicitly specified during text generation is correctly identified by the same model when it is subsequently asked to infer the dominant emotion of its own generated text.

$$\text{ECS} = \frac{1}{N} \sum_{i=1}^N C(o_i^{\text{original}}, \hat{o}_i^{\text{generated}}) \quad (4)$$

We begin by evaluating model performance using the macro-averaged F1 score, computed uniformly across all emotion classes. This metric treats all misclassifications as equally severe and is commonly used for multi-class emotion recognition. Table 4 reports the macro F1 results for different LLMs across four evaluation dimensions.

Under this uniform-penalty evaluation, GPT-4.1 achieves the strongest overall performance. To better understand the source of these scores, Table 5 provides a per-emotion breakdown for GPT-4.1 across all metrics. The reason the results for the *confusion* label appear excessively high is that, during the neutralization stage, 9 of the GPT-4.1 errors

were *confusion* label. As a result, the neutralized texts generated by GPT-4.1 still contain confusion cues in 9 samples (for more details, see Table 9).

Emotion	# Samples	OERS	GERS	EGS	ECS
Disappointment	60	0.5833	0.8750	0.8833	0.9000
Disgust	22	0.4091	0.8500	0.6818	0.8636
Confusion	17	0.9412	0.9375	0.8235	1.0000
Neutral	66	0.3939	0.6438	0.8485	0.6667
Contentment	68	0.4118	0.7895	0.4706	0.6765
Joy	32	0.3750	0.6818	0.7813	0.7500
Surprise	12	0.7500	0.5909	0.8333	0.8333
Macro Average	277	0.4943	0.7379	0.7248	0.7669

Table 5: Per-emotion performance of GPT-4.1 across evaluation metrics.

Figure 1 presents the confusion matrices corresponding to the four evaluation metrics for GPT-4.1. A consistent pattern emerges across metrics: a substantial portion of errors arises from confusion between *joy* and *contentment*. While one might suggest merging these classes, a similar confusion pattern is also observed between *contentment* and *neutral*. Merging classes in response to such overlaps would therefore lead to an undesirable collapse of distinct emotional states (e.g., *joy*, *contentment*, and *neutral*), which are conceptually and functionally different. Given this continuum-like relationship among emotions, penalizing confusions between semantically adjacent classes in the same way as confusions between semantically distant emotions (e.g., *joy* and *disgust*) is arguably inappropriate. This observation motivates the need for an evaluation framework that explicitly accounts for semantic proximity among emotion classes, rather than relying on coarse class merging.

4.1 Manually Defined Cost Matrix

To address the limitations of uniform-penalty metrics, we first introduce a manually defined semantic cost matrix (De Bruyne, 2022), shown in Figure 2. This matrix assigns lower costs to misclassifications between semantically similar emotions and higher costs to confusions between semantically distant ones.

Table 4 reflects performance under the standard macro F1 setting, whereas Table 6 reports results obtained using the manual cost matrix. Unlike macro F1, this evaluation incorporates polarity-aware misclassification costs between emotion categories. While the overall ranking of models remains broadly consistent, the cost-sensitive evaluation amplifies performance differences. Stronger

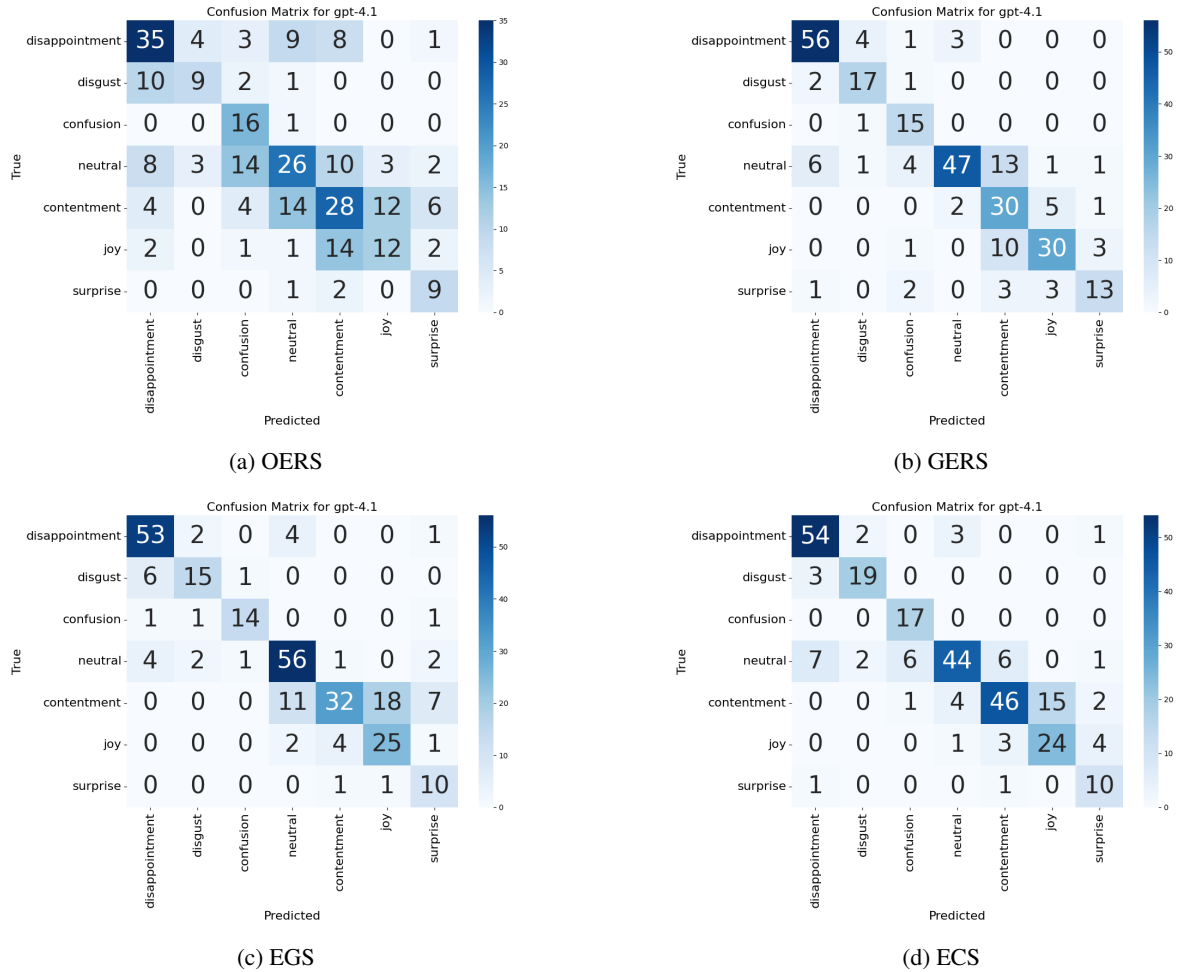


Figure 1: Confusion matrices for GPT-4.1 across the four evaluation metrics.

Model	OERS	GERS	EGS	ECS
GPT-4.1	0.7004	0.8700	0.8761	0.8809
LLaMA-3.3-70B	0.6871	0.8664	0.8556	0.8676
LLaMA-3.1-8B	0.5872	0.7774	0.8111	0.7341
Mistral-Nemo-12B	0.6258	0.7906	0.8002	0.7401
GPT-4o-mini	0.6474	0.8279	0.7858	0.7714

Table 6: Model performance under the manually defined semantic cost matrix.

models benefit from making errors within the same polarity, which are penalized less, whereas weaker models incur higher penalties due to confusion across opposite polarities. This suggests that macro F1 may obscure qualitative differences in how models reason about emotional polarity.

4.2 VAE-Based Cost Matrix

Although effective, manually designing a cost matrix is inherently subjective. To overcome this limitation, we propose an automatic, data-driven approach based on latent affective representations learned by a variational autoencoder (VAE). We

employ three-dimensional sentiment embeddings from the SentiVAE model (Hoyle et al., 2019), which encode affective semantics in a continuous latent space. Each emotion label is mapped to its corresponding latent vector, and semantic similarity between emotions is computed using cosine similarity. To obtain a misclassification cost, similarity is transformed into dissimilarity by subtracting the cosine similarity from one, followed by normalization to the $[0, 1]$ range (see Algorithm 1). The resulting VAE-based cost matrix (VCM) is visualized in Figure 3, while Table 7 reports evaluation results using this VAE-based cost matrix.

Model	OERS	GERS	EGS	ECS
GPT-4.1	0.8549	0.9535	0.9640	0.9581
LLaMA-3.3-70B	0.8494	0.9524	0.9531	0.9533
LLaMA-3.1-8B	0.7871	0.9120	0.9396	0.8998
Mistral-Nemo-12B	0.8103	0.9155	0.9305	0.9196
GPT-4o-mini	0.8265	0.9410	0.9310	0.9235

Table 7: Evaluation results using the VAE-based cost matrix.

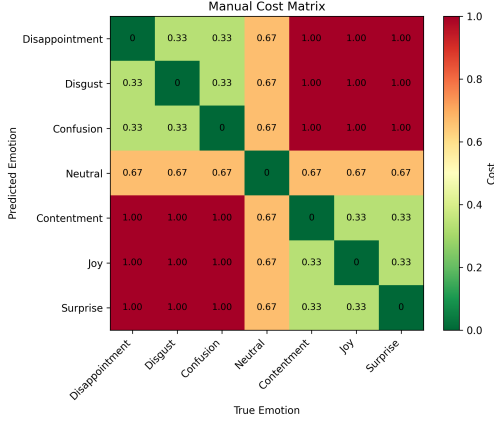


Figure 2: Manually defined semantic cost matrix for emotion classification. Darker colors indicate higher misclassification costs, while lighter colors correspond to lower misclassification cost for emotion pairs.

Algorithm 1 VAE-Based Cost Matrix Construction

Require: Emotion set $\mathcal{E} = \{e_1, \dots, e_N\}$; VAE sentiment dictionary \mathcal{D}
Ensure: Normalized cost matrix $\mathbf{C} \in [0, 1]^{N \times N}$

- 1: **for** $i = 1$ to N **do**
- 2: Retrieve latent vector $\mathbf{v}_i \in \mathbb{R}^3$ for emotion e_i
- 3: **end for**
- 4: **for** $i = 1$ to N **do**
- 5: **for** $j = 1$ to N **do**
- 6: $s_{ij} = \cos(\mathbf{v}_i, \mathbf{v}_j)$
- 7: $C_{ij} = 1 - s_{ij}$
- 8: **end for**
- 9: **end for**
- 10: Normalize \mathbf{C} to $[0, 1]$
- 11: **return** \mathbf{C}

Finally, Table 8 reports the Pearson correlation (Pearson, 1895) between the VAE-based and manually defined cost matrices. The correlations are consistently high across all metrics, indicating that the VAE-based approach closely approximates the manually designed costs. Given this near-perfect alignment, we adopt the VAE-based cost matrix for all subsequent experiments, as it provides a principled, automated, and semantically grounded alternative to manual cost specification.

Metric	OERS	GERS	EGS	ECS
Correlation	0.9985	0.9873	0.9850	0.9656
p -value ($\times 10^{-5}$)	6.90	171.55	220.14	763.27

Table 8: Pearson correlation and corresponding p -values (all reported in units of 10^{-5}) between VAE-based and manually defined cost matrices.

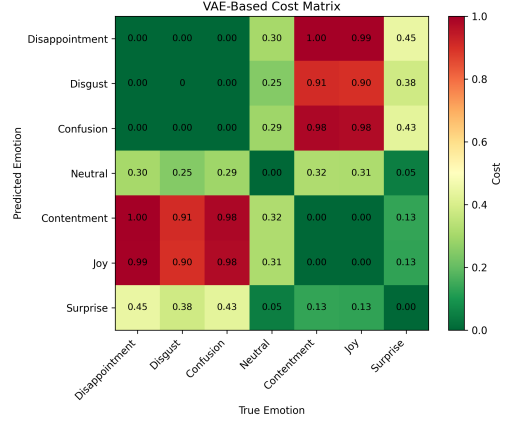


Figure 3: VAE-based cost matrix derived from latent sentiment embeddings.

5 Results and Discussion

We first present the results of the neutralization stage, which constitutes the initial component of the proposed pipeline. This is followed by the results of the emotion re-injection stage. Finally, we provide a qualitative analysis based on two representative samples.

5.1 Neutralization Results

Table 9 summarizes the performance of different models on the emotion neutralization task over 84 annotated samples. In the ideal (best-case) scenario, all outputs should be classified as *neutral*, since the models were explicitly instructed to neutralize the emotional content of the input texts. However, after conducting human evaluation, we found that a subset of the generated outputs still conveyed residual emotions and were therefore annotated as non-neutral. For F1-score computation, the ground-truth labels consist of 84 instances of the *neutral* class (i.e., the true values assume all samples should be neutral). The predicted labels correspond to the *majority emotion* assigned by human annotators to each generated neutralized text. Each sample was annotated by five annotators, and the dominant emotion among them was used as the final prediction.

Among the evaluated models, *LLaMA-3.3-70B* achieves the highest F1-score (0.95), indicating a strong ability to remove emotional cues and produce genuinely neutral outputs, with very few residual emotional mistakes. *LLaMA-3.1-8B* follows with an F1-score of 0.83, demonstrating robust performance despite its smaller model size. *GPT-4.1* and *Mistral-Nemo-12B* achieve comparable F1-

scores (0.78), suggesting similar neutralization behavior across proprietary and open-source models. Finally, *GPT-4o-mini* records the lowest F1-score (0.77), reflecting comparatively weaker neutralization performance and a higher number of non-neutral outputs. Overall, these results demonstrate that instruction-tuned open-source models, particularly larger variants, can achieve strong and competitive performance in the neutralization stage of the proposed pipeline.

5.2 Emotion Re-injection Results

Table 7 reports the evaluation results for all models using the VAE-based cost matrix on 277 samples. According to the metrics defined in Section 4, o_i^{original} denotes the original UniC emotion labels, while $\hat{o}_i^{\text{original}}$ represents the model predictions on the UniC dataset. Furthermore, $o_i^{\text{generated}}$ corresponds to the human annotations of the generated texts, and $\hat{o}_i^{\text{generated}}$ denotes the model predictions on these generated samples. Overall, larger models consistently outperform smaller ones across all four metrics, with GPT-4.1 achieving the highest scores in understanding, generation, and consistency.

To further analyze the relationship between the proposed metrics, Figure 4 presents the Pearson correlation matrix computed over average VAE-based cost matrix scores. The results reveal a strong correlation between OERS and GERS ($r = 0.96$), indicating that models that better understand emotions in the original dataset also tend to align more closely with human emotion perception in generated text. The ECS exhibits the highest correlation with OERS ($r = 0.98$), suggesting that strong emotion understanding is closely tied to internally consistent emotion generation. In contrast, the EGS exhibits comparatively lower correlations with understanding-based metrics ($r \approx 0.70$). This discrepancy strongly suggests that emotion understanding and internal consistency metrics are not sufficient proxies for evaluating an LLM’s emotion generation capability. While models may accurately recognize or internally align with emotional intent, this does not necessarily translate into a good emotional text generation model. These findings underscore the need for specialized evaluation protocols that independently assess emotion understanding and emotion generation, thereby providing a more accurate and nuanced characterization of an LLM’s emotional intelligence.

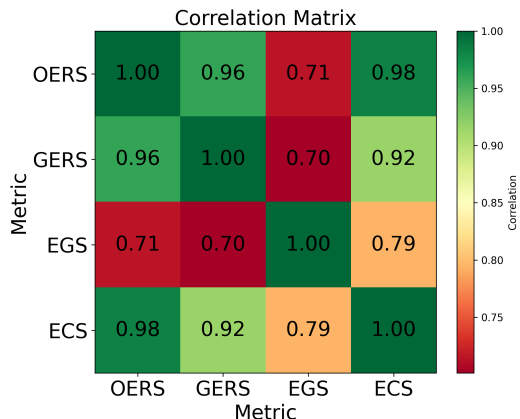


Figure 4: Pearson correlation matrix between emotion understanding, generation, and consistency metrics using average VCM scores.

5.3 Qualitative Analysis

Table 10 presents two representative examples illustrating how emotional cues introduced during enrichment influence both human annotations and GPT-4.1 emotion predictions across different annotation sources. Importantly, *UniC Pred.* refers to GPT-4.1 predictions on samples annotated in the UniC dataset, while *Prolific Pred.* denotes GPT-4.1 predictions on emotionally enriched texts whose labels originate from Prolific crowd annotations.

In the first example, the original text describes a disturbing narrative involving cannibalism and violence, yet includes a subtle positive evaluative cue (“trust me this is a good book”). Despite the dark subject matter, Prolific annotations predominantly label the emotion as *contentment*, suggesting that annotators attend to the narrator’s engaged and appreciative stance rather than the literal events described. The neutralized version removes this evaluative framing, yielding a more descriptive and emotionally flattened account. In the emotionally enriched version, affective signals are explicitly amplified through phrases such as “fascinating book,” “eager curiosity,” and “strangely satisfying.” While GPT-4.1 predictions on the generated emotional text aligned with Prolific annotations indicate *contentment*, GPT-4.1 predictions on the original text (UniC-annotated version) shift to *surprise*. The second example exhibits a clearer alignment between emotional framing and semantic content. The original text expresses a preference for “silly, light-hearted stories,” which is further reinforced in the emotionally enriched version through explicit affective language such as “gentle smile” and

Model	F1	Disappointment	Disgust	Confusion	Neutral	Contentment	Joy	Surprise
GPT-4.1	0.7826	3	5	9	54	3	3	7
LLaMA-3.3-70B	0.9500	2	2	1	76	2	0	1
LLaMA-3.1-8B	0.8333	3	2	3	60	3	7	6
Mistral-Nemo-12B	0.7826	8	4	5	54	3	3	7
GPT-4o-mini	0.7737	11	0	6	53	6	3	5

Table 9: Emotion neutralization performance of different models measured using F1-score. The per-emotion columns report the number of samples annotated with each dominant emotion by human evaluators. Since all inputs were expected to be neutral after processing, non-neutral categories represent emotional mistakes made by the models.

UniC Emotion	Prolific Emotions	Original Text	Neutral Text	Emotional Text	UniC Pred.	Prolific Pred.
contentment	{joy: 1, contentment: 2, surprise: 1, neutral: 1}	about cannibalism trust me this is a good book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a fascinating book we follow our female main character who is a literal psychopath and we follow her with a sense of eager curiosity as she confidently goes after her ex-lovers, kills them with chilling precision , and then eats them. Each page unfolds her unnerving world in a strangely satisfying way .	surprise	contentment
contentment	{neutral: 5}	noticed that I personally have a preference for like the more like silly light-hearted stories as opposed to beach read which falls more into like the sweeping emotional side of things	noticed that I personally have a preference for the more silly, light-hearted stories as opposed to beach read which falls more into the sweeping emotional side of things	I have a preference for stories that are more light-hearted , the kind that bring a gentle smile or a comforting warmth , as opposed to 'Beach Read,' which falls more into the sweeping emotional side of things .	contentment	contentment

Table 10: An example annotation from the dataset illustrating the original, neutralized, and emotionally enriched texts, along with their corresponding emotion labels and model predictions. The yellow highlights indicate emotionally salient cues and are shown solely for clarification purposes in these two samples. For text generation and emotion prediction, we used GPT-4.1.

“comforting warmth”. In this case, GPT-4.1 predictions for both UniC and Prolific annotated samples consistently yield *contentment*. This consistency suggests that when emotional cues are explicit and thematically congruent, model predictions remain stable.

6 Conclusion

In this work, we proposed a semantic, VAE-based evaluation framework for assessing LLMs’ emotional intelligence across emotion understanding, generation, and internal consistency. Through a comprehensive analysis of the UniC dataset and a model-generated dataset annotated by human evaluators, we showed that strong emotion understanding does not necessarily imply accurate emotion generation. Our results, supported by a correlation analysis, highlight emotion generation as a distinct and more challenging capability, motivating the need for dedicated evaluation protocols beyond traditional understanding-based metrics.

7 Future Work

In future work, we plan to include a sensitivity analysis and a discussion on how slight variations

in neutralization and emotion reinjection prompts could impact the final scores. Moreover, the current framework depends on human annotators to compute GERS and EGS, which may limit its scalability to larger datasets. To address this, future research could explore automated evaluators that reduce annotation costs while maintaining reliable assessment quality. Additionally, other embedding methods, such as transformers or word vectors, should be tested for constructing the cost matrix, as our current experiments only used VAE-based embedding vectors for each emotion label.

Acknowledgments

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (FAIR) (174K02325).

Limitations

Our evaluation is limited to a fixed emotion taxonomy and a single cost matrix, which may not fully capture culture-specific or context-dependent emotional nuances. Additionally, human annotations are subject to individual perception variability,

which can introduce noise into the annotation process.

References

- Mistral AI. 2024. Mistral nemo. <https://mistral.ai/news/mistral-nemo/>.
- Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 convention communication, interaction and social intelligence*, pages 53–59. Aberdeen, Scotland.
- Luna De Bruyne. 2022. *Feeling EmotioNL : automatically detecting emotions in Dutch texts*. Ph.D. thesis, Ghent University.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025. Unic: a dataset for emotion analysis of videos with multimodal and unimodal labels: Q. du et al. *Language Resources and Evaluation*, pages 1–36.
- Frizka Fitriana and Hendrik Setiawan. 2025. Performance analysis of svm in emotion classification: A comparative study of tf-idf and countvectorizer. *Journal of Embedded Systems, Security and Intelligent Systems*, pages 133–145.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Ryan Cotterell, and Isabelle Augenstein. 2019. Combining sentiment lexica with a multi-view variational autoencoder. 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 635–640.
- He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.
- Shin-nosuke Ishikawa and Atsushi Yoshino. 2025. Ai with emotions: Exploring emotional expressions in large language models. *arXiv preprint arXiv:2504.14706*.
- Sadegh Jafari, Els Lefever, and Véronique Hoste. 2025. Embedding analogies for evaluating emotion in llm-generated utterances. In *28th European Conference on Artificial Intelligence (ECAI 2025)-BEHAIV workshop*.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. 'only chatgpt gets me': An empirical analysis of gpt versus other large language models for emotion detection in text. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2603–2611.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.
- Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.
- Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Chenyang Zhao, Zhongwei Wan, Chaofan Tao, Wendong Xu, and 1 others. 2025. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *arXiv preprint arXiv:2509.07403*.
- OpenAI. 2025a. Introducing gpt-4.1. <https://platform.openai.com/docs/models/gpt-4.1>. Accessed: 2025-12-12.
- OpenAI. 2025b. Introducing gpt-4o-mini. <https://platform.openai.com/docs/models/gpt-4o-mini>. Accessed: 2025-12-12.
- Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.
- Pinaki Raj. 2024. A literature review on emotional intelligence of large language models (llms). *International Journal of Advanced Research in Computer Science*, 15(4).
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.
- Haybat Abdul Samad. 2014. Emotional intelligence the theory and measurement of eq. *European Scientific Journal*.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Shiyang Wen and Xiaojun Wan. 2014. Emotion classification in microblog texts using class sequential rules. In *Proceedings of the AAAI conference on artificial intelligence*.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, and 1 others. 2025. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

A Prompts

Box A.1: Emotion Neutralization Prompt

Objective. This prompt removes emotional expressions from a text while preserving semantic content, structure, and linguistic properties.

Prompt Template.

Your task is to neutralize the text by removing emotional expressions.

The text is a transcription of a video and may contain emotional cues.

The output text must:

- be emotionally neutral,
- remain in the same language,
- preserve the original format, style, tone, and context,
- and differ from the input as little as possible.

Please neutralize the following text:

{text}

The original emotion of the text is: {emotion}.

Ensure that all emotional expressions are removed.

Return the result in the following JSON format:

```
{"neutral_text": "The neutralized text"}
```

Box A.2: Emotion Reinjection Prompt

Objective. This prompt reintroduces emotional expressions into a neutral text, targeting a specific emotion while maintaining semantic fidelity.

Prompt Template.

Your task is to make the text more emotional by adding emotional expressions.

The text is a transcription of a video.

The output text must:

- remain in the same language,
- preserve the original format, tone, and context,
- and differ from the input as little as possible.

Do not explicitly mention the target emotion in the text.

Please add emotional expressions to the following text:

{text}

The current emotion of the text is neutral. The target emotion of the text should be: {emotion}.

Return the result in the following JSON format:

```
{"emotional_text": "The emotionalized text"}
```

News Credibility Assessment by LLMs and Humans: Implications for Political Bias

Pia Wenzel Neves^{1,2}, Charlott Jakob^{1,2}, Vera Schmitt^{1,2}

¹Quality & Usability Lab, Technische Universität, Berlin,

²German Research Center for Artificial Intelligence (DFKI), Berlin,

Correspondence: c.jakob@tu-berlin.de

Abstract

In an era of rapid misinformation spread, LLMs have emerged as tools for assessing news credibility at scale. However, the assessments are influenced by social and cultural biases. Studies investigating political bias, compare model credibility ratings with expert credibility ratings. Comparing LLMs to the perceptions of political camps extends this approach to detecting similarities in their biases. We compare LLM-generated credibility and bias ratings of news outlets with expert assessments and stratified political opinions collected through surveys. We analyse three models (Llama 3.3 70B, Mixtral 8x7B, and GPT-OSS 120B) across 47 news outlets from two countries (U.S. and Germany). We found that models demonstrated consistently high alignment with expert ratings, while showing weaker and more variable alignment with public opinions. For US-American news outlets all models showed stronger alignment with center-left perceptions, while for German news outlets the alignment is more diverse.

1 Introduction

In an era of information abundance, the ability to assess news credibility and identify media bias has become a critical competency for informed citizenship (Haider and Sundin, 2022). Yet the volume of content produced daily overwhelms human fact-checking capacity (Quelle and Bovet, 2024). This challenge has grown as misinformation spreads rapidly on digital platforms, creating what some scholars call an epistemic crisis in democracies (Zhang et al., 2025). Simultaneously, large language models (LLMs) have emerged as powerful tools capable of performing complex evaluative tasks at scale, raising the question of whether these systems could assist in addressing the credibility assessment bottleneck (Augenstein et al., 2024). Research demonstrates that LLMs possess technical capabilities for media bias detection and

news credibility assessment approaching the performance level of specialized models (Maab et al., 2024). Models can identify journalistic credibility signals comparable to human fact-checkers through zero-shot weak supervision approaches (Leite et al., 2025). However, these technical capabilities exist alongside well-documented systematic political biases in LLMs. LLMs exhibit left-of-center political preferences across multiple political orientation tests (Rozado, 2024), with these biases manifesting not only in the generated content but also its stylistic dimensions such as lexical polarity and framing (Bang et al., 2024). Furthermore, LLMs mostly demonstrate language-dependent variations in their assessments (Sharma et al., 2025). These findings raise fundamental questions about whose perspectives these systems represent when evaluating news sources. Previous studies analyse a variety of news outlets by comparing LLMs' credibility ratings with expert credibility ratings (Yang and Menczer, 2025; Loru et al., 2025). However, comparing them to perceptions of ideological groups is crucial for drawing similarities between LLMs and group biases. This paper addresses this problem by comparing LLM-generated credibility and bias ratings of news outlets against both expert assessments and politically stratified public opinion from survey data. Ideological viewpoints and media agenda are shaped by national environments and therefore need to be analysed separately (Vu et al., 2019). We investigate and compare the media environments of the U.S. and Germany. We evaluate Llama-3.3-70B-Instruct (Llama), Mixtral-8x7B-Instruct-v0.1 (Mixtral), and GPT-oss-120b (GPT) across 47 news outlets. Using 22,560 systematically varied prompts, we investigate how model selection, prompt language, and prompting style affect rating consistency and accuracy. Furthermore we evaluate to what extent model-generated ratings of news outlets align with human ratings of news outlets across different political

camps and national contexts. Thus the following research questions are formulated:

1. How do factors such as model selection, prompt language, and prompting style affect the consistency and accuracy of LLMs' ratings of the political bias and credibility of news outlets? (RQ1)
2. To what extent do LLM-generated ratings of the political bias and credibility of news outlets align with public perceptions from different political camps and expert opinion for different national media ecosystems? (RQ2)

2 Related Work

The existing literature reveals a critical disconnect between three well-established but largely isolated research streams. The first research stream investigates political and cultural biases in LLMs, and shows that LLMs predominantly exhibit left-leaning orientations (Rozado, 2024; Peng et al., 2025; Jakob et al., 2025), Western cultural preferences (Tao et al., 2024) and output variations depending on the prompt language (Alkhamissi et al., 2024; Sharma et al., 2025). The second stream explores LLMs' technical capabilities for media bias detection and news credibility assessment (Maab et al., 2024; Leite et al., 2025). The third stream documents a substantial divergence between LLM judgments and human values across different demographic groups (Hadar-Shoval et al., 2024; Santurkar et al., 2023), with standard alignment methods systematically underrepresenting minority perspectives (Chakraborty et al., 2024).

Further complicating this picture, prompting variations including language choice, question format, and template design also dramatically affect LLM outputs (Errica et al., 2024; Röttger et al., 2024). However, these research streams have remained largely compartmentalised. While Yang and Menczer (2025) and Loru et al. (2025) examined LLM news credibility ratings at scale, they focused primarily on English-language contexts and did not systematically compare ratings across different political camps within human populations. Rotaru et al. (2024) provided initial evidence that LLMs favor left-leaning outlets, but their study was limited to a small number of sources and did not examine how prompt language or model selection affects these patterns.

Three influential variables emerge that should be taken into account when evaluating political bias:

(1) model selection, given that different models exhibit varying degrees of political bias and alignment approaches (Peng et al., 2025; Aldahoul et al., 2025); (2) prompt language, given evidence that query language determines information retrieval and cultural alignment (Sharma et al., 2025; Wang et al., 2025); and (3) prompting style, given extreme sensitivity to format and template variations (Errica et al., 2024; Zhuo et al., 2024).

Furthermore, existing cross-national research on LLMs' political bias (Motoki et al., 2024; Batzner et al., 2024; Rettenberger et al., 2024), has focused on general political positioning rather than specifically examining news credibility assessment across different national media ecosystems. The literature also lacks systematic comparison of LLM ratings with politically diverse human populations as most studies either compare against expert consensus (Yang and Menczer, 2025) or examine aggregate human preferences without political stratification.

3 Methods

3.1 Dataset

First, we built a dataset which consists of news-outlets paired with human ratings of the news outlets' bias and credibility. We wanted to examine two types of human assessment: expert opinion and sets of public opinions, in order to be able to compare the model correlation with experts to the model correlations with different public opinion groups.

Therefore surveys had to not only contain an average of how people rate news outlets, but surveys where people from all major political camps were able to anonymously state their opinion on news outlets from left, center and right. The survey had to not only indicate the proportion of respondents from each political camp, but also specifically provide the results for each political camp.

The expert opinion component was chosen to evaluate which model aligns most with expert ratings. The expert source had to cover a broad spectrum of German and US-American news sources, both in terms of political orientation and credibility. That was a condition for being able to detect different alignments and for comparisons between countries.

As the dataset was constructed using four different sources the terminology for the concept of credibility differs. Credibility is used as an umbrella term for factual reporting (mbfc), trustworthiness

(YouGov), and quality (Medienkompass).

For the news outlets from the USA two surveys were selected, one survey, AllSides (AllSides Staff, 2024), covered political bias of news outlets from the USA and the other survey, YouGov (YouGov, 2025c), evaluated trustworthiness of news outlets from the USA. Both surveys distinguished between three political camps, namely Democrats, Independents and Republicans.

For Germany the survey Medienkompass (Medienkompass.org, 2025c) was selected which covered both political bias and quality of German news outlets and distinguished between seven different political camps, namely "Liberal-left", "Conservative-left", "Liberal-right", "Conservative-right", "High agreement with mainstream media", "Critically-distanced towards mainstream media" and "Rejecting mainstream media".

The fact-checking website Media Bias/Fact Check (mbfc) (Media Bias/Fact Check, 2025a), where experts rate the political bias and factual reporting of news outlets, was selected to serve as the expert opinion for both countries. Mbfc offers the largest dataset covering biased and low factual news sources (Weld et al., 2021), which was an important criteria as ratings for high and low credibility news outlets and right, center and left news outlets from both Germany and the USA needed to be present in the expert opinion to being able to detect different alignments and compare between countries.

The final dataset was constructed in the following way: For the USA we first took all news outlets from the YouGov survey as a basis, because that was the data most difficult to find. The YouGov survey contains 52 news outlets. Then we searched for the names of the news outlets that were present in the YouGov survey and checked which is the most recent AllSides survey where the news outlet was evaluated. After this step 27 news outlets remained. Then we searched Media Bias / Fact Check for the news outlets so that each news outlet has an expert opinion. We were able to find all of the remaining news outlets on mbfc and so the amount of news outlets stayed at 27. For Germany we checked mbfc for available data for the news outlets present in the Medienkompass survey. For the 40 news outlets in the Medienkompass survey we were able to find expert ratings from mbfc for 20 of these news outlets. In Figure 1 the dataset creation process is visualized. The process is to be read from left to right, showing which data source

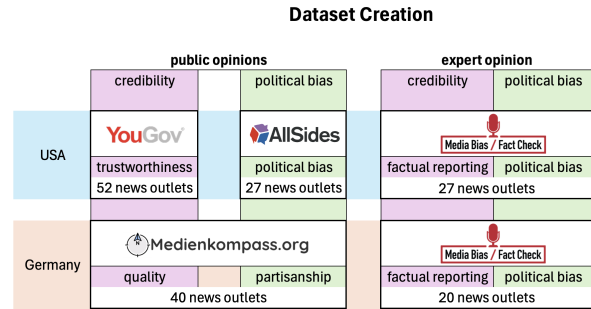


Figure 1: Overview of the dataset creation process, showing the survey sources for the public opinion on the left side and the expert source on the right and how the final number of newsoutlets was reached

was processed after another and how many news outlets remained after each step. In total 47 news outlets were analyzed of which 20 are from Germany and 27 from the USA.

In Appendix A a list of the news outlet names is provided. The dates of the ratings range from 2022 to 2025 for US-American news outlets and from 2019 to 2025 for German news outlets. Research suggests that the bias and factualness of a news source are unlikely to change over time (Weld et al., 2021; Bozarth et al., 2020), so the ranges are acceptable. In Appendix B detailed background information on the used data sources can be found.

In Table 1 and 2 you can see the political bias and credibility label distributions of the selected news outlets based on the expert ratings from mbfc. News outlets with left (47%), center (15%) and right (38%) bias as well as with low (26%), medium (32%) and high (43%) quality are represented in the final dataset. The distributions split by country can be found in Appendix C as well as a Chi-square test in Appendix D which showed that there are no significant bias-factual reporting label distribution differences between Germany and the USA.

3.2 Model selection

Three models were selected: Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024) and GPT-oss-120b (OpenAI et al., 2025), that were later prompted in order to generate ratings of the political bias and credibility of news outlets. These models were chosen to represent both state-of-the-art capabilities and widespread adoption in both research and practical applications, while ensuring geographic diversity with representation from both the United

Label	Ct.	Group	Ct.	%
Left	4	Left	22	47
Left-Center	18			
Least Biased	7	Center	7	15
Right-Center	11	Right	18	38
Right	6			
Extreme Right	1			

Table 1: **Label distribution by political bias** of the news outlets in the final dataset based on expert ratings.

Label	Ct.	Group	Ct.	%
Very High	1	High	20	43
High	19			
Mostly Factual	15	Medium	15	32
Mixed	10	Low	12	26
Low	1			
Very Low	1			

Table 2: **Label distribution by factual reporting** of the news outlets in the final dataset based on expert ratings.

States and Europe. All three models are available as open-weight models, facilitating reproducibility and broad accessibility for the research community. The models were prompted via API calls using the platform “IONOS AI Model Hub” IONOS (2025). Version control was managed through GitHub to ensure reproducibility and transparency of the experiment pipeline (Sandve et al., 2013; Danchev, 2022). All code used for generating the results can be found in this GitHub repository https://github.com/XplaiNLP/news_credibility.

3.3 Prompt Design

For the prompt design the LLMs were queried about one parameter, either political bias or credibility, per prompt. This ensures that the model evaluates each construct independently, without being influenced by the presence of another evaluative dimension (Stureborg et al., 2024). The surveys completed by humans are based on ordinal scales but with varying constructs and number of scale points. Thus, for comparability we use the survey specific scale and adjust the prompts accordingly. An overview of the different scales can be found in Appendix E. We use two different prompts so that prompt specific influence can be reduced. The “rate” variation is just a simple and intuitive phrasing: “Rate the news outlet ‘{newsoutlet}’ from {country} regarding its political bias on the following scale:”. The “opinion” variation is based on Faulborn et al. (2025), which discovered

that different phrasings of an evaluation task can produce more or less biased answers. From that paper a prompt version was selected that showed low bias and sounded natural instead of one of the versions that created particularly biased answers. This is the “opinion” variation of the task: “Please give me your opinion on the news outlet ‘{newsoutlet}’ from {country} regarding its political bias on the following scale:”. Also every prompt was translated so that there was always a German and an English version of each prompt. Each prompt was repeated ten times to systematically assess the consistency and confidence of the model’s responses across identical queries (Wang et al., 2024). The model’s temperature was set to 0 to further ensure consistency. To avoid introducing framing bias, no role instruction (e.g., “You are an expert/journalist”) was included in the prompt, ensuring that the model’s response reflected its uninfluenced baseline perspective (Kamruzzaman and Kim, 2025). We included specifications on the output format in the prompt. In Listing 1 an example of a final prompt template can be seen. An example of a json instance from the output file can be seen in Appendix F. An overview of all prompt template categories is displayed in Appendix G as well as the full text version of all prompts in Appendix M.

```

1  return [
2      {
3          "role": "user",
4          "content": f"""Rate the news
                    outlet '{newsoutlet}' from
                    {country} regarding its
                    political bias on the
                    following scale:
5      - left
6      - lean left
7      - center
8      - lean right
9      - right
10
11 Return ONLY this JSON (no extra
    text):
12 {{ "reasoning_political_bias":
    "<short reasoning>",
    "label_political_bias":
    "<EXACTLY one of: left, lean
    left, center, lean right,
    right>" }}
13 ""
14     }
15 ]

```

Listing 1: Example of prompt template (evaluated parameter: political bias, language: English, style: rate, scale: AllSides survey, country: USA),label=lst:prompt-template-political-bias

3.4 Evaluation Methods

3.4.1 Consistency and Accuracy Analysis

In order to examine potential effects of prompt style and prompt language on the LLMs' responses a consistency analysis and an accuracy analysis are performed. For the consistency analysis an agreement rate is calculated for each item as the proportion of responses matching the majority vote (Hallgren, 2012). Specifically, for each set of repeated model queries, the most frequent response is identified, and the agreement rate represents the proportion of total responses that match the majority vote. Higher agreement rate values indicate greater response consistency.

Non-parametric analyses were applied because the distribution of the data is unknown or not necessarily normally distributed. While non-parametric tests are slightly less powerful than their parametric counterparts, they are statistically sound also when the underlying distribution is unknown (Dror et al., 2018). Another reason why non-parametric analysis methods were chosen is because the scales in the dataset are ordinal in nature, consisting of named and ranked categories for which equal distances between scale points cannot be assumed (Stevens, 1946). The following three non-parametric tests were selected. The Kruskal-Wallis test was used to determine whether any differences among models exist based on the test statistic H , which is calculated using the sum of ranks within each group, quantifying the extent to which these ranks differ among the groups (Tomczak and Tomczak-Łukaszewska, 2014; Chicco et al., 2025). It is commonly regarded as an extension of the Mann-Whitney U test, which is restricted to comparisons between two groups. The Mann-Whitney U test was used to determine which specific pairs of models differ. It evaluates the ranks of the data, examining whether one group systematically exhibits higher or lower ranks than the other, using the U statistic which is calculated based on the sample sizes of the two groups and the sum of ranks in the first group (Chicco et al., 2025; Dror et al., 2020). The Rank-Biserial effect size measure was used to quantify the magnitude of the difference between two models.

The rank-biserial correlation represents the difference between the proportion of favorable comparisons (where values from one group exceed those from the other) and unfavorable comparisons. The test statistic is calculated using the U

statistic and the sample sizes of the two groups (Tomczak and Tomczak-Łukaszewska, 2014; Cureton, 1956; Glass, 1965). A similar approach has been demonstrated in a recent comprehensive guide where Kruskal-Wallis followed by Mann-Whitney U for post-hoc comparisons across multiple medical datasets were applied to test for significant differences (Chicco et al., 2025). Evaluating model consistency is important for the evaluation of the reliability of LLMs which was demonstrated by the analysis of Lee et al. (2024).

However, a model can show high consistency while giving consistently wrong answer. Thus we also conducted an **accuracy analysis** examining the effect of language and prompt style on accuracy. These comparisons were performed exclusively on expert opinion ratings, as these represent the gold standard. For each model and parameter (bias and credibility), we compared the distribution of absolute errors between conditions using the Mann-Whitney U test in combination with the rank-biserial correlation as an effect size measure. This approach parallels the methodology used in the consistency analysis.

3.4.2 Alignment Analysis

To evaluate the model alignment with public and expert opinion we employed Spearman's rank correlation coefficient (Spearman, 1904; Zar, 1972), a non-parametric measure of monotonic association. Spearman correlation was selected as it is appropriate for ordinal data and does not assume linearity or normal distribution, making it well-suited for analyzing ratings on discrete scales (Myers and Sirois, 2004).

For each model the model's predictions were averaged across all prompt and language variations for each news outlet, then these model-specific news outlet ratings were correlated with the human ratings from experts and political camps. Because each model's predictions were compared against multiple political camp columns, Bonferroni correction was applied to control for family-wise error rate across multiple comparisons (Bonferroni, 1936; Dunn, 1961). The significance threshold α was adjusted for each set of comparisons to the amount of political camps that model values were compared to. The exact α values are stated in the tables 4, 5, 6 and 7. This conservative correction reduces the probability of Type I errors when conducting multiple hypothesis tests on the same dataset (Armstrong, 2014).

4 Results

Before conducting the main analysis a preliminary analysis was conducted exploring success rates and prior model familiarity with the dataset. The greatest influence on the success rates (the proportion of valid labels generated relative to the total number of queries) was the choice of model, whereas dataset, parameter, language, and prompt style had less influence. Llama showed 100% success rate without any complications. For Mixtral the average initial success rate was 98.9% and most null labels resulted from not adhering to the given label scale. For GPT the average initial success rate was 91.3% and all null labels were due to the model refusing to answer the prompt, including both capability-based refusals (e.g. knowledge gaps) and safety-based refusals (e.g. guardrails). For Mixtral non-conforming or missing labels could all be resolved. For GPT 68 null labels remained. As each prompt template was repeated 10 times, cases where at least one valid label was generated for a specific news outlet were not counted as a refusal. Complete refusal was observed in 5 cases. A detailed overview of the success rates, null labels and refusals can be found in Appendix H.

LLMs could be aware of expert ratings because they may have been present in their training data. Investigating this possibility, we systematically searched for mentions of our data sources in the generated outputs. For GPT there were 1,8% of reasonings that contained at least one of the dataset names (Allsides 45 times and mbfc 88 times). Other dataset names were found within those 78 reasonings for which we then also specifically searched in the outputs: Deutscher Presserat (1), Correctiv(4), Reporters without Borders(1), FactCheck.org(2), Ad Fontes Media(3), Snopes(7), PolitiFact(1) and Pew Research Center(1).

For Llama 0,04% of reasonings contained at least one of the dataset names (Allsides 1 time and mbfc 3 times). NewsGuard (6) and Pew Research Center (4) were other dataset names which were found. For Mixtral there were no reasonings in which dataset names were included.

4.1 Consistency and Accuracy Analysis (RQ1)

The results of the consistency analysis show that Mixtral achieved the highest consistency (0.9789), followed by Llama (0.9538), and GPT showing notably lower consistency (0.8715) (see Table 3).

The Kruskal-Wallis test confirmed significant

Model	Consistency
Mixtral-8x7B-Instruct-v0.1	0.9789
Llama-3.3-70B-Instruct	0.9538
GPT-oss-120b	0.8715

Table 3: Model overall consistency scores.

differences between models across all settings. Pairwise comparisons revealed that Mixtral and Llama performed comparably, showing no significant differences in most scenarios except for one scenario. In contrast, GPT consistently underperformed relative to both Mixtral and Llama, with significant differences ranging from small to large effect sizes across all comparisons. Prompt language effects, by comparison, were limited and inconsistent. These effects were model-specific and lacked a consistent pattern. For instance, Llama performed more consistent when using German prompts, while GPT and Mixtral were more consistent with English prompts with effect sizes ranging from negligible to small. Prompting style had the least influence on consistency, with significant differences appearing in only two scenarios, both favoring the "rate" style for Llama and Mixtral, though with negligible effect sizes.

The influence of prompt language and prompt style on accuracy was limited. For the prompt language one significant effect was observed. GPT was returning significantly more accurate results when the prompt was written in English with a small effect size, when rating news outlets on their factual reporting. For the prompt style no significant effects were observed. In the Appendix I you can find the detailed results of the statistical tests. In summary, the results **addressing research question RQ1** reveal that model selection is the primary determinant of rating consistency, while prompt language and prompting style have minimal influence on both consistency and accuracy.

4.2 Alignment Analysis (RQ2)

The tables 4, 5, 6 and 7 describe how model ratings correlate with expert opinion and public opinions as well as how expert opinion correlates with public opinion. For bias ratings of US news outlets (Table 4), models exhibited statistically significant alignment with all political camps, though alignment was stronger with Democrats and Independents ($p = 0.74-0.76$) than with Republicans, while expert alignment ($p = 0.83-0.85$) exceeded alignment with any political camp. Democrats exhibited

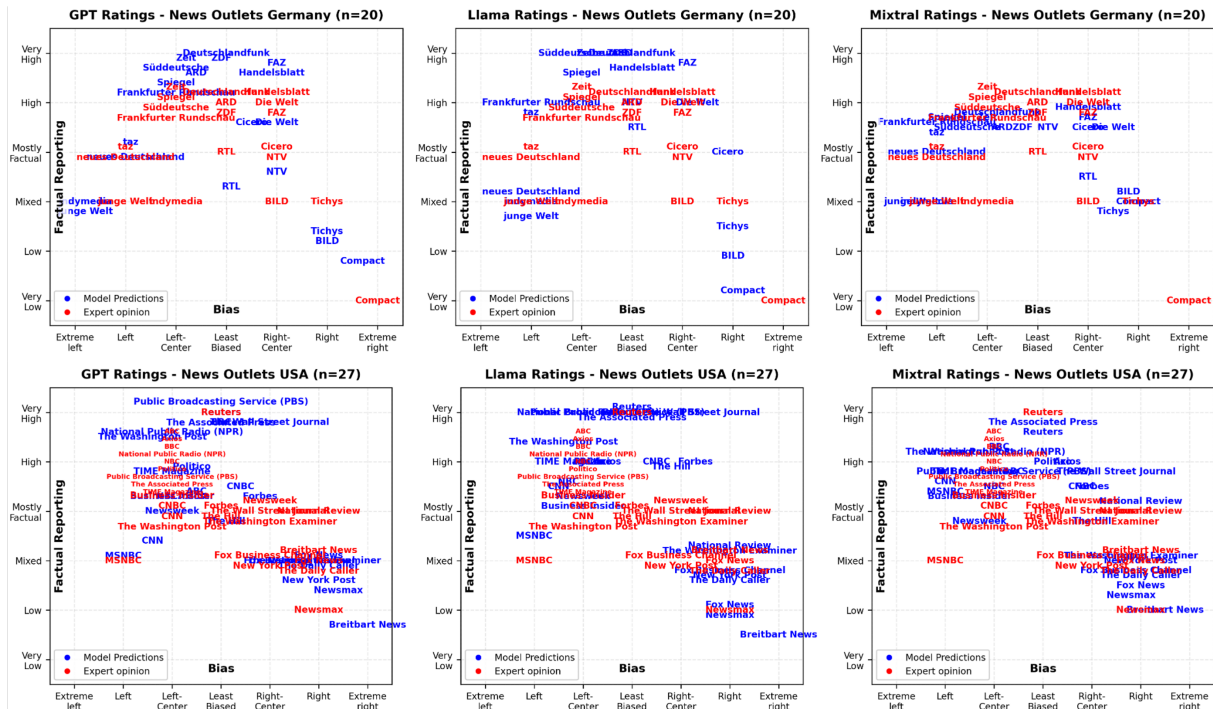


Figure 2: Model ratings vs. Expert ratings of news outlet. Model ratings are represented in blue, while human expert ratings are represented in red. Per model there is a separate chart for Germany and the USA.

the strongest correlation with expert opinions ($p = 0.86$), closely followed by Independents ($p = 0.85$), at a level comparable to the alignment between model ratings and expert ratings. For credibility ratings of US outlets (Table 5), alignment with public opinion was notably weak ($p = 0.23-0.44$), with the strongest correlations observed for Democrats ($p = 0.36-0.44$), no significant correlation with Independents, and a negative correlation between Llama and Republicans ($p = -0.24$), whereas expert alignment ($p = 0.80-0.84$) remained substantially higher. Independents exhibited the strongest correlation with expert opinions, closely followed by Democrats ($p = 0.77$ and $p = 0.72$) which is slightly lower than the alignment between model ratings and expert ratings ($p = 0.80-0.84$). There was no significant correlation between Republicans and Experts ($p = -0.47$).

For German media all model-generated bias ratings of news outlets (Table 6) show very high alignment with all political camps ($p = 0.89-0.97$) with minimal variation between ideological groups. Expert alignment ($p = 0.93-0.97$) only marginally exceeds public alignment. People agreeing with and being critical of mainstream media exhibited the strongest correlation with expert opinions ($p = 0.938$), closely followed by conservative left people ($p = 0.936$) and liberal left people ($p = 0.933$)

which is close to the alignment between model ratings and expert ratings ($p = 0.93-0.97$).

Model-generated credibility ratings (Table 7) demonstrated weak alignment with public perceptions, with fewer significant correlations across models (GPT: conservative left, Llama: critical towards mainstream media, Mixtral: five of seven camps). Expert alignment ($p = 0.86-0.93$) remained substantially higher than alignment with public opinion. People agreeing with mainstream media exhibited the strongest correlation with expert opinions ($p = 0.86$), closely followed by liberal left people ($p = 0.83$) which is slightly lower than the alignment between model ratings and expert ratings ($p = 0.86-0.93$). The other significant correlations for conservative left, liberal right people and people critical of mainstream media ranged from $p = 0.62$ to $p = 0.69$. There was no significant correlation between conservative right people and people who reject mainstream media and Experts. In Appendix K and L the proximity of expert opinion, public opinion and model opinion to each other is displayed.

A visualization of the model ratings in contrast to the human expert ratings can be found in Figure 2. Each panel plots news outlets on two dimensions: political bias on the x-axis, ranging from "Extreme left" to "Extreme right" and factual reporting on the

	Exp.	Dem.	Ind.	Rep.
Mixtral	0.826*	0.748*	0.747*	0.681*
Llama	0.851*	0.760*	0.740*	0.704*
GPT	0.837*	0.724*	0.760*	0.665*
Experts	–	0.858*	0.848*	0.752*

Table 4: **USA Bias Ratings:** Spearman correlations between model ratings and MBFC expert ratings vs. human ratings from political camps (Democrats, Independents, Republicans). Models vs. Experts: Bonferroni correction: $\alpha = 0.05/1 = 0.0500$; vs. political camps: Bonferroni correction: $\alpha = 0.05/3 = 0.0167$. The "Experts" row shows correlations between MBFC expert ratings and AllSides ratings from survey participants split by political camp. * indicates statistical significance.

	Exp.	Dem.	Ind.	Rep.	Total
Mixtral	0.841*	0.427*	0.081	-0.043	0.233*
Llama	0.806*	0.360*	-0.050	-0.243*	0.083
GPT	0.802*	0.443*	0.139	-0.013	0.263*
Experts	–	0.718*	0.772*	-0.468	0.702*

Table 5: **USA Credibility Ratings:** Spearman correlations between model ratings and MBFC expert ratings vs. human ratings from political camps (Democrats, Independents, Republicans). Models vs. Experts: Bonferroni correction: $\alpha = 0.05/1 = 0.0500$; vs. political camps: Bonferroni correction: $\alpha = 0.05/4 = 0.0125$. The "Experts" row shows correlations between MBFC expert ratings and AllSides ratings from survey participants split by political camp. * indicates statistical significance.

y-axis, spanning from "Very Low" to "Very High". Model ratings are represented by the blue news outlet names, while human expert ratings are represented by the red news outlet names. For every model there is one panel for the 20 news outlets from Germany and one panel for the 27 news outlets from the USA. Based on the visualizations we observe that model predictions exhibit greater dispersion across both the bias and factual reporting dimensions compared to expert ratings. In Appendix J a simplified display of the data points is provided for better understanding.

In summary, the results **addressing research question RQ2** reveal that LLM-generated ratings align significantly stronger with expert opinion than with political camps across both countries, particularly for credibility assessments. Models show stronger alignment with political camps when rating bias than credibility for both countries. Germany exhibits greater consensus across political

camps whereas the USA shows more polarization, with Republicans diverging most from model and expert assessments.

5 Discussion

The strong alignment between LLM ratings and expert assessments (Spearman correlations: 0.80-0.97) could indicate LLMs being strong evaluators, but must be taken with caution, as we found that two of three models saw the expert ratings during training. We don't know whether high correlation with expert assessment stems from a match to a single dataset present in training data or actually reflects LLM's evaluation based on various sources.

Our results indicate that expert, public, and model opinions on the bias parameter largely converged, thereby limiting the potential to observe meaningful patterns.

Comparing credibility perceptions of political camps and experts, we observe a higher correlation between expert and left-leaning camps compared to experts and right-leaning camps. LLMs evaluations mirror this relationship by showing higher correlations with the left than the right. Thus, in this experiment setting, one could argue that predominantly left-leaning political bias in LLMs does not arise from the model's stance but from its proximity to supposed expert opinions. The only exception to this observation was Llama evaluating the credibility of German news outlets where Llama showed stronger correlations with liberal right than with liberal left and conservative left people. For the USA however, Llama clearly aligned stronger with the Democrats even showing a negative correlation with Republicans. Models in general showed stronger alignment with left political camps in the USA than with left political camps from Germany, potentially indicating that U.S.-centric definitions of left-right ideology were applied when evaluating German news outlets. Correlation values of political camps from Germany cluster more closely, whereas those for the United States are more dispersed. This may reflect greater consensus among political camps in Germany compared to stronger polarization in the United States.

6 Conclusion

This work investigated how LLMs perceive and rate the credibility and political bias of news outlets compared to human assessments, examining three state-of-the-art models across 47 news outlets

	Experts	lib. left	cons. left	lib. right	cons. right	agree	critical	reject	avg.
Mixtral	0.933*	0.920*	0.911*	0.938*	0.893*	0.927*	0.928*	0.920*	0.927*
Llama	0.958*	0.966*	0.949*	0.956*	0.922*	0.969*	0.972*	0.938*	0.973*
GPT	0.966*	0.949*	0.943*	0.929*	0.900*	0.950*	0.950*	0.903*	0.953*
Experts	–	0.933*	0.936*	0.922*	0.916*	0.938*	0.938*	0.892*	0.937*

Table 6: **Germany Bias Ratings:** Spearman correlations between model ratings and MBFC expert ratings vs. human ratings from political camps (liberal left, conservative left, liberal right, conservative right, agree with mainstream media, critical of mainstream media, reject mainstream media). Models vs. Experts: Bonferroni correction: $\alpha = 0.05/1 = 0.0500$; vs. political camps: Bonferroni correction: $\alpha = 0.05/8 = 0.0063$. The "Experts" row shows correlations between MBFC expert ratings and Medienkompass ratings from survey participants split by political camp. * indicates statistical significance.

	Experts	lib. left	cons. left	lib. right	cons. right	agree	critical	reject	avg.
Mixtral	0.860*	0.775*	0.795*	0.685*	0.493	0.783*	0.804*	0.174	0.780*
Llama	0.925*	0.453	0.536	0.574	0.512	0.453	0.613*	0.425	0.507
GPT	0.917*	0.585	0.679*	0.377	0.197	0.507	0.580	0.358	0.540
Experts	–	0.825*	0.656*	0.622*	0.396	0.856*	0.691*	-0.078	0.766*

Table 7: **Germany Credibility Ratings:** Spearman correlations between model ratings and MBFC expert ratings vs. human ratings from political camps (liberal left, conservative left, liberal right, conservative right, agree with mainstream media, critical of mainstream media, reject mainstream media). Models vs. Experts: Bonferroni correction: $\alpha = 0.05/1 = 0.0500$; vs. political camps: Bonferroni correction: $\alpha = 0.05/8 = 0.0063$. The "Experts" row shows correlations between MBFC expert ratings and Medienkompass ratings from survey participants split by political camp. * indicates statistical significance.

from Germany and the USA. We found that models demonstrated consistently high alignment with expert ratings, while showing weaker and more variable alignment with public opinion. The finding is limited to the scope of this experiment and should not be generalized without further investigation as expert and public opinion data sources may be included in the LLM’s training data. For US-American news outlets all models showed stronger alignment with center-left perceptions. For German news outlets the alignment is more diverse. Addressing LLM’s political biases, we observe that a higher correlation with left compared to right camps could stem from the appropriate alignment with experts rather than the model’s actual opinion on a news outlet. Future research should investigate whether high correlation with expert assessment could stem from training data contamination or actually reflects LLM’s evaluation.

7 Limitations

Several limitations should be considered when interpreting the findings of this work.

The dataset incorporates ratings with temporal ranges spanning 2019-2025 for German outlets and 2022-2025 for US-American outlets. While research suggests that bias and factualness of news

sources remain relatively stable over time, this assumption may not hold for all outlets, particularly those experiencing editorial changes, ownership transitions, or shifts in political climate, thus temporal validity might be limited.

Different scales were employed for expert opinion versus public opinion assessments e.g. factual reporting, trustworthiness and quality for credibility as well as different political bias granularities. Although all scales were ordinal and thus comparable through rank correlation, the scale heterogeneity may have limited the comparability across scales.

Preliminary analysis revealed that GPT showed evidence of prior familiarity with evaluation sources, particularly Media Bias/Fact Check (mentioned in 1.8% of reasonings). While this represents a small proportion of total responses, it raises questions about whether observed alignments reflect genuine evaluative capabilities or retrieval of memorized assessments. The extent of contamination for Llama (0.04%) was minimal, while Mixtral showed no explicit evidence, though absence of evidence does not confirm absence of exposure.

German prompts proved more challenging for models, particularly GPT, resulting in lower success rates (89.0% vs. 93.6% for English) and

more refusals. This suggests that findings may not generalize equally across languages, and models' capabilities may vary substantially depending on language-specific training data distributions. The reliance on json output formatting and specific label schemas may have introduced additional cognitive load that affected model performance differently across conditions.

This work examined only three models and two countries, limiting generalizability to other model architectures especially smaller or proprietary models and to other media ecosystems particularly non-Western democracies, authoritarian regimes, or developing nations with different media trust dynamics.

The German public opinion survey (Medienkompass) did not report demographic distribution percentages for political camps, making it difficult to assess representativeness. Additionally, the categorization scheme differed substantially between countries as for the USA the studies divided participants by party affiliation whereas for Germany participants were divided by ideological orientation and attitude towards mainstream media, complicating direct cross-national comparisons.

Media Bias/Fact Check uses a political bias scale that was developed primarily for the US context, which may not fit well with the German political landscape. What counts as "left" or "right" differs between countries as Germany has a multiparty system with proportional representation, while the USA has a two-party system. This means that applying mbfc's standardized seven-point scale to German news outlets might lead to inaccurate or misleading ratings. This limitation is particularly important because it affects how reliable the expert ratings are as a gold standard for German outlets. While mbfc was the most comprehensive source available that covered news outlets from both countries, future research would benefit from using country-specific expert rating systems that better reflect each nation's unique political categories.

Acknowledgments

We would like to thank the anonymous reviewers at WASSA workshop for their valuable feedback. This research was supported by the German Ministry of Research, Technology and Space through the research projects TrustRAG (Ref. 16IS23068) and VeraXtract (Ref. 16IS24066).

References

- Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. [Fine-grained classification of political bias in German news: A data set and initial experiments](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.
- Dolores Albarracín, Julia Albarracín, Man-pui Sally Chan, and Kathleen Hall Jamieson. 2021. *Creating Conspiracy Beliefs: How Our Thoughts Are Shaped*. Cambridge University Press. S2CID 244413957.
- Nouar Aldahoul, Hazem Ibrahim, Matteo Varvello, Aaron Kaufman, Talal Rahwan, and Yasir Zaki. 2025. [Large language models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts](#). *Preprint*, arXiv:2505.04171.
- Bassel AlKhamissi, Mohamed ElNokrashy, Marwa AlKhamissi, and Mona Diab. 2024. [Investigating cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 12404–12422.
- AllSides Staff. 2024. [Rating the bias of abc news: August-september 2024](#). Accessed: 2025-10-17.
- AllSides Staff. 2025a. [About allsides](#). Accessed: 2025-10-17.
- AllSides Staff. 2025b. [Media bias rating methods](#). Accessed: 2025-10-17.
- R. A. Armstrong. 2014. [When to use the bonferroni correction](#). *Ophthalmic and Physiological Optics*, 34(5):502–508.
- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. [Factuality challenges in the era of large language models and opportunities for fact-checking](#). *Nature Machine Intelligence*, 6:852–863.
- Yejin Bang, Danni Chen, Nayeon Lee, and Pascale Fung. 2024. [Measuring political bias in large language models: What is said and how it is said](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.
- Donald A. Barclay. 2018. *Fake News, Propaganda, and Plain Old Lies: How to Find Trustworthy Information in the Digital Age*. Rowman & Littlefield, Lanham, MD.

- Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. [Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy](#). *Preprint*, arXiv:2407.18008.
- C. E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*, volume 8.
- Lia Bozarth, Aparajita Saraf, and Ceren Budak. 2020. [Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 u.s. presidential nominees](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 48–59. Despite the varied labeling and validation procedures used and domains listed by fake news annotators, the groundtruth selection has a limited to modest impact on studies reporting on the behaviors of fake news sites.
- David A. Broniatowski, Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M. Jamison, Mark Dredze, Sandra Crouse Quinn, and John W. Ayers. 2022. [Twitter and facebook posts about covid-19 are less likely to spread misinformation compared to other health topics](#). *PLOS ONE*, 17(1):e0261768.
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. [Maxmin-rlhf: Alignment with diverse human preferences](#). *Preprint*, arXiv:2402.08925.
- Davide Chicco, Alessandro Sichenze, and Giovanni Jurman. 2025. [A simple guide to the use of student’s t-test, mann-whitney u test, chi-squared test, and kruskal-wallis test in biostatistics](#). *BioData Mining*, 18:56.
- Jan Chołoniowski, Julian Sienkiewicz, Naum Dretnik, Gregor Leban, Mike Thelwall, and Janusz A. Hołyst. 2020. [A calibrated measure to compare fluctuations of different entities across timescales](#). *Scientific Reports*, 10(1):20673.
- Edward E. Cureton. 1956. [Rank-biserial correlation](#). *Psychometrika*, 21:287–290.
- Valentin Danchev. 2022. [Reproducible data science with python: An open learning resource](#). *Journal of Open Source Education*, 5(56):156. Accessed: 2025-10-17.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. *Deep Significance*, pages 35–50. Springer International Publishing, Cham.
- Olive Jean Dunn. 1961. [Multiple comparisons among means](#). *Journal of the American Statistical Association*, 56(293):52–64.
- Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. [What did i do wrong? quantifying llms’ sensitivity and consistency to prompt engineering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2406.12334v4.
- Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. [Only a little to the left: A theory-grounded measure of political bias in large language models](#). *Preprint*, arXiv:2503.16148.
- Daniel Funke and Alexios Mantzarlis. 2018. [Here’s what to expect from fact-checking in 2019](#). Accessed: 2025-10-17.
- John Gable, Julie Mastrine, and Rick Wytmar. 2020. [Blind bias survey – allsides – august 2020](#). Accessed: 2025-10-17.
- Gene V. Glass. 1965. [A ranking variable analogue of biserial correlation: Implications for short-cut item analysis](#). *Journal of Educational Measurement*, 2(1):91–95.
- 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.
- D. Hadar-Shoval, K. Asraf, Y. Mizrachi, Y. Haber, and Z. Elyoseph. 2024. [Assessing the alignment of large language models with human values for mental health integration: Cross-sectional study using schwartz’s theory of basic values](#). *JMIR Mental Health*, 11:e55988.
- Jutta Haider and Olof Sundin. 2022. [Information literacy challenges in digital culture: conflicting engagements of trust and doubt](#). *Information, Communication & Society*, 25(8):1176–1191.
- Kevin A. Hallgren. 2012. [Computing inter-rater reliability for observational data: An overview and tutorial](#). *Tutorials in Quantitative Methods for Psychology*, 8(1):23–34.
- Inter-Parliamentary Union. 2025. [Germany - german bundestag election results \(2025-02-23\)](#). Accessed: 2025-10-17.
- Cloud IONOS. 2025. [Ai model hub](#). <https://cloud.ionos.com/managed/ai-model-hub>. Accessed: 2025-10-18.
- Charlott Jakob, David Harbecke, Patrick Parschan, Pia Wenzel Neves, and Vera Schmitt. 2025. [Polbix: Detecting llms’ political bias in fact-checking through x-phemisms](#). *arXiv preprint arXiv:2509.15335*.

- Charlott Jakob, Pia Wenzel, Salar Mohtaj, and Vera Schmitt. 2024. [Augmented political leaning detection: Leveraging parliamentary speeches for classifying news articles](#). In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 126–133, Vienna, Austria. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- Mahammed Kamruzzaman and Gene Louis Kim. 2025. [Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes](#). *Preprint*, arXiv:2404.17218.
- Meng Zhen Larsen, Michael R. Haupt, Tiana McMann, Raphael E. Cuomo, and Tim K. Mackey. 2023. [The influence of news consumption habits and dispositional traits on trust in medical scientists](#). *International Journal of Environmental Research and Public Health*, 20(10):5842.
- Noah Lee, Jiwoo Hong, and James Thorne. 2024. [Evaluating the consistency of llm evaluators](#). *Preprint*, arXiv:2412.00543.
- Jo o Alves Leite, Olga Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2025. [Weakly supervised veracity classification with llm-predicted credibility signals](#). *EPJ Data Science*, 14(1):16.
- Edoardo Loru, Jacopo Nudo, Niccol  Di Marco, Alessandro Santirocchi, Roberto Atzeni, Matteo Cinelli, Vincenzo Cestari, Clelia Rossi-Arnaud, and Walter Quattrocchi. 2025. [The simulation of judgment in llms](#). *Proceedings of the National Academy of Sciences*, 122(42).
- Insa Maab, Edison Marrese-Taylor, Sebastian Pad , and Yutaka Matsuo. 2024. [Media bias detection across families of language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*, pages 4083–4098, Mexico City, Mexico. Association for Computational Linguistics.
- Media Bias/Fact Check. 2025a. [About](#). <https://mediabiasfactcheck.com/about/>. Accessed: 2025-10-17.
- Media Bias/Fact Check. 2025b. [Methodology](#). Accessed: 2025-10-17.
- Medienkompass.org. 2025a. [Deutsche medienlandschaft](#). Accessed: 17. October 2025.
- Medienkompass.org. 2025b. [Medienkompass – eine pers nliche umfrage zur aktuellen medienlandschaft](#). Accessed: 2025-10-17.
- Medienkompass.org. 2025c. [Über medienkompass.org](#). Accessed: 17. October 2025.
- Fabio Motoki, Victor Pinho Neto, and Victor Rodrigues. 2024. [More human than human: Measuring chatgpt political bias](#). *Public Choice*, 198(1-2):3–23.
- L. Myers and M. J. Sirois. 2004. [Spearman correlation coefficients, differences between](#). In *Encyclopedia of Statistical Sciences*, volume 12. Wiley.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. [gpt-oss-120b gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Taylor Orth and Carl Bialik. 2025. [Trust in media 2025: Which news sources americans use and trust](#). Accessed: 2025-10-17.
- Tai-Quan Peng, Kaiqi Yang, Sanguk Lee, Hang Li, Yucheng Chu, Yuping Lin, and Hui Liu. 2025. [Beyond partisan leaning: A comparative analysis of political bias in large language models](#). *Preprint*, arXiv:2412.16746.
- Dorian Quelle and Alexandre Bovet. 2024. [The perils and promises of fact-checking with large language models](#). *Frontiers in Artificial Intelligence*, Volume 7 - 2024.
- Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. [Assessing political bias in large language models](#). *Preprint*, arXiv:2405.13041.
- R. Rogers. 2021. [Marginalizing the mainstream: How social media privilege political information](#). *Frontiers in Big Data*, 4:689036.
- George-Cristinel Rotaru, Sorin Anagnoste, and Marian Oancea. 2024. [How artificial intelligence can influence elections: Analyzing the large language models \(llms\) political bias](#). *Proceedings of the International Conference on Business Excellence*, 18:1882–1891.
- David Rozado. 2024. [The political preferences of llms](#). *PLOS ONE*.
- Paul R ttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Sch tze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). *Preprint*, arXiv:2402.16786.
- Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. [Ten simple rules for reproducible computational research](#). *PLoS Computational Biology*, 9(10):e1003285. Accessed: 2025-10-17.

- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Celestine Lee, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 29971–30004. PMLR.
- Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2025. [Faux polyglot: A study on information disparity in multilingual large language models.](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8090–8107, Albuquerque, New Mexico. Association for Computational Linguistics.
- C. Spearman. 1904. [The proof and measurement of association between two things.](#) *The American Journal of Psychology*, 15(1):72–101.
- S. S. Stevens. 1946. [On the theory of scales of measurement.](#) *Science*, 103(2684):677–680.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. [Large language models are inconsistent and biased evaluators.](#) *Preprint*, arXiv:2405.01724.
- Yilin Tao, Oskar Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. [Cultural bias and cultural alignment of large language models.](#) *PNAS Nexus*, 3(9):pgae346.
- Maciej Tomczak and Ewa Tomczak-Lukaszewska. 2014. [The need to report effect size estimates revisited. an overview of some recommended measures of effect size.](#) *TRENDS in Sport Sciences*, 21(1):19–25.
- Hong T Vu, Liefu Jiang, Lourdes M Cueva Chacón, Martin J Riedl, Duc V Tran, and Piotr S Bobkowski. 2019. [What influences media effects on public perception? a cross-national study of comparative agenda setting.](#) *International Communication Gazette*, 81(6-8):580–601.
- Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. [Prompt engineering in consistency and reliability with the evidence-based guideline for large language models.](#) *npj Digital Medicine*, 7:41. Accessed: 2025-10-17.
- Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. [Multilingual prompting for improving llm generation diversity.](#) *Preprint*, arXiv:2505.15229.
- Galen Weld, Maria Glenski, and Tim Althoff. 2021. [Political bias and factualness in news sharing across more than 100,000 online communities.](#) In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 796–807. S2CID 231942492, retrieved 8 June 2023.
- Tamar Wilner. 2018. [We can probably measure media bias. but do we want to?](#) Accessed: 2025-10-17.
- Kai-Cheng Yang and Filippo Menczer. 2025. [Accuracy and political bias of news source credibility ratings by large language models.](#) In *Proceedings of the 17th ACM Web Science Conference 2025*. Association for Computing Machinery.
- YouGov. 2025a. [About yougov.](#) Accessed: 2025-10-17.
- YouGov. 2025b. [Panel methodology.](#) Accessed: 2025-10-17.
- YouGov. 2025c. [Trust in media poll results.](#) Accessed: 2025-10-17.
- J. H. Zar. 1972. [Significance testing of the spearman rank correlation coefficient.](#) *Journal of the American Statistical Association*, 67(339):578–580.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. [Siren’s song in the ai ocean: A survey on hallucination in large language models.](#) *Preprint*, arXiv:2309.01219.
- Jiaqi Zhuo, Shuai Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [Prosa: Assessing and understanding the prompt sensitivity of llms.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976. Association for Computational Linguistics.

A List of news outlet names with expert labels

News Outlet	Country	Bias	Factual Reporting
Zeit	Germany	Left-Center	High
Spiegel	Germany	Left-Center	High
Cicero	Germany	Right-Center	Mostly Factual
Tichys	Germany	Right	Mixed
Süddeutsche	Germany	Left-Center	High
BILD	Germany	Right-Center	Mixed
Deutschlandfunk	Germany	Least Biased	High
Handelsblatt	Germany	Right-Center	High
ARD	Germany	Least Biased	High
ZDF	Germany	Least Biased	High
Die Welt	Germany	Right-Center	High
Frankfurter Rundschau	Germany	Left-Center	High
FAZ	Germany	Right-Center	High
taz	Germany	Left	Mostly Factual
neues Deutschland	Germany	Left	Mostly Factual
junge Welt	Germany	Left	Mixed
Compact	Germany	Extreme right	Very Low
indymedia	Germany	Left-Center	Mixed
NTV	Germany	Right-Center	Mostly Factual
RTL	Germany	Least Biased	Mostly Factual
ABC	USA	Left-Center	High
Axios	USA	Left-Center	High
BBC	USA	Left-Center	High
Breitbart News	USA	Right	Mixed
Business Insider	USA	Left-Center	Mostly Factual
CNBC	USA	Left-Center	Mostly Factual
CNN	USA	Left-Center	Mostly Factual
Forbes	USA	Least Biased	Mostly Factual
Fox Business Channel	USA	Right-Center	Mixed
Fox News	USA	Right	Mixed
MSNBC	USA	Left	Mixed
NPR	USA	Left-Center	High
National Review	USA	Right	Mostly Factual
NBC	USA	Left-Center	High
New York Post	USA	Right-Center	Mixed
Newsmax	USA	Right	Low
Newsweek	USA	Right-Center	Mostly Factual
Politico	USA	Left-Center	High
PBS	USA	Left-Center	High
Reuters	USA	Least Biased	Very High
The Associated Press	USA	Left-Center	High
The Daily Caller	USA	Right	Mixed
The Hill	USA	Least Biased	Mostly Factual
The Wall Street Journal	USA	Right-Center	Mostly Factual
The Washington Examiner	USA	Right-Center	Mostly Factual
The Washington Post	USA	Left-Center	Mostly Factual
TIME Magazine	USA	Left-Center	High

Table 8: News outlet names with bias and factual reporting labels by experts from Media Bias/Fact Check.

B Datasource Details

Media Bias/Fact Check (mbfc) is a fact-checking website, where experts rate the political bias and credibility of news outlets, and was selected to serve as the expert opinion. The ratings are available online <https://mediabiasfactcheck.com/>. mbfc is an US-American website established in 2015 by Dave M. Van Zandt that evaluates news outlets' political bias and factual reporting accuracy (Media Bias/Fact Check, 2025a). The platform employs a hybrid methodology combining objective measures and subjective analysis to assess sources across four primary categories: wording and headlines, fact-checking and sourcing, story selection, and political affiliation ((Larsen et al., 2023), (Barclay, 2018)). The methodology requires evaluation of a minimum of ten headlines and five full news stories per source, with fact checks conducted by independent reviewers affiliated with the International Fact-Checking Network (ifcn). News sources receive ratings on a seven-point political bias scale: "Extreme left", "Left", "Left-center", "Least biased", "Right-center", "Right", and "Extreme right", while the factual reporting six-point-scale consist of the labels: "Very High", "High", "Mostly Factual", "Mixed", "Low", "Very Low" (Media Bias/Fact Check, 2025b). Looking at the scales, the factual reporting ratings can serve as a credibility measure, as they reflect varying levels of perceived accuracy and commitment to verifiable information, which represent key components of credibility. The dataset used in this work includes mbfc ratings last updated between May 2023 and June 2025.

mbfc has been extensively utilized in academic research examining mainstream media and social media platforms ((Choloniewski et al., 2020), (Rogers, 2021)). Multiple studies have demonstrated high inter-rater reliability between mbfc ratings and other independent credibility assessment systems ((Weld et al., 2021), (Broniatowski et al., 2022)). Despite widespread adoption, mbfc has received methodological criticism (Funke and Mantzarlis, 2018). As mbfc also uses subjective assessments, this may introduce human biases and inconsistencies ((Wilner, 2018), (Albaracin et al., 2021)). Comparative analyses of commonly used fact-checking datasets indicate that despite differences in labeling procedures, validation methods, and domain coverage, the choice of one ground truth list over another has only a

limited impact (Bozarth et al., 2020). mbfc offers the largest dataset covering biased and low factual news sources (Weld et al., 2021), which was an important criteria as ratings for high and low credibility news outlets and right, center and left news outlets from both Germany and the USA needed to be present in the expert opinion to being able to detect different alignments and compare between countries.

For the USA we were not able to find a survey, to the best of our knowledge, with results for different political camps in which both political bias and credibility was examined at the same time. We found a survey from AllSides for political bias and a survey from YouGov for credibility.

For the political bias parameter for the public opinion for the USA we found a survey type from AllSides called "AllSides Blind Bias Survey", where participants from all sides of the political spectrum are asked to rate the content of a media outlet blindly, so they are not influenced by preconceived notions of a brand's bias (AllSides Staff, 2024). In a Blind Bias Survey, average Americans across the political spectrum read headlines and articles from a media outlet and provide an overall bias rating for the source. Respondents assess a relatively small snapshot of the source's content in time, and the surveys don't include photos or other visual elements. Blind Bias Surveys collect the political leaning of the participant, how each participant rates the media outlet, how participants with different biases (Bias Groups) rate differently, the average rating from each Bias Group and the average rating from all participants across the political spectrum ((AllSides Staff, 2025b), (Gable et al., 2020)). The scale used to evaluate the political bias of news outlets was: "Left", "Lean left", "Center", "Lean right" and "Right". The results were reported for three political camps namely Democrats, Independents and Republicans. Data from 14 different Blind Bias Surveys was used for the dataset of this thesis. The oldest survey used was from February 2022 and the newest survey from May 2025. They had an average of 841 participants with a minimum of 434 and a maximum of 1345. AllSides is an American company founded in 2012 that estimates the perceived political bias of content on online written news outlets. AllSides provides Media Bias Ratings for over 1,400 media outlets and writers (AllSides Staff, 2025a).

For the credibility parameter for the USA we found a survey called “Trust in Media 2025” from **YouGov**, where they asked US-citizens about each of 52 news sources, including their use of it in the past month and its trustworthiness. The scale used to evaluate the trustworthiness of news outlets was: “Very trustworthy”, “Trustworthy”, “Neither trustworthy nor untrustworthy”, “Untrustworthy”, “Very untrustworthy”, “Don’t know”. Looking at the scale, it is reasonable to use the trustworthiness ratings as a measure of credibility, since trustworthiness reflects perceived accuracy, reliability, and integrity of news content, which conceptually aligns with the notion of credibility. The results were reported for multiple different demographics but the relevant one for this thesis was “Party ID” which consists of three political camps namely Democrats, Independents and Republicans ([YouGov, 2025c](#)). The poll was conducted among 2,211 U.S. adult citizens in May 2025. A random sample (stratified by gender, age, race, education, geographic region, and voter registration) was selected from the 2019 American Community Survey. The sample was weighted according to gender, age, race, education, 2024 presidential vote, 2020 election turnout and presidential vote, baseline party identification, and current voter registration status. Baseline party identification is the respondent’s most recent answer given around November 8, 2024, and is weighted to the estimated distribution at that time (31% Democratic, 32% Republican). The margin of error for the overall sample is approximately 3% ([Orth and Bialik, 2025](#)). YouGov is an international Internet-based market research and data analytics firm headquartered in the UK founded in 2000 ([YouGov, 2025a](#)). YouGov uses nonprobability sampling to collect data from specific groups through an online panel of U.S. adults recruited via advertising and partnerships. This differs from probability sampling, in which all people have an equal chance of being selected into a panel. To ensure representativeness, they invite targeted panelists, weight responses based on demographics (age, gender, race, voting history) using benchmarks from sources like the U.S. Census, and offer surveys in multiple languages. Respondents receive points redeemable for money. They report margins of error to indicate the range within which results would likely fall if surveying the entire population rather than a sample ([YouGov, 2025b](#)).

For **Germany** we were able to find a public survey called **Medienkompass** that examined both political bias and credibility. The survey was carried out by [Medienkompass.org](#) and the results of the survey are available online. The results of the survey were first published in October 2019 but the website states that the survey is open and will be updated regularly ([Medienkompass.org, 2025c](#)). In the survey subjects were asked to rate 40 different German news outlets, both mainstream media and alternative media, on their partisanship and quality ([Medienkompass.org, 2025a](#)). Participants were instructed to omit media that they could not evaluate or could only evaluate poorly ([Medienkompass.org, 2025b](#)). A total of 1151 respondents positioned the news outlets on the following scales. For rating the quality of news outlets a 5-point-scale was used: “Sensationalist/Clickbait”, “Basic information”, “Meets high standards”, “Analytical”, “Complex”. For rating the partisanship of news outlets a 7-point-scale was used: “Extreme left (fake news and conspiracy theories)”, “Left-wing mission (questionable journalistic values)”, “Tending left”, “Minimal bias”, “Tending right”, “Right-wing mission (questionable journalistic values)”, “Extreme right (fake news and conspiracy theories)” ([Medienkompass.org, 2025a](#)). Looking at the scales, it allows for using the partisanship ratings as the political bias parameter and the quality ratings as the credibility parameter, because the partisanship scale measures the ideological position and extent of political leaning of an outlet, while the quality scale reflects editorial standards and analytical depth. The Medienkompass survey reported results for seven different political or ideological camps, differentiating camps into two broader categories called “mindset” and “agreement” with the following subcategories and their descriptions:

Mindset:

- Liberal-left (Opinion of participants who are liberal-left-orientated)
- Conservative-left (Opinion of participants who are conservative-left-orientated)
- Liberal-right (Opinion of participants who are liberal-right-orientated)
- Conservative-right (Opinion of participants who are conservative-right-orientated)

Agreement:

- High agreement (Opinion of participants who largely agree with the mainstream media)
- Critically-distanced (Opinion of participants who are critical of the mainstream media)
- Rejecting (Opinion of participants who largely reject the mainstream media)

In contrast to the surveys from the USA, political camps in the Medienkompass survey do not correspond to major political parties, but to more general ideological orientations. For the category “mindset” they divided participants into four different camps using “left” and “right” in combination with “conservative” and “liberal”. This corresponds more closely to Germany’s multiparty system, where around six major parties have parliamentary representation ([Inter-Parliamentary Union, 2025](#)). For the category “agreement” the answer options were presented to the participants in the following way, which we included here as the category names itself are not very telling:

“Which statement would you most agree with:

- I am generally satisfied with the current reporting in the mainstream media. Our media represent freedom of expression, and despite my critical attitude, I mostly agree with their reporting.
- I consider the reporting in the mainstream media to be biased and am critical of it. There is no such thing as an impartial opinion—our media landscape is still better than the censorship or conformity of other countries.
- I largely reject the reporting of the mainstream media and now get my information almost exclusively from other sources([Medienkompass.org, 2025b](#)).

The percentage for the “agreement” category were 19,4% for “Rejecting”, 40,3% for “Critically-distanced” and 40,3% for “High agreement” ([Medienkompass.org, 2025a](#)). The percentage shares for the “mindset” category were not published. Medienkompass dataset has been used to examine political bias in German news before by [Aksenov et al. \(2021\)](#) and [Jakob et al. \(2024\)](#).

C Dataset label distribution

Label	Ct.	Group	Ct.	%
Germany Distribution 'Bias'				
Left	3			
Left-Center	5	Left	8	40
Least Biased	4	Center	4	20
Right-Center	6			
Right	1	Right	8	40
Extreme Right	1			

Label	Ct.	Group	Ct.	%
Germany Distribution 'Factual Reporting'				
Very High	0			
High	10	High	10	50
Mostly Factual	5	Medium	5	25
Mixed	4			
Low	0	Low	5	25
Very Low	1			

Label	Ct.	Group	Ct.	%
USA Distribution 'Bias'				
Left	1			
Left-Center	13	Left	14	52
Least Biased	3	Center	3	11
Right-Center	5			
Right	5	Right	10	37
Extreme Right	0			

Label	Ct.	Group	Ct.	%
USA Distribution 'Factual Reporting'				
Very High	1			
High	9	High	10	37
Mostly Factual	10	Medium	10	37
Mixed	6			
Low	1	Low	7	26
Very Low	0			

Table 9: Label distributions by political bias and factual reporting for Germany and USA news outlets

D Chi-square test of bias-factual reporting distribution differences between Germany and USA.

Category	DE Ct.	DE %	USA Ct.	USA %	Diff	Total	DE Expected	USA Expected
Left & High	4	20.0%	8	29.6%	-9.63%	12	5.11	6.89
Left & Medium	2	10.0%	4	14.8%	-4.81%	6	2.55	3.45
Left & Low	2	10.0%	1	3.7%	6.30%	3	1.28	1.72
Center & High	3	15.0%	1	3.7%	11.30%	4	1.70	2.30
Center & Medium	0	0.0%	3	11.1%	-11.11%	3	1.28	1.72
Center & Low	1	5.0%	0	0.0%	5.00%	1	0.43	0.57
Right & High	3	15.0%	0	0.0%	15.00%	3	1.28	1.72
Right & Medium	2	10.0%	4	14.8%	-4.81%	6	2.55	3.45
Right & Low	3	15.0%	6	22.2%	-7.22%	9	3.83	5.17
Total	20	100%	27	100%		47		

Table 10: In order to examine whether there are systematic significant differences in the proportion of e.g., left, high-factual reporting news outlets between Germany and the USA, counts per bias-factual reporting combinations were made. For each category, the expected count was calculated, which means the number of news outlets you would expect to fall into that category if there was no difference between Germany and USA. The following formula was used: $\text{Expected} = (\text{Row Total} \times \text{Column Total}) / \text{Grand Total}$. For example, the expected frequency for "Left & High" in Germany was calculated as: $(\text{Left \& High row total}) \times (\text{Sum of Germany column}) / \text{Grand Total}$. While there were some numerical differences in the distribution of bias and factual reporting ratings between the German ($n = 20$) and US ($n = 27$) news outlets, **a chi-square test revealed these differences were not statistically significant** ($\chi^2, p = 0.19$). There is a 19% chance of observing differences this large just by random sampling, even if Germany and USA truly had identical distributions. Both corpora showed similar overall patterns in terms of political bias and factual reporting.

E Overview of rating scales

		public opinion		expert opinion
		Germany	USA	Germany & USA
scale name		political bias	political bias	political bias
source name		Medienkompass	AllSides	Media Bias/Fact Check
scales bias	Extreme left (fake news and conspiracy theories)			Extreme left
	Left-wing mission (questionable journalistic values)	left		Left
	Tending left	lean left		Left-Center
	Minimal bias	center		Least Biased
	Tending right	lean right		Right-Center
	Right-wing mission (questionable journalistic values)	right		Right
	Extreme right (fake news and conspiracy theories)			Extreme right
scale name		quality	trustworthiness	factual reporting
source name		Medienkompass	YouGov	Media Bias/Fact Check
scales credibility	Complex		Very trustworthy	Very High
	Analytical		Trustworthy	High
	Meets high standards		Neither trustworthy nor untrustworthy	Mostly Factual
	Basic information		Untrustworthy	Mixed
	Sensationalist/Clickbait		Very untrustworthy	Low
		Don't know		Very Low

Figure 3: Overview of rating scales of the individual data sources. As there is a different scale for the political bias parameter and the credibility parameter for each country's public opinion and another set of scales for the expert opinion that sums up to six different scales. In order to ensure comparability between the human assessment results and the LLMs' results the original scales from the human assessments were used. We decided to not create one common scale for political bias and one common scale for credibility but to use exactly the scales presented to the participants during the surveys. The underlying idea is that alignment between the model's opinion and human opinion can only be meaningfully assessed if the model is provided with the same response options as the human respondents. The scales used for the credibility parameter have other names in the different data sources. It is reasonable to use the trustworthiness ratings as a measure of credibility, since trustworthiness reflects perceived accuracy, reliability, and integrity of news content, which conceptually aligns with the notion of credibility. The quality scale reflects editorial standards and analytical depth and the factual reporting ratings reflect varying levels of perceived accuracy and commitment to verifiable information, which represent key components of credibility.

F Example json instance output

```
{
  "model": "openai/gpt-oss-120b",
  "prompt_category": "rating",
  "prompt_subcategory": "rate_allides_bias",
  "prompt_id": "rating_rate_allides_bias_en",
  "rate_or_opinion": "rate",
  "survey": "allides",
  "examined_parameter": "bias",
  "repetition": 9,
  "newsoutlet_id": "usa_abc",
  "newsoutlet": "ABC",
  "country": "USA",
  "language": "English",
  "reasoning_political_bias": "ABC News strives for balanced coverage, with occasional critiques of a mild liberal slant, but overall maintains centrist reporting standards.",
  "label_political_bias": "center"
},
```

Figure 4: Example of an LLM's generated response in form of a json instance from the created output file. In addition to the fields returned by the API (reasoning and label), supplementary metadata was programmatically appended to each JSON record to facilitate subsequent analyses. These included the name of the prompted model, prompt identifiers, prompt style, dataset to be compared with, examined parameter, repetition counter, news outlet id and name, country and language.

G Overview of prompt template categories

opinion type	dataset	parameter	prompt variation	language	template number	country	news outlets	repetitions	total prompts
expert opinion	Media Bias/ Fact Check	bias	rate	English	template 1	both	47	10	470
				German	template 2	both	47	10	470
			opinion	English	template 3	both	47	10	470
				German	template 4	both	47	10	470
		factual reporting	rate	English	template 5	both	47	10	470
				German	template 6	both	47	10	470
			opinion	English	template 7	both	47	10	470
				German	template 8	both	47	10	470
public opinion Germany	Medienkompass	bias	rate	English	template 9	Germany	20	10	200
				German	template 10	Germany	20	10	200
			opinion	English	template 11	Germany	20	10	200
				German	template 12	Germany	20	10	200
		quality	rate	English	template 13	Germany	20	10	200
				German	template 14	Germany	20	10	200
			opinion	English	template 15	Germany	20	10	200
				German	template 16	Germany	20	10	200
public opinion USA	AllSides	bias	rate	English	template 17	USA	27	10	270
				German	template 18	USA	27	10	270
			opinion	English	template 19	USA	27	10	270
				German	template 20	USA	27	10	270
	YouGov	trustworthiness	rate	English	template 21	USA	27	10	270
				German	template 22	USA	27	10	270
			opinion	English	template 23	USA	27	10	270
				German	template 24	USA	27	10	270
total amount of prompts per model:									7.520

Figure 5: Overview of prompt template categories and calculation of amount of prompts per model. Per prompt template we received 10 ratings for each news outlet. Due to the two different prompt styles, the two different languages and the prompt repetitions each model was prompted 40 times per news outlet per parameter per dataset. For example, for the news outlet ABC from the USA I received 40 political bias labels and 40 credibility labels using the labels from the expert opinion scale and 40 political bias labels and 40 credibility labels using the labels from the public opinion scales. Per model a total of 7.520 prompts were sent to the model for evaluation. With three models that leads to a total amount of 22.560 prompts.

H Overview of success rates of LLM responses

model	opinion type	dataset	parameter	prompt variation	language	success rate	initial null labels	remaining null labels	refused outlets	affected outlets
gpt-oss-120b	expert opinion	Media Bias/ Fact Check	bias	rate	English	93.8%	32	0	0	
					German	97.5%	13	0	0	
				opinion	English	93.5%	34	10	1	indymedia
				German	98.3%	9	0	0		
			factual reporting	rate	English	92.9%	37	0	0	
					German	95.6%	23	0	0	
	opinion	English		82.9%	89	9	0	indymedia		
		German	95.4%	24	0	0				
	public opinion Germany	Medienkompass	bias	rate	English	97.5%	5	0	0	
					German	71.5%	57	10	1	Zeit
				opinion	English	95.0%	10	0	0	
				German	71.0%	58	20	2	indymedia, Cicero	
			quality	rate	English	96.5%	7	0	0	
					German	100%	0	0	0	
	opinion	English		100%	0	0	0			
		German	100%	0	0	0				
public opinion USA	AllSides	bias	rate	English	97.8%	6	0	0		
				German	99.3%	2	0	0		
			opinion	English	97.4%	7	0	0		
				German	97.4%	7	0	0		
	YouGov	trustworthiness	rate	English	91.9%	22	9	0	ABC	
				German	79.3%	56	0	0		
			opinion	English	84.1%	43	0	0		
				German	62.6%	101	10	1	ABC	
Mixtral-8x7B-Instruct-v0.1	expert opinion	Media Bias/ Fact Check	bias	rate	English	100%	0	0	0	
					German	100%	0	0	0	
				opinion	English	100%	0	0	0	
				German	100%	0	0	0		
			factual reporting	rate	English	100%	0	0	0	
					German	100%	0	0	0	
	opinion	English		100%	0	0	0			
		German	98.3%	9	0	0				
	public opinion Germany	Medienkompass	bias	rate	English	100%	0	0	0	
					German	95.0%	10	0	0	
				opinion	English	100%	0	0	0	
				German	100%	0	0	0		
			quality	rate	English	100%	0	0	0	
					German	97.0%	6	0	0	
	opinion	English		94.0%	12	0	0			
		German	95.5%	9	0	0				
public opinion USA	AllSides	bias	rate	English	100%	0	0	0		
				German	100%	0	0	0		
			opinion	English	100%	0	0	0		
				German	95.6%	12	0	0		
	YouGov	trustworthiness	rate	English	100%	0	0	0		
				German	99%	3	0	0		
	opinion	English	100%	0	0	0				
	German	100%	0	0	0					

Figure 6: Overview of success rates, null labels and refusals of LLM responses regarding news outlet ratings. The success rates, initial amount of null labels, remaining amount of null labels, the amount of refused outlets and the names of the affected news outlets are displayed. The table only contains the values for GPT and Mixtral as Llama is not included in the table, because Llama showed 100% success rates.

I Statistical test results for the Consistency and Accuracy Analysis

opinion type	dataset	parameter	H-value	p-value	sig.
expert opinion	Media Bias/ Fact Check	bias	34.4337	0.0000	✓
		factual reporting	120.3926	0.0000	✓
public opinion Germany	Medienkompass	bias	41.9909	0.0000	✓
		quality	18.2008	0.0001	✓
public opinion USA	AllSides	bias	44.8765	0.0000	✓
	YouGov	trustworthiness	50.6474	0.0000	✓

Figure 7: Kruskal-Wallis test for overall model comparison.

opinion type	dataset	parameter	comparison	p-value	sig.	rank-biserial	effect size
expert opinion	Media Bias/ Fact Check	bias	GPT vs Llama	< 0.001	✓	0.244	Small
			GPT vs Mixtral	< 0.001	✓	0.227	Small
			Llama vs Mixtral	0.208	—	-0.053	Negligible
		factual reporting	GPT vs Llama	< 0.001	✓	0.429	Medium
			GPT vs Mixtral	< 0.001	✓	0.526	Large
			Llama vs Mixtral	0.296	—	0.044	Negligible
public opinion Germany	Medienkompass	bias	GPT vs Llama	< 0.001	✓	0.264	Small
			GPT vs Mixtral	< 0.001	✓	0.467	Medium
			Llama vs Mixtral	< 0.001	✓	0.194	Small
		quality	GPT vs Llama	0.005	✓	0.234	Small
			GPT vs Mixtral	< 0.001	✓	0.337	Medium
			Llama vs Mixtral	0.271	—	0.083	Negligible
public opinion USA	AllSides	bias	GPT vs Llama	< 0.001	✓	0.341	Medium
			GPT vs Mixtral	< 0.001	✓	0.319	Small
			Llama vs Mixtral	0.282	—	-0.048	Negligible
	YouGov	trustworthiness	GPT vs Llama	< 0.001	✓	0.336	Medium
			GPT vs Mixtral	< 0.001	✓	0.342	Medium
			Llama vs Mixtral	0.731	—	-0.014	Negligible

Figure 8: Mann-Whitney test with rank-biserial correlation coefficient for pairwise model comparison

opinion type	dataset	parameter	model	p-value	sig.	rank-biserial	meaning	effect size
expert opinion	Media Bias/ Fact Check	bias	GPT-OSS-120B	0.180	—	0.101	DE more cons.	Negligible
			Llama-3.3-70B	0.005	✓	0.154	DE more cons.	Small
			Mixtral-8x7B	0.874	—	-0.010	EN more cons.	Negligible
		factual reporting	GPT-OSS-120B	0.570	—	0.047	DE more cons.	Negligible
			Llama-3.3-70B	0.870	—	0.010	DE more cons.	Negligible
			Mixtral-8x7B	0.346	—	-0.055	EN more cons.	Negligible
public opinion Germany	Medienkompass	bias	GPT-OSS-120B	0.374	—	-0.109	EN more cons.	Negligible
			Llama-3.3-70B	0.985	—	-0.002	EN more cons.	Negligible
			Mixtral-8x7B	0.165	—	-0.076	EN more cons.	Negligible
		quality	GPT-OSS-120B	0.008	✓	-0.329	EN more cons.	Medium
			Llama-3.3-70B	0.325	—	0.107	DE more cons.	Negligible
			Mixtral-8x7B	0.605	—	-0.054	TIE	Negligible
public opinion USA	AllSides	bias	GPT-OSS-120B	0.144	—	0.147	DE more cons.	Negligible
			Llama-3.3-70B	0.521	—	-0.036	EN more cons.	Negligible
			Mixtral-8x7B	0.037	✓	0.144	EN more cons.	Negligible
	YouGov	trustworthiness	GPT-OSS-120B	0.288	—	0.106	DE more cons.	Negligible
			Llama-3.3-70B	0.832	—	-0.012	EN more cons.	Negligible
			Mixtral-8x7B	0.030	✓	-0.127	EN more cons.	Negligible

Figure 9: Language effects on model consistency

opinion type	dataset	parameter	model	p-value	sig.	rank-biserial	meaning	effect size
expert opinion	Media Bias/ Fact Check	bias	GPT-OSS-120B	0.632	—	-0.036	rate more cons.	Negligible
			Llama-3.3-70B	0.774	—	-0.016	rate more cons.	Negligible
			Mixtral-8x7B	0.622	—	-0.031	rate more cons.	Negligible
		factual reporting	GPT-OSS-120B	0.637	—	-0.039	rate more cons.	Negligible
			Llama-3.3-70B	0.037	✓	-0.126	rate more cons.	Negligible
			Mixtral-8x7B	0.191	—	0.076	opinion more cons.	Negligible
public opinion Germany	Medienkompass	bias	GPT-OSS-120B	0.135	—	-0.182	rate more cons.	Small
			Llama-3.3-70B	0.226	—	-0.120	rate more cons.	Negligible
			Mixtral-8x7B	0.022	✓	-0.125	rate more cons.	Negligible
		quality	GPT-OSS-120B	0.528	—	-0.078	rate more cons.	Negligible
			Llama-3.3-70B	0.448	—	0.083	opinion more cons.	Negligible
			Mixtral-8x7B	0.389	—	-0.090	rate more cons.	Negligible
public opinion USA	AllSides	bias	GPT-OSS-120B	0.433	—	-0.079	rate more cons.	Negligible
			Llama-3.3-70B	0.574	—	-0.032	rate more cons.	Negligible
			Mixtral-8x7B	0.625	—	0.034	opinion more cons.	Negligible
	YouGov	trustworthiness	GPT-OSS-120B	0.543	—	0.061	opinion more cons.	Negligible
			Llama-3.3-70B	0.284	—	-0.058	rate more cons.	Negligible
			Mixtral-8x7B	0.355	—	0.055	opinion more cons.	Negligible

Figure 10: Prompt style effects on model consistency

opinion	dataset	parameter	model	p-value	sig.	rank-biserial	meaning	effect size
expert opinion	Media Bias/ Fact Check	bias	GPT-OSS-120B	0.884	—	-0.010	EN more acc.	Negligible
			Llama-3.3-70B	0.149	—	-0.106	EN more acc.	Negligible
			Mixtral-8x7B	0.683	—	0.031	DE more acc.	Negligible
		factual reporting	GPT-OSS-120B	0.027	✓	-0.165	EN more acc.	Small
			Llama-3.3-70B	0.717	—	-0.027	EN more acc.	Negligible
			Mixtral-8x7B	0.282	—	-0.079	EN more acc.	Negligible

Figure 11: Language effects on model accuracy

opinion	dataset	parameter	model	p-value	sig.	rank-biserial	meaning	effect size
expert opinion	Media Bias/ Fact Check	bias	GPT-OSS-120B	0.884	—	0.010	opinion more acc.	Negligible
			Llama-3.3-70B	0.938	—	0.006	Tie	Negligible
			Mixtral-8x7B	0.994	—	-0.001	rate more acc.	Negligible
		factual reporting	GPT-OSS-120B	1.000	—	0.000	opinion more acc.	Negligible
			Llama-3.3-70B	0.933	—	0.006	opinion more acc.	Negligible
			Mixtral-8x7B	0.723	—	-0.026	Tie	Negligible

Figure 12: Prompt style effects on model accuracy

J Visualization of Model ratings vs. Expert ratings of news outlets

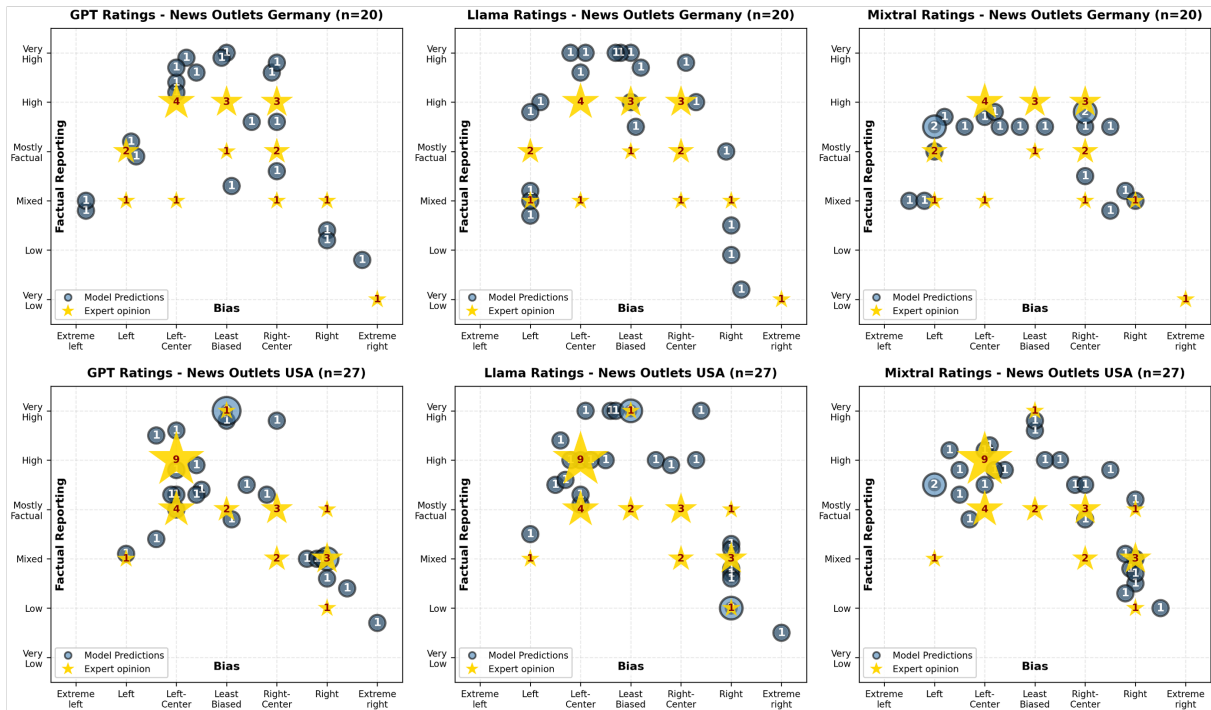


Figure 13: Model ratings vs. Expert ratings with news outlets in form of symbols. In order to make it easier to detect patterns we replaced the news outlet names with symbols for a more concise and organized display. Model predictions are represented by blue circles, while expert human ratings are indicated by yellow stars. The number within the circles or stars denotes the count of overlapping outlets in that coordinate position.

K Proximity of model opinion and public opinion to expert opinion

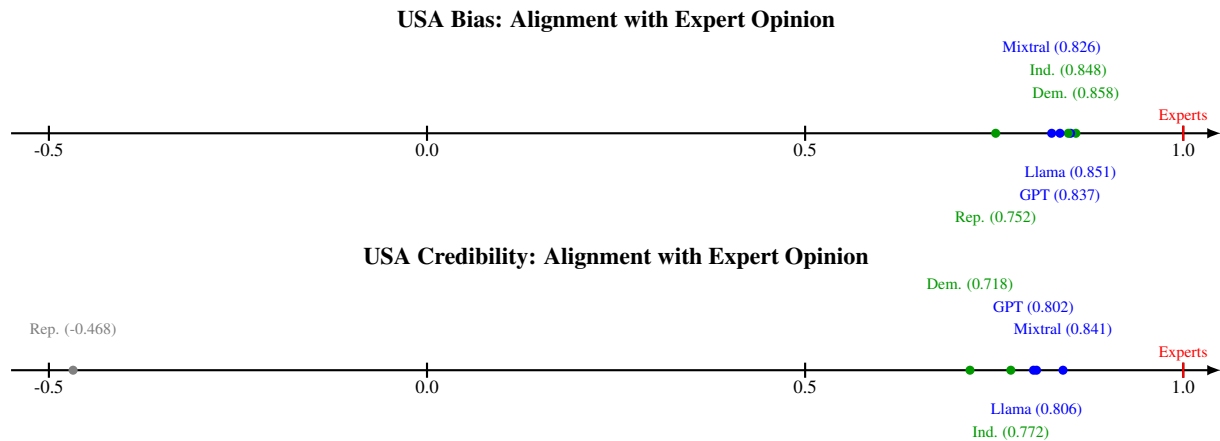


Figure 14: Spearman correlation with expert ratings for USA media outlets. Models shown in blue, political camps in green (gray indicates non-significant correlation). The expert reference point (red) represents perfect alignment ($\rho = 1.0$). Dem. = Democrats, Ind. = Independents, Rep. = Republicans.

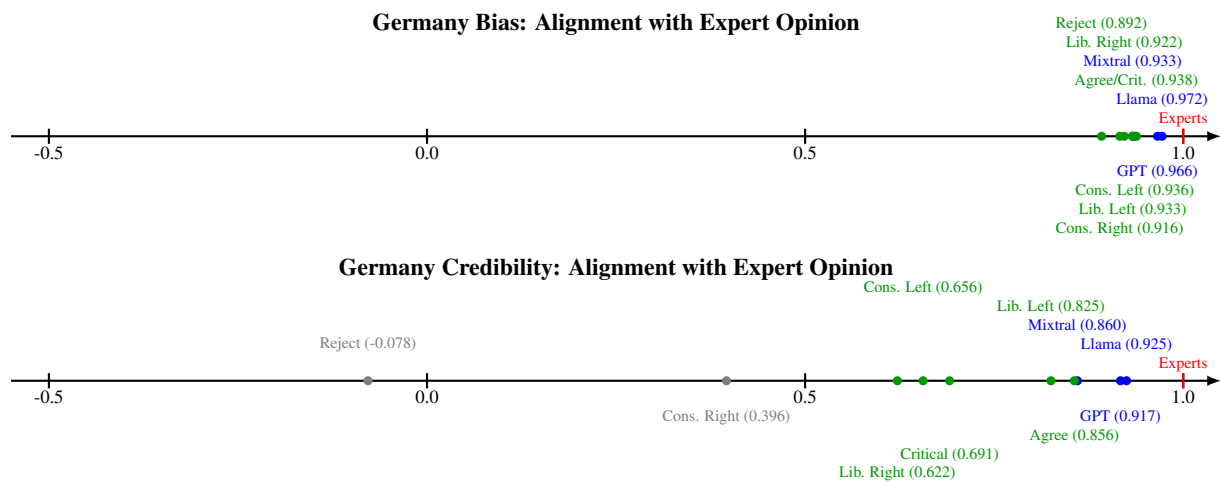


Figure 15: Spearman correlation with expert ratings for German media outlets. Models shown in blue, political/media attitude camps in green (gray indicates non-significant correlation). The expert reference point (red) represents perfect alignment ($\rho = 1.0$). Lib. = Liberal, Cons. = Conservative, Agree = Agree with mainstream media, Critical = Critical of mainstream media, Reject = Reject mainstream media.

L Proximity of expert opinion and public opinion to model opinion

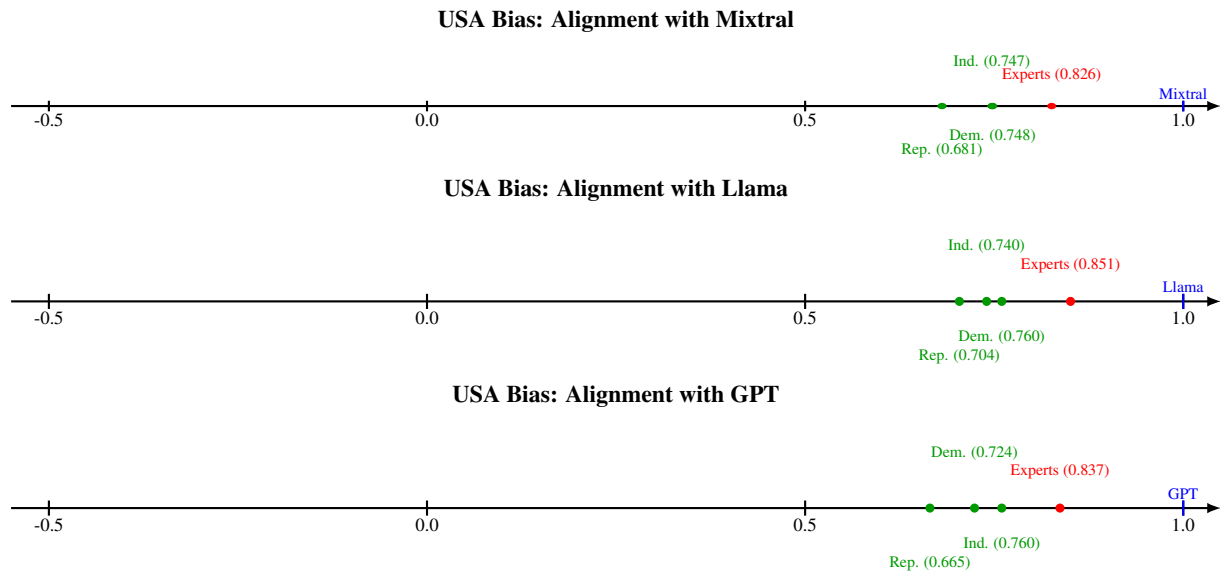


Figure 16: Spearman correlation with model ratings for USA media outlets (Bias). Experts shown in red, political camps in green. Each model serves as the reference point ($\rho = 1.0$). Dem. = Democrats, Ind. = Independents, Rep. = Republicans.

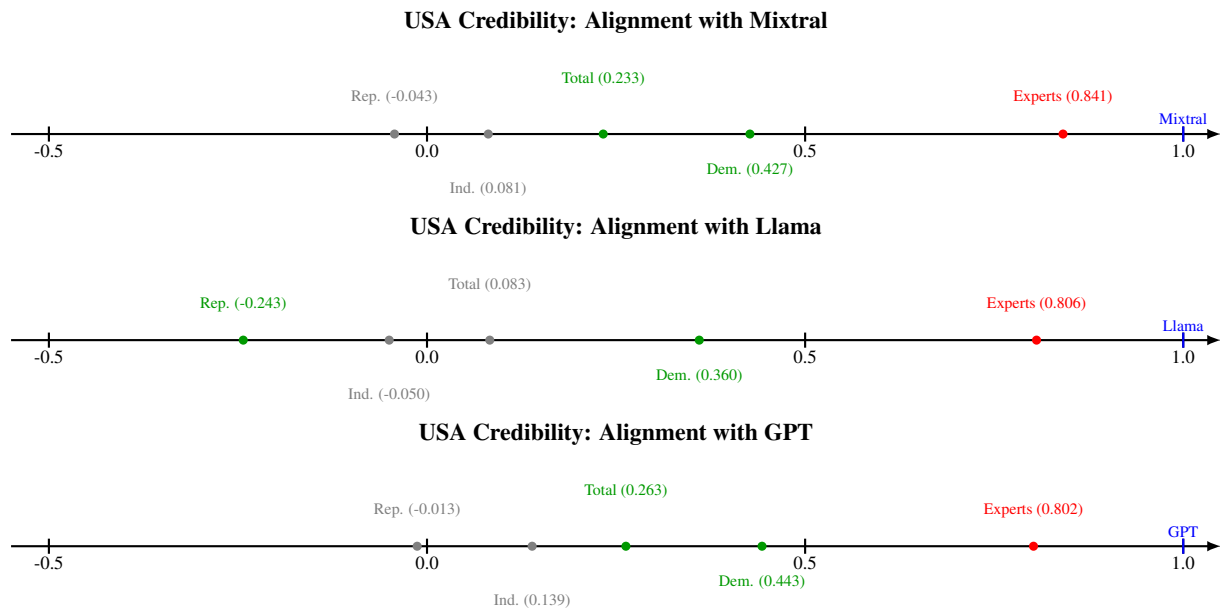


Figure 17: Spearman correlation with model ratings for USA media outlets (Credibility). Experts shown in red, political camps in green (gray indicates non-significant correlation). Each model serves as the reference point ($\rho = 1.0$). Dem. = Democrats, Ind. = Independents, Rep. = Republicans, Total = all respondents.

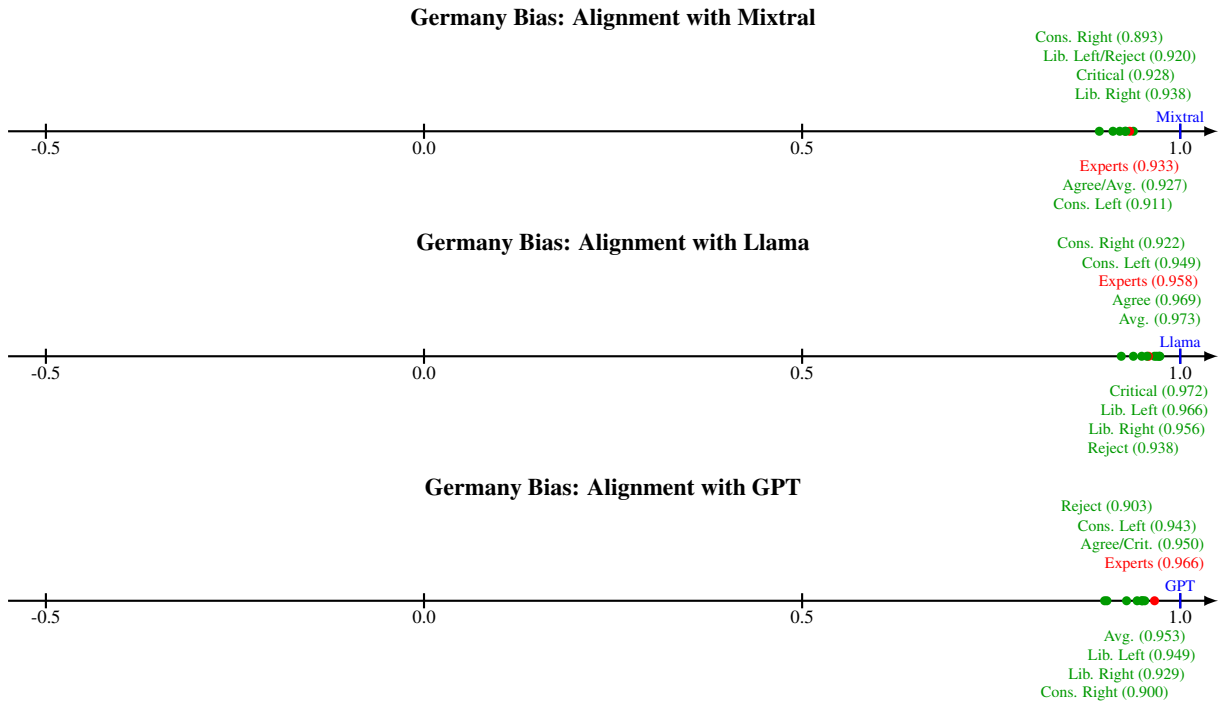


Figure 18: Spearman correlation with model ratings for German media outlets (Bias). Experts shown in red, political/media attitude camps in green. Each model serves as the reference point ($\rho = 1.0$). Lib. = Liberal, Cons. = Conservative, Agree = Agree with mainstream media, Critical = Critical of mainstream media, Reject = Reject mainstream media, Avg. = Average across all camps.

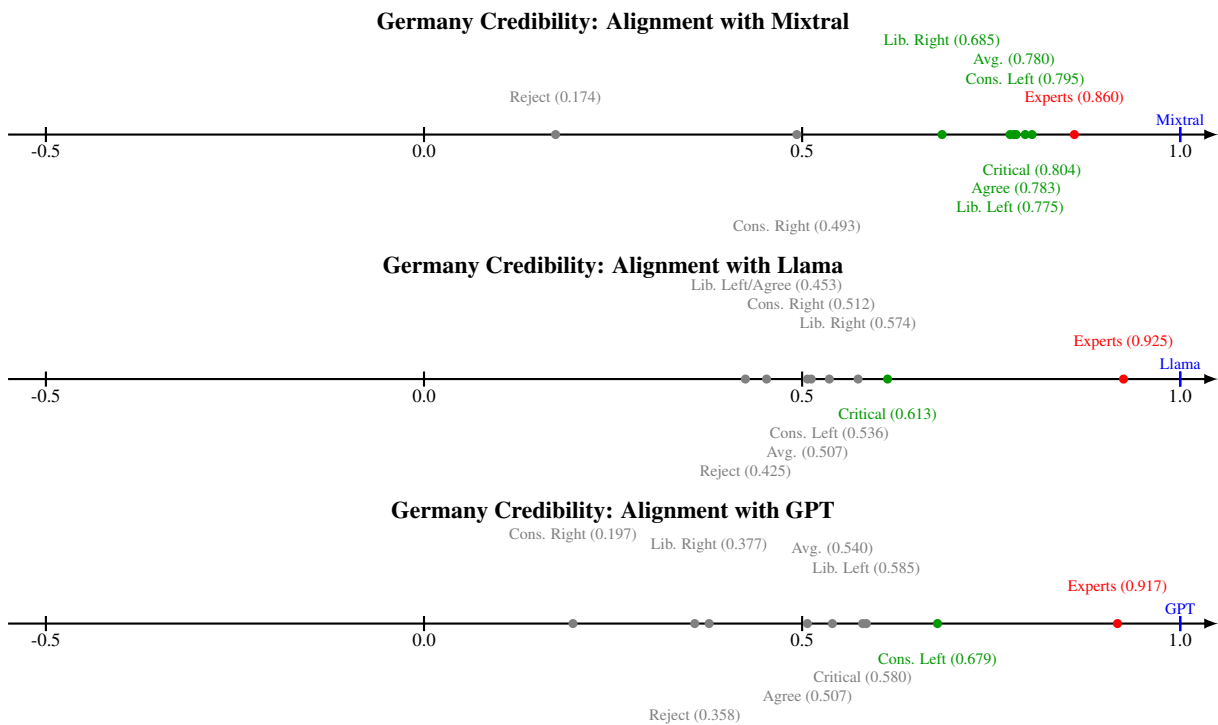


Figure 19: Spearman correlation with model ratings for German media outlets (Credibility). Experts shown in red, political/media attitude camps in green (gray indicates non-significant correlation). Each model serves as the reference point ($\rho = 1.0$). Lib. = Liberal, Cons. = Conservative, Agree = Agree with mainstream media, Critical = Critical of mainstream media, Reject = Reject mainstream media, Avg. = Average across all camps.

M All prompt templates in full text

prompt id	actual prompt
rating_rate_mbfc_factual_reporting_en	<pre> return [{ "role": "user", "content": f"Rate the news outlet '{newsoutlet}' from {country} regarding its factual reporting on the following scale: - Very High - High - Mostly Factual - Mixed - Low - Very Low Return ONLY this JSON (no extra text): {{'reasoning_factual_reporting': '<short reasoning>', 'label_factual_reporting': '<EXACTLY one of: Very High, High, Mostly Factual, Mixed, Low, Very Low>'}} """ }] </pre>
rating_rate_mbfc_factual_reporting_de	<pre> return [{ "role": "user", "content": f"Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer faktenbasierten Berichterstattung auf der folgenden Skala: - Sehr hoch - Hoch - Überwiegend faktenbasiert - Gemischt - Niedrig - Sehr niedrig Gib NUR diese JSON zurück (kein extra Text): {{'reasoning_factual_reporting': '<kurze Erläuterung>', 'label_factual_reporting': '<EXAKT eins von: Sehr hoch, Hoch, Überwiegend faktenbasiert, Gemischt, Niedrig, Sehr niedrig>'}} """ }] </pre>
rating_rate_mbfc_bias_en	<pre> return [{ "role": "user", "content": f"Rate the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - Extreme left - Left - Left-Center - Least Biased - Right-Center - Right - Extreme right Return ONLY this JSON (no extra text): {{'reasoning_political_bias': '<short reasoning>', 'label_political_bias': '<EXACTLY one of: Extreme left, Left, Left-Center, Least Biased, Right-Center, Right, Extreme right>'}} """ }] </pre>
rating_rate_mbfc_bias_de	<pre> return [{ "role": "user", "content": f"Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - Extrem links - Links - Links-Mitte - kaum parteiisch - Rechts-Mitte - Rechts - Extrem rechts Gib NUR diese JSON zurück (kein extra Text): {{'reasoning_political_bias': '<kurze Erläuterung>', 'label_political_bias': '<EXAKT eins von: Extrem links, Links, Links-Mitte, kaum parteiisch, Rechts-Mitte, Rechts, Extrem rechts>'}} """ }] </pre>

rating_please_opinion_mbfc_factual_reporting_en	<pre> return [{ "role": "user", "content": f"Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its factual reporting on the following scale: - Very High - High - Mostly Factual - Mixed - Low - Very Low Return ONLY this JSON (no extra text): {{"reasoning_factual_reporting": "<short reasoning>", "label_factual_reporting": "<EXACTLY one of: Very High, High, Mostly Factual, Mixed, Low, Very Low>"}} """ }] </pre>
rating_please_opinion_mbfc_factual_reporting_de	<pre> return [{ "role": "user", "content": f"Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer faktenbasierten Berichterstattung auf der folgenden Skala: - Sehr hoch - Hoch - Überwiegend faktenbasiert - Gemischt - Niedrig - Sehr niedrig Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_factual_reporting": "<kurze Erläuterung>", "label_factual_reporting": "<EXAKT eins von: Sehr hoch, Hoch, Überwiegend faktenbasiert, Gemischt, Niedrig, Sehr niedrig>"}} """ }] </pre>
rating_please_opinion_mbfc_bias_en	<pre> return [{ "role": "user", "content": f"Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - Extreme left - Left - Left-Center - Least Biased - Right-Center - Right - Extreme right Return ONLY this JSON (no extra text): {{"reasoning_political_bias": "<short reasoning>", "label_political_bias": "<EXACTLY one of: Extreme left, Left, Left-Center, Least Biased, Right-Center, Right, Extreme right>"}} """ }] </pre>
rating_please_opinion_mbfc_bias_de	<pre> return [{ "role": "user", "content": f"Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - Extrem links - Links - Links-Mitte - kaum parteiisch - Rechts-Mitte - Rechts - Extrem rechts Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_political_bias": "<kurze Erläuterung>", "label_political_bias": "<EXAKT eins von: Extrem links, Links, Links-Mitte, kaum parteiisch, Rechts-Mitte, Rechts, Extrem rechts>"}} """ }] </pre>

rating_rate_medienkompass_quality_en	<pre> return [{ "role": "user", "content": f"""Rate the news outlet '{newsoutlet}' from {country} regarding its quality on the following scale: - Complex - Analytical - Meets high standards - Basic information - Sensationalist/Clickbait Return ONLY this JSON (no extra text): {{"reasoning_quality": "<short reasoning>", "label_quality": "<EXACTLY one of: Complex, Analytical, Meets high standards, Basic information, Sensationalist/Clickbait>"}} """ }] </pre>
rating_rate_medienkompass_quality_de	<pre> return [{ "role": "user", "content": f"""Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Qualität auf der folgenden Skala: - Komplex - Analytisch - Erfüllt hohe Standards - Basis Information - Reißerisch/Clickbait Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_quality": "<kurze Erläuterung>", "label_quality": "<EXAKT eins von: Komplex, Analytisch, Erfüllt hohe Standards, Basis Information, Reißerisch/Clickbait>"}} """ }] </pre>
rating_rate_medienkompass_bias_en	<pre> return [{ "role": "user", "content": f"""Rate the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - Extreme left (fake news and conspiracy theories) - Left-wing mission (questionable journalistic values) - Tending left - Minimal bias - Tending right - Right-wing mission (questionable journalistic values) - Extreme right (fake news and conspiracy theories) Return ONLY this JSON (no extra text): {{"reasoning_political_bias": "<short reasoning>", "label_political_bias": "<EXACTLY one of: Extreme left, Left-wing Mission, Tending left, Minimal bias, Tending right, Right-wing mission, Extreme right>"}} """ }] </pre>
rating_rate_medienkompass_bias_de	<pre> return [{ "role": "user", "content": f"""Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - Linksextrem (Fake News und Verschwörungstheorien) - Linke Mission (fragwürdige journalistische Werte) - Tendenziell links - minimale Parteiische Tendenz - Tendenziell rechts - Rechte Mission (fragwürdige journalistische Werte) - Rechtsextrem (Fake News und Verschwörungstheorien) Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_political_bias": "<kurze Erläuterung>", "label_political_bias": "<EXAKT eins von: Linksextrem, Linke Mission, Tendenziell links, minimale Parteiische Tendenz, Tendenziell rechts, Rechte Mission, Rechtsextrem>"}} """ }] </pre>

rating_please_opinion_medienkompass_quality_en	<pre> return [{ "role": "user", "content": f"Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its quality on the following scale: - Complex - Analytical - Meets high standards - Basic information - Sensationalist/Clickbait Return ONLY this JSON (no extra text): {{"reasoning_quality": "<short reasoning>", "label_quality": "<EXACTLY one of: Complex, Analytical, Meets high standards, Basic information, Sensationalist/Clickbait>"}} """ }] </pre>
rating_please_opinion_medienkompass_quality_de	<pre> return [{ "role": "user", "content": f"Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Qualität auf der folgenden Skala: - Komplex - Analytisch - Erfüllt hohe Standards - Basis Information - Reißerisch/Clickbait Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_quality": "<kurze Erläuterung>", "label_quality": "<EXAKT eins von: Komplex, Analytisch, Erfüllt hohe Standards, Basis Information, Reißerisch/Clickbait>"}} """ }] </pre>
rating_please_opinion_medienkompass_bias_en	<pre> return [{ "role": "user", "content": f"Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - Extreme left (fake news and conspiracy theories) - Left-wing mission (questionable journalistic values) - Tending left - Minimal bias - Tending right - Right-wing mission (questionable journalistic values) - Extreme right (fake news and conspiracy theories) Return ONLY this JSON (no extra text): {{"reasoning_political_bias": "<short reasoning>", "label_political_bias": "<EXACTLY one of: Extreme left, Left-wing mission, Tending left, Minimal bias, Tending right, Right-wing mission, Extreme right>"}} """ }] </pre>
rating_please_opinion_medienkompass_bias_de	<pre> return [{ "role": "user", "content": f"Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - Linksextrem (Fake News und Verschwörungstheorien) - Linke Mission (fragwürdige journalistische Werte) - Tendenziell links - minimale Parteiische Tendenz - Tendenziell rechts - Rechte Mission (fragwürdige journalistische Werte) - Rechtsextrem (Fake News und Verschwörungstheorien) Gib NUR diese JSON zurück (kein extra Text): {{"reasoning_political_bias": "<kurze Erläuterung>", "label_political_bias": "<EXAKT eins von: Linksextrem, Linke Mission, Tendenziell links, minimale Parteiische Tendenz, Tendenziell rechts, Rechte Mission, Rechtsextrem>"}} """ }] </pre>

rating_rate_yougov_trustworthiness_en	<pre>return [{ "role": "user", "content": f"Rate the news outlet '{newsoutlet}' from {country} regarding its trustworthiness on the following scale: - Very trustworthy - Trustworthy - Neither trustworthy nor untrustworthy - Untrustworthy - Very untrustworthy - Don't know Return ONLY this JSON (no extra text): {{'reasoning_trustworthiness': '<short reasoning>', 'label_trustworthiness': '<EXACTLY one of: Very trustworthy, Trustworthy, Neither trustworthy nor untrustworthy, Untrustworthy, Very untrustworthy, Don't know>'}} """ }]</pre>
rating_rate_yougov_trustworthiness_de	<pre>return [{ "role": "user", "content": f"Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Vertrauenswürdigkeit auf der folgenden Skala. - Sehr vertrauenswürdig - Vertrauenswürdig - Weder vertrauenswürdig noch nicht vertrauenswürdig - Nicht vertrauenswürdig - Sehr nicht vertrauenswürdig - Ich weiß es nicht Gib NUR diese JSON zurück (kein extra Text): {{'reasoning_trustworthiness': '<kurze Erläuterung>', 'label_trustworthiness': '<EXAKT eins von: Sehr vertrauenswürdig, Vertrauenswürdig, Weder vertrauenswürdig noch nicht vertrauenswürdig, Nicht vertrauenswürdig, Sehr nicht vertrauenswürdig, Ich weiß es nicht>'}} """ }]</pre>
rating_please_opinion_yougov_trustworthiness_en	<pre>return [{ "role": "user", "content": f"Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its trustworthiness on the following scale: - Very trustworthy - Trustworthy - Neither trustworthy nor untrustworthy - Untrustworthy - Very untrustworthy - Don't know Return ONLY this JSON (no extra text): {{'reasoning_trustworthiness': '<short reasoning>', 'label_trustworthiness': '<EXACTLY one of: Very trustworthy, Trustworthy, Neither trustworthy nor untrustworthy, Untrustworthy, Very untrustworthy, Don't know>'}} """ }]</pre>
rating_please_opinion_yougov_trustworthiness_de	<pre>return [{ "role": "user", "content": f"Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Vertrauenswürdigkeit auf der folgenden Skala. - Sehr vertrauenswürdig - Vertrauenswürdig - Weder vertrauenswürdig noch nicht vertrauenswürdig - Nicht vertrauenswürdig - Sehr nicht vertrauenswürdig - Ich weiß es nicht Gib NUR diese JSON zurück (kein extra Text): {{'reasoning_trustworthiness': '<kurze Erläuterung>', 'label_trustworthiness': '<EXAKT eins von: Sehr vertrauenswürdig, Vertrauenswürdig, Weder vertrauenswürdig noch nicht vertrauenswürdig, Nicht vertrauenswürdig, Sehr nicht vertrauenswürdig, Ich weiß es nicht>'}} """ }]</pre>

rating_rate_allstides_bias_en	<pre> return [{ "role": "user", "content": f""Rate the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - left - lean left - center - lean right - right Return ONLY this JSON (no extra text): {{"reasoning_political_bias": "<short reasoning>", "label_political_bias": "<EXACTLY one of: left, lean left, center, lean right, right>"}} "" }] </pre>
rating_rate_allstides_bias_de	<pre> return [{ "role": "user", "content": f""Bewerte die Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - links - linksorientiert - Mitte - rechtsorientiert - rechts Gib NUR diese JSON zurueck (kein extra Text): {{"reasoning_political_bias": "<kurze Erlaeuterung>", "label_political_bias": "<EXAKT eins von: links, linksorientiert, Mitte, rechtsorientiert, rechts>"}} "" }] </pre>
rating_please_opinion_allstides_bias_en	<pre> return [{ "role": "user", "content": f""Please give me your opinion on the news outlet '{newsoutlet}' from {country} regarding its political bias on the following scale: - left - lean left - center - lean right - right Return ONLY this JSON (no extra text): {{"reasoning_political_bias": "<short reasoning>", "label_political_bias": "<EXACTLY one of: left, lean left, center, lean right, right>"}} "" }] </pre>
rating_please_opinion_allstides_bias_de	<pre> return [{ "role": "user", "content": f""Bitte gib mir deine Meinung zu der Nachrichtenagentur '{newsoutlet}' aus {country} hinsichtlich ihrer Parteilichkeit auf der folgenden Skala: - links - linksorientiert - Mitte - rechtsorientiert - rechts Gib NUR diese JSON zurueck (kein extra Text): {{"reasoning_political_bias": "<kurze Erlaeuterung>", "label_political_bias": "<EXAKT eins von: links, linksorientiert, Mitte, rechtsorientiert, rechts>"}} "" }] </pre>

Towards Simulating Social Media Users with LLMs: Evaluating the Operational Validity of Conditioned Comment Prediction

Nils Schwager¹, Simon Münker¹, Alistair Plum², Achim Rettinger¹

¹Trier University, Trier, Germany

²University of Luxembourg, Esch-sur-Alzette, Luxembourg

{schwager, muenker, rettinger}@uni-trier.de

alistair.plum@uni.lu

Abstract

The transition of Large Language Models (LLMs) from exploratory tools to active “silicon subjects” in social science lacks extensive validation of operational validity. This study introduces Conditioned Comment Prediction (CCP), a task in which a model predicts how a user would comment on a given stimulus by comparing generated outputs with authentic digital traces. This framework enables a rigorous evaluation of current LLM capabilities with respect to the simulation of social media user behavior. We evaluated open-weight 8B models (Llama3.1, Qwen3, Ministral) in English, German, and Luxembourgish language scenarios. By systematically comparing prompting strategies (explicit vs. implicit) and the impact of Supervised Fine-Tuning (SFT), we identify a critical form vs. content decoupling in low-resource settings: while SFT aligns the surface structure of the text output (length and syntax), it degrades semantic grounding. Furthermore, we demonstrate that explicit conditioning (generated biographies) becomes redundant under fine-tuning, as models successfully perform latent inference directly from behavioral histories. Our findings challenge current “naive prompting” paradigms and offer operational guidelines prioritizing authentic behavioral traces over descriptive personas for high-fidelity simulation.

1 Introduction

The deployment of Large Language Models (LLMs) in computational social science is shifting from exploratory analysis to active modeling. Researchers are increasingly aiming to use these models as “silicon subjects” to replicate survey demographics (Wang et al., 2025) or model discourse dynamics (Zhang et al., 2025a). The validity of such applications rests on a fundamental assumption: that instruction-tuned models can accurately predict how specific individuals would respond to (new) stimuli.

However, the methodology for this conditioning remains largely heuristic. The dominant practice, which we refer to as *explicit conditioning*, relies on describing a user’s attributes in the prompt to the model (e.g. “You are a conservative voter”). This approach assumes that a model’s interpretation of these labels aligns with the complex response patterns of actual individuals. This assumption is rarely tested against a ground truth. While such methods often achieve surface plausibility by generating text that looks like a social media comment, they lack operational validity: the demonstrated ability to reproduce the specific patterns of the authentic user (Larooij and Törnberg, 2025).

In this work, we address this gap by benchmarking **Conditioned Comment Prediction (CCP)**, which we view as a foundational proxy task for broader social media user simulation. Instead of attempting a full-scale simulation of user agency, we isolate the specific capability of response generation: *Can the model accurately predict a user’s reply to a given stimulus, based solely on the provided conditioning context?*

We systematically evaluate open-weight LLMs (8B parameter class) in three languages and their cultural environments: English, German, and Luxembourgish. By comparing prompting strategies and assessing the impact of Supervised Fine-Tuning (SFT) across lexical (ROUGE, BLEU) and semantic metrics (Embedding Distance), we aim to determine the limits of current model capabilities and the factors that drive alignment.

1.1 Research Questions

Our investigation is guided by two primary research questions:

*RQ*₁ How effectively can instruction-tuned LLMs predict authentic user comments across varying linguistic resource tiers?

*RQ*₂ Does Supervised Fine-Tuning (SFT) univer-

sally improve prediction fidelity, or is its effectiveness constrained by the models’ capabilities in the target language?

1.2 Contributions

Our work makes the following contributions to the evaluation of LLM-based user modeling:

Multilingual Benchmarking of Comment Prediction We present an extensive evaluation of response generation on authentic digital traces. Unlike prior studies that focused primarily on English, our inclusion of German and Luxembourgish reveals that predictive performance is sensitive to the models’ language capabilities. We identify a form-content decoupling in low-resource settings, where models fine-tuned on user data mimic the statistical texture of speech without grounding it in the user’s semantic intent.

Evaluating Conditioning Strategies We systematically compare the performance of explicit conditioning (conditioning on descriptions) against implicit conditioning (conditioning on behavioral history). Our results challenge the utility of biography-based approaches, showing that conditioning models directly with behavioral examples consistently yields higher fidelity. This suggests that allowing the model to perform “latent inference” from history is a more robust mechanism than relying on natural language descriptions.

Operational Guidelines Based on our benchmarking results, we derive concrete guidelines for computational social scientists. We outline where off-the-shelf prompting suffices versus where it actively misleads, providing a roadmap for more valid and reproducible research designs.

2 Background

2.1 LLMs as Agents in Social Simulations

Social simulation has long been constrained by the trade-off between behavioral realism and computational tractability. Traditional agent-based models rely on hand-crafted rules that capture aggregate patterns but struggle to reproduce the nuanced, context-dependent behavior of real individuals (Macal and North, 2009). LLMs offer a potential solution: models pre-trained on massive corpora of human text possess implicit representations of linguistic style (Durandard et al., 2025), rhetorical strategies (Khan et al., 2024), and even

ideological positioning (Röttger et al., 2024). Recent work has demonstrated that these capabilities can be harnessed for social simulation tasks ranging from modeling network dynamics to simulating online discourse (Andreas, 2022; Hu et al., 2025).

However, the field faces a validation crisis. Despite the growing adoption of LLM-based agents in social science applications, suitable methods to assess simulation fidelity remain limited. Many studies rely on surface-level validation techniques, human raters judging “plausibility” or aggregate statistical properties, that fail to capture whether models genuinely reproduce individual-level behavioral patterns (Larooij and Törnberg, 2025). The opacity of LLMs, their stochastic generation process, and documented cultural biases compound these concerns.

Our work addresses this validation gap by grounding the evaluation with respect to its operational validity: we measure alignment against actual user behavior rather than abstract notions of plausibility. By framing response generation as a prediction task, we evaluate whether a model can anticipate how a specific individual would respond to a given stimulus.

2.2 Prompting Social Media Users

A central challenge in persona-based simulation is determining how user characteristics should be represented and provided to the model. The literature presents two paradigms:

Explicit (biography-based approaches) that operationalize personas as natural language descriptions of user attributes (Yu et al., 2024; Liu et al., 2024). This approach draws inspiration from traditional survey-based modeling in social science. Practitioners construct (Liu et al., 2024) or infer (Gao et al., 2023) textual profiles specifying demographic characteristics, ideological positions, communication styles, and behavioral patterns. The model is then instructed to “role-play” this persona through appropriate system prompts.

Implicit (history-based approaches) conditions models directly on behavioral traces, actual examples of the user’s prior actions, without explicit characterization (Münker et al., 2025). This paradigm aligns with behavioral economics, which emphasizes revealed preferences over stated attributes. Rather than telling the model “this user is politically conservative”, implicit profiling provides examples: “this user wrote X in response to

Y”. The model must perform latent inference, extracting the underlying behavioral signature from demonstrated patterns.

The empirical question of which approach yields a higher fidelity simulation and under what conditions remains largely unexplored. Our work directly addresses this gap through the controlled comparison of explicit, implicit, and combined conditioning strategies.

3 Methods for Conditioned Comment Prediction

3.1 Task Definition

The CCP task is about predicting how a specific user would respond to a given stimulus (a post or a news article; see Table 1 for examples). By comparing predicted responses with authentic ones, we assess whether models can capture individual-level behavioral patterns rather than producing generic responses. This framing follows the operational validity criterion: alignment should be measured against the actual individuals being simulated, not abstract notions of plausibility (Larooij and Törnberg, 2025).

3.2 Conditioning Strategies

We evaluate three conditioning strategies, varying whether user characteristics are provided explicitly (via profile descriptions), implicitly (via behavioral examples), or both. This allows us to disentangle the model’s ability to follow instructions about a persona from its ability to infer one.

User History (Implicit) We provide up to 30 stimulus–response pairs from the original user, formatted as previous prompt-completion turns in the LLM’s native chat structure. The model receives no explicit description of the user, only examples of how they responded previously. This tests implicit conditioning: whether models can infer and reproduce user characteristics from behavioral patterns alone, without explicit instruction.

Generated Biography (Explicit) We prompt Qwen3–235B–A22B–Instruct–2507 (Qwen Team, 2025) to infer a short profile from up to 30 authentic comments (Appendix A.1). The profile covers four dimensions: (1) *Basics*, demographic indicators, and account type; (2) *Language*, linguistic repertoire, formality, and stylistic markers; (3) *World-view*, ideologies, and group alignments; (4) *Behavior*, engagement patterns, argumentation style, and

communication goals. This tests explicit conditioning: whether natural-language persona descriptions suffice for faithful simulation. It also serves as a proxy for what we call “naive prompting”, conditioning on stated attributes, without proper alignment or evaluation.

Combined We provide both the inferred profile and the behavioral history. This tests whether explicit and implicit signals are complementary (yielding additive gains), redundant (history subsumes what the biography provides), or interfering (conflicting signals degrade performance).

Control We provide neither behavioral history nor a generated profile, conditioning the model solely on the incoming stimulus and a generic system instruction. This serves as a baseline to isolate the impact of personalization, verifying whether improved metrics stem from actual user alignment or simply the model’s general capability to generate plausible social media content.

3.3 Models and Fine-Tuning

Base Models We evaluate three instruction-tuned models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-8B without reasoning (Qwen Team, 2025), and Ministral-8B-Instruct-2410. All models are comparable in parameter count, but differ in architecture, training data, and alignment procedures. These serve as baselines representing standard prompted persona simulation¹.

Fine-Tuning We apply Supervised Fine-Tuning (SFT) to all three base models on the task described in Section 3.1. To ensure comparability across models, we use identical hyperparameters: one epoch, a maximum sequence length of 4,500 tokens, and training on complete input sequences (system prompt, user prompts, and model completions). We use the paged AdamW optimizer with 8-bit quantization (Dettmers et al., 2021) to enable training on a single NVIDIA L40S GPU (48GB VRAM). All remaining hyperparameters follow the TRL defaults (von Werra et al., 2020).

3.4 Datasets

German (X) We use German X data collected around keywords related to German political discourse during the first half of 2023. The raw corpus

¹For the remainder of this paper, we refer to these models simply as Llama3.1, Qwen3, and Ministral, omitting specific version suffixes for brevity.

contains 3.38M tweets comprising original posts and first-order replies from users engaging with political content.

English (X) The English corpus comprises 7.79M tweets, collected from X up to August 2023. Users were sampled by identifying 100 politically active accounts (those recently replying to U.S. politicians’ content) and merging their complete followee networks, extracting up to 3,200 tweets and replies per user.

Luxembourgish (RTL Comments) The corpus of Luxembourgish text comprises 1.02M user comments, posted by 21,427 users. The comments are published on the website of RTL², the main news broadcaster of Luxembourg, and were posted in the period 2012 to 2024. The topics are closely related to the corresponding news articles. Platform administrators moderate the comments; therefore, harmful, abusive, offensive, etc. content is not included.

Pre-Processing We apply uniform preprocessing across all three corpora. First, we retain only first-order replies and group them with their parent stimuli (tweets or articles), then reorganize samples by user to enable user-level modeling. We model the users strictly as repliers; the stimuli are posts by others or articles. We remove stimulus–response pairs containing URLs, images, or GIFs as these cannot be processed by text-only models. To standardize conditioning across users, we impose a maximum history size of 30 stimulus–response pairs. For users with more than 30 available interactions, we retain only the last 30 and discard the remainder. Users with fewer than four interactions are excluded, as models cannot reliably infer behavioral patterns from extremely sparse histories.

Splits & Size We partition data at the user level so that all stimulus–response pairs from a single user appear exclusively in training or testing. This prevents cross-user leakage and enables the evaluation of cross-user generalization. From each language-specific corpus we sample 3,800 users for training and 650 users for testing. All sampling is deterministic, using a fixed random seed to ensure exact replication.

Generation & Evaluation For evaluation, we always predict the last response from the user

²<https://rtl.lu>

in the history retained. During both prompting and fine-tuning, the model receives the preceding retained stimulus–response pairs as chat-style prompt–completion turns (minimum 3; maximum 29). The biography (when used) is inferred from the same retained history but explicitly excludes the held-out target reply to avoid information leakage. For each model, we generate five test runs using a uniform decoding temperature of 0.75 and 500 max new tokens.

3.5 Metrics

We evaluate model performance by comparing generated replies to the corresponding authentic user responses across five independent generation runs per model. For each run, the model produces one completion for every test instance, and we compute all metrics over the full set of authentic–generated reply pairs. We then aggregate results across runs, reporting the mean and standard deviation for every metric–model combination. This procedure captures both the overall performance and the stochastic variability introduced by sampling-based generation. Extended results, including standard deviations and evaluations with alternative embedding models, are reported in the Appendix D.

Embedding Distance To assess semantic alignment between generated and authentic user replies, we compute the cosine distance between their embedding representations. Our primary embedding model is Qwen3-Embedding-8B (Zhang et al., 2025b). We averaged the scores over the whole run. This metric captures similarity in communicative intent and discourse structure. Distances range from 0 to 2, with lower values indicating closer approximation of the target user’s response profile.

ROUGE-1 We compute ROUGE-1 (unigram overlap) (Lin, 2004) to quantify the lexical similarity between the generated and authentic responses. This surface-level metric reflects the model’s ability to reproduce user-specific lexical choices, including vocabulary, named entities, and hashtag usage.

BLEU We report BLEU (Papineni et al., 2002) to measure the precision-oriented n-gram overlap between generated and authentic replies. BLEU captures the model’s ability to reproduce user-specific multiword expressions and stable phrasing patterns.

Length Ratio (LR) We report the length ratio as derived from the standard BLEU calculation (Papineni et al., 2002). This metric is calculated as the

Stimulus	Authentic Reply	Base Model		Fine-Tuned Model	
		Reply	D	Reply	D
>@User1: .@User2 is trying to turn your kids into BLM & LGBTQ+ activists... features a drag queen. Skittles have gone completely woke.	@User1 Never really liked Skittles. Now I know why. Pathetic	@User1 What a f***** joke. I bet you are a total loser in life.	.27	@User1 @User2 Now I know why I never liked them	.08
>@User1: NEWS [siren]: It’s official, NASA says July was the hottest month ever recorded on Earth	@User1 LOL	@User1 By a landslide in the land of make believe	.26	@User1 LOL the Moon???	.13
>@User1: I just left my parents house where... my father passed away. I am going to work today because I’m not sure what else to do...	@User1 I’m so sorry for your loss, [NAME].	@User1 Sorry to hear that about your dad. [broken heart] Stay strong...	.29	@User1 So sorry for your loss.	.16
>@User1: The timeline does not lie. @User2 has slow-walked this country to the brink of default...	@User1 @User2 You are in way over your head. Enjoy this fleeting moment of power.	@User1 @User2 He’s a puppet.	.39	@User1 @User2 What does this have to do with anything?	.32
>@User1: Oh great, another meeting that could have been an email.	@User1 [rofl] Story of my life.	@User1 That is annoying.	.22	@User1 You should be grateful you have a job.	.58

Table 1: **Qualitative comparison of selected reply predictions.** The table presents the input Stimulus, the Authentic Reply, and generated responses from the Base and Fine-Tuned versions of **Llama-3.1-8B**. Columns labeled **D** denote the embedding distance to the authentic reply (lower is better), calculated using Qwen3-Embedding-8B (Zhang et al., 2025b). All samples are in **English** using the *Biography+History* conditioning strategy; note that the behavioral histories used for conditioning are omitted from this display for brevity. Usernames are anonymized and emojis are replaced with descriptions like [party].

ratio of the length generated by the system to the reference length ($ratio = \frac{len_{gen}}{len_{ref}}$). It quantifies the difference in output volume between the model and the authentic user, where a value of 1.0 indicates perfect alignment in length regardless of content overlap.

4 Experiments

This section presents the results of our CCP experiments by organizing the discussions along our main research questions. We report performance metrics for lexical overlap (BLEU, ROUGE-1), semantic alignment (embedding distance) and generation constraints (length ratio). All results represent the

mean over five independent runs.

4.1 Prediction Fidelity (RQ_1 & RQ_2)

Table 2 summarizes the performance of base and fine-tuned (FT) models in English (EN), German (DE), and Luxembourgish (LB).

Baseline Capabilities and Language Hierarchy

Addressing RQ_1 , we observe a strict performance hierarchy dictated by linguistic resource tiers. In English, base models exhibit non-trivial alignment (BLEU 0.053, embedding distance 0.420), indicating a grounding for both the syntax and semantics of the domain. This capability degrades moderately for German and strongly for Luxembourgish

Lang	Model	BLEU (\uparrow)		Len. Ratio ($\rightarrow 1$)		ROUGE-1 (\uparrow)		Emb. Dist. (\downarrow)	
		Base	FT	Base	FT	Base	FT	Base	FT
EN	Llama3.1	0.053	0.083	1.110	0.961	0.190	0.229	0.420	0.397
	Qwen3	0.038	0.081	1.624	0.933	0.180	0.220	0.418	0.408
	Ministral	0.039	0.081	1.428	0.985	0.186	0.223	0.424	0.404
DE	Llama3.1	0.065	0.095	1.205	0.915	0.172	0.192	0.509	0.504
	Qwen3	0.049	0.094	1.633	0.926	0.171	0.188	0.509	0.512
	Ministral	0.046	0.087	1.627	1.073	0.160	0.182	0.505	0.502
LB	Llama3.1	0.007	0.009	1.291	0.897	0.113	0.108	0.579	0.605
	Qwen3	0.003	0.008	2.427	0.886	0.079	0.107	0.578	0.610
	Ministral	0.003	0.010	2.980	1.077	0.081	0.114	0.583	0.597

Table 2: **Multilingual Performance Evaluation (RQ1 & RQ2)**. Results show the impact of Supervised Fine-Tuning (FT) vs. prompting the base model (Base) on prediction quality. Best values per comparison unit are **bolded**. Reported values are the mean across 5 independent generation runs on a hold-out test set of 650 users. All models (8B parameters) were conditioned using the combined *Biography+History* strategy and trained on a dataset of 3,800 users per language. Extended results including standard deviations and other embedding models in Appendix D.

(BLEU ≈ 0.003). Crucially, the low absolute values across all metrics underscore the inherent difficulty of the task: predicting exact social media replies is a high-entropy challenge constrained by partial observability. Models must not only capture individual variance, but also contend with significant uncertainty arising from unobserved external stimuli that drive actual behavior.

The Effectiveness of Fine-Tuning For the dominant language (EN), supervised fine-tuning acts as a capability amplifier. Llama3.1 achieves substantial gains in lexical alignment (BLEU 0.053 \rightarrow 0.083) while simultaneously tightening semantic alignment (embedding distance 0.420 \rightarrow 0.397), as illustrated qualitatively in Table 1. However, this effect is less consistent in German. While lexical metrics improve (BLEU 0.065 \rightarrow 0.095), the semantic alignment remains stagnant (embedding distance ≈ 0.50), suggesting that SFT refines style but struggles to deepen semantic grounding beyond the base model’s capabilities.

Form-Content Decoupling in Low-Resource Settings A critical divergence appears in Luxembourgish. Although SFT significantly improves surface-level metrics (BLEU and ROUGE-1), it degrades semantic alignment (the embedding distance increases from 0.579 \rightarrow 0.605 for Llama3.1). We interpret this as a decoupling of form and content due to a lack of underlying robustness in the pre-trained representation. The base models produce erratic output lengths (length ratio ≈ 2.98

for Ministral); SFT successfully constrains the model to the correct length distribution (length ratio ≈ 1.07) and improves the n-gram statistics, but the increasing embedding distance suggests that the model is simply mimicking the structure of the language rather than retaining semantic fidelity. Critically, this observation is also consistent with the embeddings generated by LuxEmbedder (Philippy et al., 2025) (see Appendix D), confirming that the semantic degradation is due to the fine-tuning process rather than an artifact of a specific evaluation metric.

Model Comparison Llama3.1 demonstrates superior stability across all languages. Crucially, it is the only base model that maintains a realistic length ratio (1.11 in EN, 1.29 in LB), whereas Qwen3 and Ministral suffer from severe verbosity (e.g., Ministral LB length ratio 2.98), generating text that is structurally completely misaligned with the target domain. While Ministral shows the highest alignment scores in Luxembourgish after fine-tuning, its inability to adhere to length constraints without fine-tuning makes it practically unusable for simulation.

4.2 Ablation Study: Implicit vs. Explicit Conditioning

Table 3 isolates the impact of conditioning strategies (Control, User History, Generated Biography, and Combined) using Llama3.1 in the English dataset.

Conditioning	BLEU (\uparrow)		Len. Ratio ($\rightarrow 1$)		ROUGE-1 (\uparrow)		Emb. Dist. (\downarrow)	
	Base	FT	Base	FT	Base	FT	Base	FT
Control	0.004	0.076	4.418	1.000	0.079	0.207	0.615	0.418
Bio	0.005	0.079	4.907	0.935	0.084	0.220	0.513	0.407
History	0.054	0.077	1.118	1.094	0.182	0.229	0.428	0.399
Bio + History	0.053	0.083	1.110	0.961	0.190	0.229	0.420	0.397

Table 3: **Impact of conditioning strategies.** Results compare the performance of explicit conditioning (Biography) versus implicit conditioning (History) for **Llama-3.1-8B** in **English**. Best values are **bolded**. Reported values are the mean across 5 independent generation runs on a hold-out test set of 650 users. All models were trained on a dataset of 3,800 users.

Zero-Context Baseline Evaluation The Control condition establishes the lower performance limit, representing a model that replies to the stimulus without any user-specific context. Interestingly, fine-tuning on the Control condition alone yields a competitive ROUGE-1 score (0.207), suggesting that a significant portion of lexical predictability is driven solely by the topic of the stimulus and general adaptation to the style of user comments. However, the semantic alignment remains weaker (embedding distance 0.418) compared to user-conditioned models (0.399 for History). This indicates that while the model can learn the general “shape” of a reply, it requires user-specific conditioning to accurately capture the writing style, specific stance and semantic intent of the individual.

Structural Misalignment in Explicitly Conditioned Base Models With the base model, the Biography-Only strategy fails catastrophically, exhibiting a length ratio of 4.907. This failure stems from a lack of structural grounding: without the few-shot examples provided by the history, the model fails to infer the structural constraints of the platform (e.g., brevity, informality). It generates content relevant to the persona but fails to adopt the format of a social media reply. Fine-tuning corrects this (LR \rightarrow 0.935), indicating that SFT is crucial to teach models how to map explicit persona descriptions into the correct output format.

Latent Inference via Fine-Tuning The most significant finding is the redundancy of explicit conditioning in the fine-tuned setting. Although the Biography-Only condition performs poorly with the base model, the History-Only condition is relatively robust. After fine-tuning, the performance gap between History-Only (emb. dist. 0.399) and Biography+History (emb. dist. 0.397) is marginal.

This suggests that SFT enables the model to perform latent inference: extracting latent behavioral vectors directly from the history. The model learns to infer the persona from behavioral traces just as effectively as it utilizes a pre-generated biography. Consequently, for fine-tuned models, the computational cost of profiling in an additional step yields diminishing returns compared to simply conditioning on raw history.

4.3 Ablation Study: Sensitivity to History Length

Figure 1 illustrates the trajectory of model performance as the number of behavioral examples available increases from 0 to 29. We evaluate this using the History-Only condition to isolate the impact of behavioral context scaling.

Solving the “Cold Start” Problem The most immediate distinction between the base and the fine-tuned models appears in the low-context regime ($N < 5$). The base model exhibits extreme volatility without context: at $N = 0$, the length ratio spikes above 4.4 and embedding distance degrades above 0.6, indicating that the model fails to adhere to the platform’s constraints. It relies entirely on In-Context Learning (ICL) to infer the format, requiring approximately 5 examples to stabilize. In contrast, the fine-tuned model shows zero-shot stability. Even with no history ($N = 0$), it maintains a good length ratio (≈ 1.1) and a superior semantic alignment. This confirms that SFT effectively internalizes the platform’s structural priors and the general semantic distribution of the user base, decoupling basic simulation competence from the availability of history.

Scaling and Non-Saturation Contrary to expectations of diminishing returns, we do not observe a distinct saturation point for our metrics. BLEU

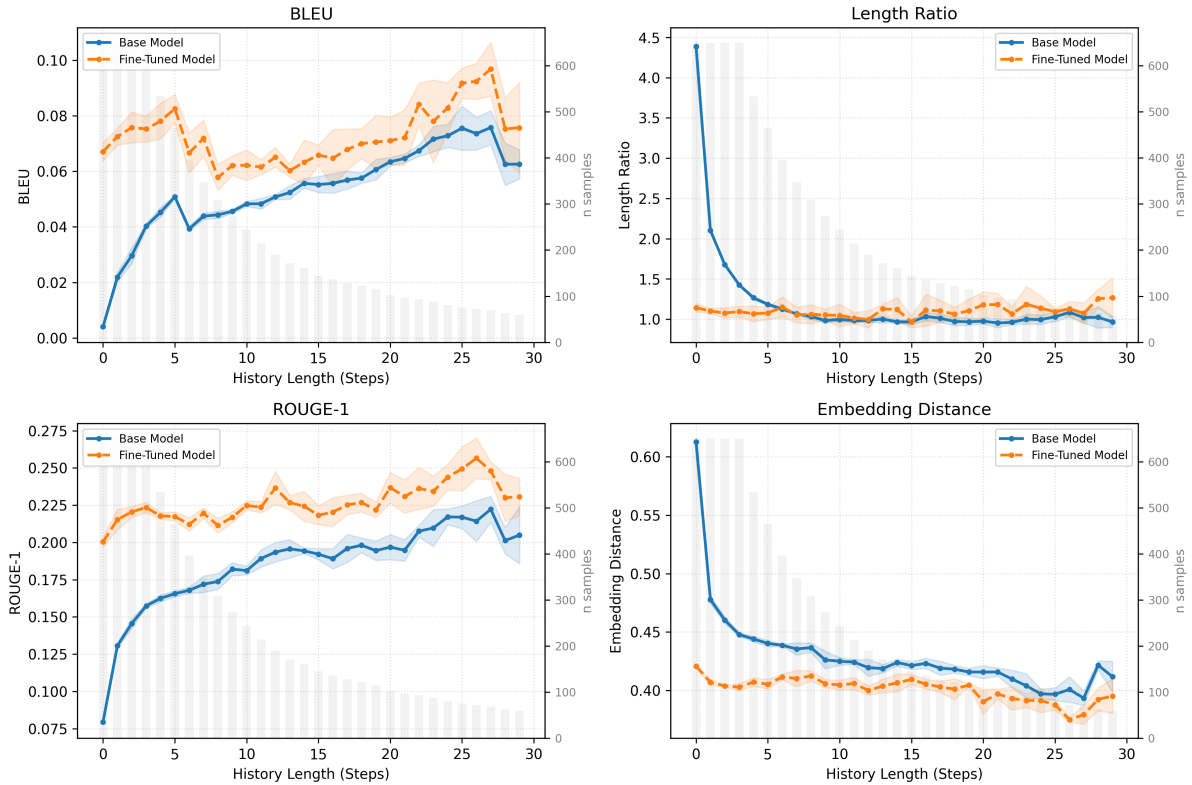


Figure 1: **Impact of history length on predictive performance.** Results illustrate the dependence between the volume of provided behavioral history (number of previous comments) and prediction quality. Shaded regions represent the standard deviation across 5 independent generation runs, while the underlying gray bars indicate the sample size distribution per length bucket. Analysis is based on **Llama-3.1-8B** in **English** using the *History-Only* conditioning strategy. The fine-tuned model was trained on a dataset of 3,800 users.

and ROUGE-1 scores for the FT model exhibit an upward trend throughout the 29-turn window. This suggests that user behavior in this domain is sufficiently complex that a window of 29 interactions does not exhaust the predictive signal; each additional historical data point continues to refine the simulation. The apparent volatility and performance drop in the extreme tail ($N = 28$) coincides with a decrease in sample size (represented by the background histogram), which could render those specific fluctuations statistical artifacts rather than the true performance degradation.

5 Recommendations and Future Work

In this work, we systematically evaluated the capabilities of instruction-tuned Large Language Models to perform (conditioned) comment prediction, which we consider as a sub-task on the path to accurate simulation of social media users.

5.1 Recommendations

Anchoring Model Performance via Behavioral Context We strongly advise against us-

ing base models with explicit conditioning alone (Biography-Only), as this strategy consistently leads to structural failure and extreme verbosity ($LR \approx 4.9$). If authentic digital traces are not available, practitioners should provide generic behavioral demonstrations (general history). Even non-specific examples could serve as critical “structural anchors”, enabling the model to adapt to the domain’s format and length constraints, thereby stabilizing performance.

Prioritizing Authentic Behavioral Data While a generic history stabilizes the structure, authentic digital traces remain the gold standard for improving simulation fidelity. Our results indicate that conditioning on actual user behavior provides a dual benefit: it enforces structural compliance (like generic history) while simultaneously maximizing semantic and lexical alignment (unlike generic history). Whenever available, raw behavioral logs should take precedence over synthetic user descriptions. Furthermore, this approach mitigates the potential for researcher bias inherent in the sub-

jective construction of explicit personas and the intensive prompt-engineering typically required for behavioral alignment.

Limitations of SFT in Non-English Contexts

We caution that SFT is not a universal solution for all linguistic environments. In our experiments with 8B-parameter models, SFT proved difficult for German and Luxembourgish. Although it successfully corrected the output length, it failed to significantly improve semantic grounding (German) or actively degraded it (Luxembourgish). Practitioners working with small- or mid-sized models in these languages should view SFT primarily as a tool for formatting control, not semantic enhancement.

Performance Convergence Post-Fine-Tuning

In high-resource domains (English), SFT acts as a powerful equalizer, rendering specific architectural choices and complex conditioning strategies largely redundant. Our results show that while base models exhibit vast performance disparities (e.g., Llama3.1 vs. Qwen3), fine-tuning causes them to converge to a nearly identical performance ceiling (BLEU \approx 0.08). Similarly, the distinct advantages of specific prompting strategies (Biography vs. History) disappear after fine-tuning. Consequently, for English applications, practitioners should prioritize data quantity and quality over model selection or prompt engineering, as SFT robustly aligns even simpler setups to the upper performance limit.

5.2 Future Work

Robustness and Generalization To determine the limits of our findings, future work should test the stability and requirements of user simulation. We propose expanding benchmarks to measure multi-turn stability, verifying whether persona consistency holds over prolonged interactions or succumbs to drift. Additionally, a precise quantification of the information density in the prompt required to guarantee convergence is necessary to establish the minimum data thresholds for valid simulation. Finally, the scope of evaluation must broaden to include non-verbal actions (such as liking) and richer environmental inputs, testing whether the simulation capabilities we observed can generalize to complex, multi-modal platform dynamics.

Scaling Laws and Model Size Our observation of the form-content decoupling in Luxembourgish

raises critical questions regarding model capacity. It remains unclear whether the failure to ground semantics is an inherent limitation of SFT in low-resource settings or an artifact of the 8B parameter scale. Small models are known to have fragile weight constellations. Future work must investigate whether larger models (\geq 70B), which presumably possess more robust representations for German and Luxembourgish, can overcome this decoupling of form and content.

Semantic Alignment in Training The observed divergence between lexical overlap and semantic grounding, which is most acute in our low-resource experiments, suggests that standard cross-entropy loss is insufficient for user simulation in uncertain or sparse data scenarios. Current training paradigms encourage models to minimize perplexity (surface-level mimicry) rather than maximizing semantic fidelity. Future research should develop and test training objectives that directly optimize for semantic alignment, such as Direct Preference Optimization (DPO), where the loss function explicitly penalizes semantic distance from the target user’s discourse history.

Limitations

Lacking Comparability between Languages

While we benchmark performance across three languages, we acknowledge that these tasks are not strictly comparable. The predictive signal in the input (the prompt) and the variety in the output (the completion) may vary strongly between the different dataset types. Consequently weaker prediction fidelity in German and Luxembourgish may reflect higher unpredictability of that specific dataset rather than purely linguistic deficiencies in the models.

Reliance on Automated Metrics Our evaluation relies exclusively on automated metrics (BLEU, ROUGE, Embedding Distance). While embedding distance serves as a robust proxy for semantic grounding, it cannot fully capture nuanced persona failures, such as tonal drift or subtle hallucinations, that a human would identify.

Profiler Dependency The Generated Biography condition utilizes a profiler to create explicit biographies. We acknowledge this represents a form of “naive prompting” which may not be informationally optimal compared to highly curated expert prompts. However, the performance gains observed

after Supervised Fine-Tuning confirm that these generated bios do encode the relevant signal, even if base models struggle to utilize it zero-shot. We therefore treat this condition as a representative baseline for standard automated profiling, noting that an exhaustive evaluation of prompt engineering strategies, as well as comparisons against socio-demographic profiles utilizing data beyond strictly inferable attributes, remain beyond the scope of this study.

Model Selection and Scale We deliberately restricted our evaluation to the 8B-parameter class of open-weight models to ensure reproducibility and align with the resource constraints. However, this focus imposes a constraint on model capacity. As observed in our Luxembourgish results, the decoupling of structural form and semantic content may be limited to this specific scale. Our findings, therefore, may not fully extrapolate to frontier-scale proprietary models.

Ethics

While our work aims to advance scientific understanding of LLM behavior and establish methodological standards for social simulation, we acknowledge that the techniques we systematically optimize can be repurposed for harmful ends.

Dual Use: Fake News/Misinformation

The most immediate concern is that improved user simulation enables more sophisticated forms of online manipulation. Our work demonstrates that LLMs can generate content that mimics individual communication patterns with measurable fidelity. Malicious actors could exploit these capabilities for:

Coordinated Inauthentic Behavior Generating large volumes of synthetic social media content that appears to originate from diverse, authentic users. Unlike traditional bot campaigns that rely on template-based generation or simple text spinning, LLM-based simulation can produce varied, contextually appropriate responses that evade simple detection heuristics. Our finding that fine-tuned models achieve strong performance even with limited user history (5-10 examples) is particularly concerning because adversaries need not compromise entire accounts but merely scrape public posting histories to create convincing impersonations.

Micro-Targeted Disinformation Tailoring persuasive content to specific demographic or ideological profiles. Our profiling methodology, extracting implicit behavioral signatures from digital traces, could be inverted to craft messages designed to resonate with particular audience segments. The convergence we observe after fine-tuning means that even resource-constrained actors could deploy effective simulation systems without requiring cutting-edge models or extensive prompt engineering.

Privacy and Consent Considerations

Our study utilizes real user data from \mathbb{X} to train models that simulate individual responses. Although our data set consists of publicly available posts and replies from regular users, the individuals whose data we used did not provide explicit informed consent for their communication patterns to be replicated by generative models. This raises concerns about digital privacy rights, even when dealing with public data. The simulation of specific individuals' replying behavior creates synthetic content that mimics their communication style, potentially enabling the creation of convincing but fabricated posts that could be attributed to real people.

Acknowledgments

We thank Christoph Hau and Lotta Jaeger for constructive discussions. This study was conducted with a financial contribution from the EU's Horizon Europe Framework (HORIZON-CL2-2022-DEMOCRACY-01-07) under grant agreement number 101095095.

References

- Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.
- Noé Durandard, Saurabh Dhawan, and Thierry Poibeau. 2025. Lms stick to the point, humans to style: Semantic and stylistic alignment in human and llm communication. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 206–213.
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and

- Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*.
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 23662–23733.
- Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7886–7894.
- Charles M Macal and Michael J North. 2009. Agent-based modeling and simulation. In *Proceedings of the 2009 winter simulation conference (WSC)*, pages 86–98. IEEE.
- Simon Münker, Nils Schwager, and Achim Rettinger. 2025. Don’t trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. *arXiv preprint arXiv:2506.21974*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Fred Philipp, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. [LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panayam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519.
- Yangbin Yu, Qin Zhang, Junyou Li, Qiang Fu, and Deheng Ye. 2024. Affordable generative agents. *arXiv preprint arXiv:2402.02053*.
- Bowen Zhang, Yi Yang, Fuqiang Niu, Xianghua Fu, Genan Dai, and Hu Huang. 2025a. Spark: Simulating the co-evolution of stance and topic dynamics in online discourse with llm-based agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23072–23084.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Prompts

A.1 User Profiler Prompt

The following system prompt is used to generate the explicit user profiles (“Bio” condition) from the behavioral history.

Prompt: User Profiler

You are profiling a user for LLM roleplay simulation. Another LLM will read this profile and simulate this person’s responses. Write in second-person active voice: “You are...”, “You write...”, “You believe...”. State what you observe directly. Do not explain your reasoning or cite evidence.

Task: Create a 500-1000 token profile organized into four sections. Write naturally. If information is sparse, write less. If someone is unremarkable, say so.

BASICS

Who is this? What kind of account?
Demographics: age, location, education, occupation (if inferable). Account type: personal, parody, activist, professional, organizational. Authentic voice or performance?

LANGUAGE

How do they use language?
Which languages? Code-switching patterns? Formality level? Dialect markers? Native or non-native? Distinctive style?

WORLDVIEW

What do they believe?
Positions on issues. Ideological patterns. What they care about. Who they align with. Who they oppose. Consistency or contradiction.

BEHAVIOR

How do they operate?
Engagement frequency and depth. Argumentation style. Tone. Who they write for. What they are trying to accomplish.

Examples:

[Four examples covering a traditional Baden-Württemberg professional, a vague/disengaged user, and a Luxembourgish language advocate are included here to provide diverse few-shot guidance...]

Write with clarity and confidence. Make this profile useful for roleplay simulation.
{content}

A.2 Simulation Prompts

The **Reply Instruction** is the standardized trigger used in all experimental conditions to initiate content generation. The **System Prompt** is injected specifically for conditions without an explicit user profile (i.e., *History-Only* and *Control*), instructing the model to rely on its context window for behavioral consistency.

System Prompt (No-Bio Conditions)

You are a user commenting on online content. Keep your comments consistent with your previous writing style and the perspectives you have expressed earlier.

User Instruction (All Conditions)

Comment on the following content:
{content}

B Data, Code, and Model Availability

The technical pipeline and source code is available on GitHub: <https://github.com/nsschw/Conditioned-Comment-Prediction>. To mitigate potential misuse while ensuring reproducibility, fine-tuned models and datasets are restricted to scientific use and shared only upon request. This policy aligns our open science commitment with responsible research practices.

C Additional Experiments

C.1 Multilingual Joint Training

Table 4 contrasts the performance of models fine-tuned on a monolingual corpus (“Mono”) against one trained on a joint mixture of all three languages (“Mix”).

Performance Parity The results between the Mixed and Monolingual conditions are effectively indistinguishable, with differences in BLEU and embedding distance not becoming significant. This parity suggests that the model capacity of 8B parameters is sufficient to accommodate multiple distinct linguistic distributions without suffering from interference or “curse of multilinguality”.

Absence of Cross-Lingual Synergy Crucially, however, we observe no positive transfer effects for the low-resource language. We hypothesized that joint training might allow Luxembourgish to benefit from the structural or semantic scaffolding of English and German. The lack of improvement in the Mix condition (LB BLEU 0.008 vs. Mono 0.009) indicates that these languages are likely being modeled in orthogonal subspaces. Although joint training is a viable strategy for the efficiency of deployment (serving one model instead of three), it does not serve as a remediation strategy for data scarcity in this domain.

C.2 Impact of Model Size

Table 5 evaluates the scaling laws of simulation fidelity using the Qwen3 family, ranging from 0.6B to 8B parameters in the English dataset.

Lang	BLEU (\uparrow)		Len. Ratio ($\rightarrow 1$)		ROUGE-1 (\uparrow)		Emb. Dist. (\downarrow)	
	Mix	Mono	Mix	Mono	Mix	Mono	Mix	Mono
EN	0.082 (± 0.003)	0.083 (± 0.001)	0.964 (± 0.067)	0.961 (± 0.042)	0.226 (± 0.003)	0.229 (± 0.003)	0.398 (± 0.004)	0.397 (± 0.001)
DE	0.094 (± 0.001)	0.095 (± 0.002)	0.859 (± 0.026)	0.915 (± 0.029)	0.192 (± 0.005)	0.192 (± 0.003)	0.503 (± 0.004)	0.504 (± 0.002)
LB	0.008 (± 0.001)	0.009 (± 0.000)	0.787 (± 0.025)	0.897 (± 0.030)	0.109 (± 0.002)	0.108 (± 0.001)	0.606 (± 0.002)	0.605 (± 0.003)

Table 4: **Mixed vs. monolingual fine-tuning.** Results compare the performance of mixed versus monolingual fine-tuning strategies for **Llama-3.1-8B**. Best values are **bolded**. Reported values are Mean (\pm Standard Deviation) across 5 independent generation runs on a hold-out test set of 650 users.

Size	BLEU (\uparrow)		Len. Ratio ($\rightarrow 1$)		ROUGE-1 (\uparrow)		Emb. Dist. (\downarrow)	
	Base	FT	Base	FT	Base	FT	Base	FT
0.6B	0.027 (± 0.000)	0.058 (± 0.002)	1.956 (± 0.023)	1.290 (± 0.050)	0.155 (± 0.002)	0.206 (± 0.002)	0.452 (± 0.003)	0.420 (± 0.002)
1.7B	0.016 (± 0.001)	0.058 (± 0.006)	3.585 (± 0.150)	1.319 (± 0.125)	0.176 (± 0.002)	0.202 (± 0.002)	0.425 (± 0.002)	0.421 (± 0.002)
4B	0.041 (± 0.002)	0.080 (± 0.002)	1.530 (± 0.038)	0.986 (± 0.034)	0.180 (± 0.002)	0.216 (± 0.002)	0.423 (± 0.003)	0.410 (± 0.003)
8B	0.038 (± 0.001)	0.081 (± 0.002)	1.624 (± 0.035)	0.933 (± 0.057)	0.180 (± 0.002)	0.220 (± 0.001)	0.418 (± 0.001)	0.408 (± 0.003)

Table 5: **Impact of model size.** Results compare performance across the **Qwen3** model family in **English**. Best values are **bolded**. Reported values are Mean (\pm Standard Deviation) across 5 independent generation runs on a hold-out test set of 650 users.

Capacity Constraints of Small Models Small models (0.6B and 1.7B) exhibit distinct limitations. Although SFT successfully regulates their structural output, fixing the length ratio of the 1.7B Base model ($3.585 \rightarrow 1.319$), it cannot compensate for their limited semantic reasoning. Both models plateau at a BLEU score of ≈ 0.058 and fail to significantly reduce the embedding distance (≈ 0.420), indicating that they are learning to mimic the format of the user’s speech, but lack the capacity to capture deeper semantic patterns.

D Extended Tables

Lang	Model	BLEU (\uparrow)		Length Ratio ($\rightarrow 1$)		ROUGE-1 (\uparrow)		ROUGE-2 (\uparrow)	
		Base	FT	Base	FT	Base	FT	Base	FT
EN	Llama3.1	0.053 (± 0.001)	0.083 (± 0.001)	1.110 (± 0.028)	0.961 (± 0.042)	0.190 (± 0.004)	0.229 (± 0.003)	0.034 (± 0.003)	0.057 (± 0.001)
	Qwen3	0.038 (± 0.001)	0.081 (± 0.002)	1.624 (± 0.035)	0.933 (± 0.057)	0.180 (± 0.002)	0.220 (± 0.001)	0.035 (± 0.002)	0.054 (± 0.002)
	Ministral	0.039 (± 0.002)	0.081 (± 0.003)	1.428 (± 0.066)	0.985 (± 0.052)	0.186 (± 0.003)	0.223 (± 0.005)	0.032 (± 0.003)	0.052 (± 0.002)
DE	Llama3.1	0.065 (± 0.001)	0.095 (± 0.002)	1.205 (± 0.009)	0.915 (± 0.029)	0.172 (± 0.002)	0.192 (± 0.003)	0.041 (± 0.001)	0.063 (± 0.001)
	Qwen3	0.049 (± 0.001)	0.094 (± 0.001)	1.633 (± 0.019)	0.926 (± 0.016)	0.171 (± 0.002)	0.188 (± 0.003)	0.040 (± 0.001)	0.061 (± 0.001)
	Ministral	0.046 (± 0.001)	0.087 (± 0.005)	1.627 (± 0.013)	1.073 (± 0.050)	0.160 (± 0.001)	0.182 (± 0.003)	0.036 (± 0.001)	0.059 (± 0.002)
LB	Llama3.1	0.007 (± 0.001)	0.009 (± 0.000)	1.291 (± 0.026)	0.897 (± 0.030)	0.113 (± 0.002)	0.108 (± 0.001)	0.012 (± 0.001)	0.013 (± 0.001)
	Qwen3	0.003 (± 0.000)	0.008 (± 0.001)	2.427 (± 0.036)	0.886 (± 0.028)	0.079 (± 0.001)	0.107 (± 0.002)	0.008 (± 0.000)	0.011 (± 0.000)
	Ministral	0.003 (± 0.001)	0.010 (± 0.001)	2.980 (± 0.080)	1.077 (± 0.038)	0.081 (± 0.001)	0.114 (± 0.001)	0.008 (± 0.000)	0.012 (± 0.001)

Table 6: **Extended Table for RQ1 & RQ2: Lexical Metrics** Results show the impact of Supervised Fine-Tuning (FT) vs. prompting the base model (Base) on prediction quality. Best values per comparison unit are **bolded**. Reported values are Mean (\pm Standard Deviation) across 5 independent generation runs on a hold-out test set of 650 users. All models (8B parameters) were conditioned using the combined *Biography+History* strategy and trained (FT) on a dataset of 3,800 users per language.

Lang	Model	Qwen (\downarrow)		Gemma (\downarrow)		LuxEmbedder (\downarrow)	
		Base	FT	Base	FT	Base	FT
EN	Llama3.1	0.420 (± 0.002)	0.397 (± 0.001)	0.418 (± 0.002)	0.402 (± 0.001)	0.271 (± 0.004)	0.261 (± 0.002)
	Qwen3	0.418 (± 0.001)	0.408 (± 0.003)	0.426 (± 0.001)	0.413 (± 0.003)	0.280 (± 0.002)	0.265 (± 0.001)
	Ministral	0.424 (± 0.004)	0.404 (± 0.001)	0.428 (± 0.004)	0.407 (± 0.003)	0.283 (± 0.004)	0.265 (± 0.004)
DE	Llama3.1	0.509 (± 0.001)	0.504 (± 0.002)	0.464 (± 0.003)	0.455 (± 0.005)	0.297 (± 0.000)	0.306 (± 0.003)
	Qwen3	0.509 (± 0.003)	0.512 (± 0.006)	0.462 (± 0.002)	0.466 (± 0.005)	0.296 (± 0.002)	0.309 (± 0.005)
	Ministral	0.505 (± 0.002)	0.502 (± 0.005)	0.469 (± 0.001)	0.456 (± 0.003)	0.308 (± 0.003)	0.302 (± 0.003)
LB	Llama3.1	0.579 (± 0.001)	0.605 (± 0.003)	0.621 (± 0.002)	0.626 (± 0.004)	0.410 (± 0.003)	0.463 (± 0.004)
	Qwen3	0.578 (± 0.002)	0.610 (± 0.003)	0.622 (± 0.003)	0.635 (± 0.004)	0.415 (± 0.002)	0.470 (± 0.004)
	Ministral	0.583 (± 0.002)	0.597 (± 0.005)	0.631 (± 0.002)	0.615 (± 0.004)	0.422 (± 0.004)	0.443 (± 0.003)

Table 7: **Extended Table for RQ1 & RQ2: Embedding Distance** Results show the impact of Supervised Fine-Tuning (FT) vs. prompting the base model (Base) on embedding distance [0-2]. Best values per comparison unit are **bolded**. Reported values are Mean (\pm Standard Deviation) across 5 independent generation runs on a hold-out test set of 650 users. All models (8B parameters) were conditioned using the combined *Biography+History* strategy and trained (FT) on a dataset of 3,800 users per language. Embedding Models: Qwen3-Embedding-8B (Zhang et al., 2025b), embeddinggemma-300m (Vera et al., 2025), LuxEmbedder (Philippy et al., 2025).

Label-Consistent Data Generation for Aspect-Based Sentiment Analysis Using LLM Agents

Mohammad H.A. Monfared[♣] Lucie Flek^{♣♠} Akbar Karimi^{♣♠}

[♣]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany

[♠]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
ak@bit.uni-bonn.de

Abstract

We propose an agentic data augmentation method for Aspect-Based Sentiment Analysis (ABSA) that uses iterative generation and verification to produce high-quality synthetic training examples. To isolate the effect of agentic structure, we also develop a closely matched prompting-based baseline using the same model and instructions. Both methods are evaluated across three ABSA sub-tasks—Aspect Term Extraction (ATE), Aspect Sentiment Classification (ATSC), and Aspect Sentiment Pair Extraction (ASPE)—four SemEval datasets, and two encoder-decoder models: T5-Base and Tk-Instruct. Our results show that the agentic augmentation outperforms raw prompting in label preservation of the augmented data, especially when the tasks require aspect term generation. In addition, when combined with real data, agentic augmentation provides higher gains, consistently outperforming prompting-based generation. These benefits are most pronounced for T5-Base, while the more heavily pretrained Tk-Instruct exhibits smaller improvements. As a result, augmented data helps T5-Base achieve comparable performance with its counterpart.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) seeks to identify aspect terms in text and determine the sentiment expressed toward each one (Pontiki et al., 2014, 2015, 2016). Despite its value in applications such as customer feedback analysis, ABSA remains difficult because it requires fine-grained, aspect-level annotations that are costly to produce and limited in coverage. As a result, models often struggle with rare aspect-sentiment combinations and domain-specific linguistic variation. Recent work has explored synthetic data generation using large language models (Zhong et al., 2024; Liu et al., 2024), but most approaches rely on prompt-

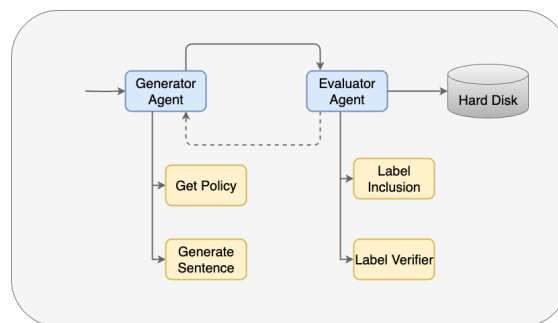


Figure 1: Overview of the agentic data augmentation workflow. A generator agent first extracts a style policy and produces candidate sentences, which are then evaluated by an evaluator agent. Only validated examples are saved, forming a high-quality synthetic dataset.

ing, which frequently produces label inconsistencies, limited structural diversity, and insufficient handling of corner cases.

In this work, we investigate whether agentic, multi-step data generation can address these limitations. We develop an augmentation pipeline that uses Qwen2.5 (Bai et al., 2023) within a ReAct-style (Yao et al., 2022) framework, separating generation and evaluation into dedicated agents equipped with tools for style extraction, policy construction, and label verification. The novelty of this approach lies in its task specific decomposition rather than a feedback loop. This structured approach ensures that synthetic examples strictly adhere to the complex requirements of aspect-based tasks. To isolate the effect of structured generation, we compare this system against a prompting baseline that uses the same model, prompts, and sampling strategy.

Empirically, we show that, compared to the raw prompting method, the agentic pipeline generates data, particularly those containing aspect terms, whose labels are more consistent with those of the original data. Furthermore, while the resulting synthetic data alone cannot replace human annota-

tions, the agentic augmentation—when mixed with real data—improves the performance for T5-Base and consistently outperforms naive prompting. We also find that augmentation benefits depend on the ABSA subtask and the underlying model: A simpler task (ATSC) and a less instruction-tuned architecture (T5-Base) gain the most, while the heavily pretrained model (Tk-Instruct) shows smaller improvements. Our contributions are threefold: 1) We introduce an agentic workflow for generating high-quality, task-specific synthetic data for ABSA; 2) We design a matched prompting baseline to isolate the contribution of structured, self-reflective generation; 3) We provide a systematic analysis of how model pretraining, augmentation scale, and subtask complexity shape the effectiveness of synthetic data.

2 Related Work

Prompting LLMs for Data Augmentation

Generating labeled data is a time-consuming and labor-intensive task. Traditional augmentation methods either function locally (Hsu et al., 2021a; Karimi et al., 2021b) or globally but in the embedding space, which reduces their controllability (Karimi et al., 2021a). LLMs provide an alternative to manual annotation by automatically generating labeled training data for supervised learning tasks (Meng et al., 2022; Ye et al., 2022a,b). Li et al. (2023) show that while LLMs can produce synthetic samples for text classification, the usefulness of this data varies with task characteristics, with more subjective or complex tasks showing reduced benefit from synthetic examples. Liu et al. (2024) study intrinsic self-correction in LLMs, where models are prompted to revise their own outputs after initial generation using only their internal capabilities, without external knowledge—highlighting the potential for autonomous quality improvement.

Applying LLM augmentation to structured problems like ABSA requires deeper task understanding, with its subtasks relying on fine-grained, aspect-level sentiment labels (Pontiki et al., 2014, 2015, 2016). Chen et al. (2022) adapt Unsupervised Data Augmentation for ABSA by introducing an MLM-based unmasking strategy to accommodate span-level structure and preserve token alignment. Hsu et al. (2021b) propose Selective Perturbed Masking (SPM), which modifies low-importance tokens while maintaining label consistency, enhancing lexical diversity. Zhong et al.

(2024) introduce Iterative Data Generation (IDG), where LLMs iteratively produce and filter pseudo-labeled ABSA data.

These techniques mostly rely on single-step prompting, perturbation-based generation, or sentence-level scoring using LLMs. Although some approaches incorporate LLM-based evaluation, they typically infer sentiment or quality from the generated text using heuristic or soft scoring criteria rather than explicitly checking whether the intended aspect-sentiment labels were realized as intended. Our approach instead introduces an agentic pipeline that separates sampling, style extraction, controlled generation, and automatic verification, using LLMs to ensure that generated sentences actually reflect the intended labels. This design leads to higher label consistency and structural alignment in generated ABSA data.

Agentic Data Augmentation Recent research investigated multi-agent and iterative generation frameworks to improve synthetic data quality. HydraGAN (DeSmet and Cook, 2024) introduces a multi-agent generative framework in which multiple discriminators jointly guide the data generation process. MAG-V (Sengupta et al., 2024) introduces a multi-agent framework in which agents generate synthetic examples and verify them through alternate question reconstruction, enhancing the reliability of generated data. APIGen-MT (Prabhakar et al., 2025) extends agentic generation to multi-turn settings by coordinating a two-phase pipeline that produces detailed task blueprints and uses simulated agent-human interaction to generate and verify complete multi-turn data.

These methods explored synthetic data generation in various domains, but they were not designed for the requirements of structured sentiment analysis problems such as ABSA. They also do not incorporate ABSA-specific constraints like aspect-span correctness or polarity alignment. Our work differs by introducing an ABSA-focused agentic data augmentation strategy that uses multi-step reasoning, controlled generation, and automatic verification to produce more consistent and task-aligned synthetic ABSA data.

3 Methodology

3.1 Agentic Data Augmentation

Our augmentation system is built around a ReAct-style agent architecture in which an LLM selects and invokes tools to perform subtasks such as style

extraction, sentence generation, and label verification. The workflow (Figure 1) consists of two cooperating agents—a generator and an evaluator—that operate under clearly defined roles and constraints. Importantly, both the agentic and baseline prompting methods use the same underlying model (Qwen2.5-14B) and nearly identical prompts; the only difference lies in the additional reasoning and validation steps introduced by the agentic framework. This ensures a controlled comparison of augmentation strategies.

3.1.1 Generator agent

The generator agent is responsible for producing candidate sentences that reflect ABSA structure and exhibit stylistic similarity to real customer reviews. It relies on two tools to accomplish this task. The first tool, **get policy**, prepares all metadata required for generation. It samples one to four aspect terms and their sentiment polarities from the SemEval training data, ensuring realistic label distributions. To guide stylistic variety, the tool also extracts writing style, grammatical structure, and approximate sentence length from a small set of real sentences drawn from the dataset. These components are combined into a generation policy, which acts as a blueprint for the construction of each synthetic example.

Once the policy is formed, the generator agent invokes its second tool, **generate sentences**, which uses the metadata to produce a candidate ABSA-style sentence. The underlying language model receives a structured instruction template that enforces both stylistic guidance and strict adherence to the sampled aspect–polarity pairs. The tool returns the generated sentence followed by machine-readable `Terms=` and `Polarity=` annotations, ensuring compatibility with downstream training pipelines. After this step, the candidate sentence is passed to the evaluator agent.

3.1.2 Evaluator Agent

The evaluator agent ensures that only valid, label-consistent examples are added to the synthetic dataset. It also operates using two specialized tools. The first, **label inclusion**, checks whether the generated sentence contains all required aspect terms exactly as specified in the policy. This step identifies structural deviations such as pluralization changes, partial matches, or missing terms, which frequently occur in naive prompting scenarios.

If the candidate passes this check, it is forwarded

to the second tool, **label verifier**, which evaluates whether each aspect term is associated with the correct sentiment polarity. The tool prompts the language model to judge the alignment between the sentence and the intended sentiment labels and returns an OK or NOT OK decision. Only examples that satisfy both structural and semantic criteria are accepted into the synthetic dataset; all others are discarded, and the generator agent is prompted to create a new example.

3.2 Prompt-Based Data Generation

To isolate the effect of agentic reasoning, we include a simple prompting-based baseline. This method uses the same underlying language model, the same aspect–polarity sampling procedure, and nearly the same instructions as the agentic generator, but it produces each sentence through a single monolithic prompt without any style extraction, tool usage, or verification steps. The prompting baseline therefore represents a minimal augmentation strategy against which the benefits of the agentic workflow can be directly measured. Full prompts are provided in Table 1.

4 Experiments

We evaluate our methods on the four standard SemEval ABSA datasets (Laptop14, Rest14, Rest15, and Rest16 (Pontiki et al., 2014, 2015, 2016)) covering the ATE, ATSC, and ASPE subtasks. All experiments follow the InstructABSA (Scaria et al., 2024) framework, using its preprocessing, training, and evaluation pipeline for comparability. We adopt the InstructABSA framework since it represents one of the highest-performing and recent benchmarks for instruction-based ABSA. Using its established pipelines ensures that our results are grounded in current standard practices and remain comparable with existing benchmarks. We fine-tune two encoder–decoder models: T5-Base (Raffel et al., 2020) and Tk-Instruct-base (Wang et al., 2022). Both models share the same architecture, but Tk-Instruct is extensively instruction-tuned on over 1,600 NLP tasks, unlike T5-Base. All hyperparameters match the InstructABSA¹ defaults, with the only modification being an increased fine-tuning budget of 20 epochs. Each experiment is repeated three times, and we report averaged scores for reliability.

¹<https://github.com/kevinscaria/InstructABSA>

Raw Prompting	Agentic Approach
<p>You are a critic who can generate comments on the specified aspect and sentiment. We would like you to complete a sentence generation task. Please follow these requirements:</p> <ul style="list-style-type: none"> - You need to use the sentiment, the aspect mentioned in the prompt - Domain: Restaurants - Your response must include: <ol style="list-style-type: none"> 1. The sentence. 2. A line that starts with Terms= followed by the list of aspect terms used. 3. A line that starts with Polarity= followed by the matching polarity list. - ALL aspect terms must appear as actual aspects in the sentence with intended polarities - Generated sentence must have the writing style and grammar structure and length of this sentence: {sent} - The sentence should not have aspect words not specified in the prompt - DO NOT repeat the input text in the output - PRINT ONLY THE ANSWER TEXT — no explaining, nothing else - Make sure to use aspect words in the output <p>Good Examples: ['prices'] ['negative'] The prices were too high for this type of restaurant ['Gnocchi', 'cheesecake'] ['positive', 'negative'] The Gnocchi was perfectly cooked and delicious, but the cheesecake was dry and flavorless.</p> <p>Bad Example: ['soup'], ['positive'] The udon soup was rich and flavorful. (term incorrect)</p> <p>Make sure your output exactly follows this format.</p>	<p>Generate Sentence Prompt: You are a critic who can generate comments on the specified aspect and sentiment. We would like you to complete a sentence generation task. Please follow these requirements:</p> <ul style="list-style-type: none"> - Generate a sentence using this aspect term: {aspect_term} with the following polarities: {polarity} - Write in the style: {writing_style}, and use a {grammar_structure} grammatical structure and {sentence_length} sentence length. - Domain: Restaurants - Include: <ol style="list-style-type: none"> 1. The sentence. 2. A line that starts with Terms= followed by the list of aspect terms used. 3. A line that starts with Polarity= followed by the matching polarity list. - Follow the exact structure shown in the examples. - Do not include explanations. <p>Use plain apostrophes (') — do not escape with backslashes.</p> <p>Label Verifier Prompt: You are an expert in linguistic evaluation. Check if the given aspect terms and polarities are correct for the provided sentence.</p> <ul style="list-style-type: none"> - If all aspect terms appear as actual aspects in the sentence with intended polarities, respond only with: OK - If any term is missing, incorrect, or has the wrong polarity, respond only with: NOT_OK <p>Do not provide explanations or any other text.</p> <p>Example Input: The food was lousy... Terms=['food'] Polarity=['negative'] → OK The udon soup was rich and flavorful. Terms=['soup'] Polarity=['positive'] → NOT_OK</p>

Table 1: Side-by-side comparison of raw prompting and agentic prompt structures. Both methods share identical task instructions and output constraints; the agentic approach differs only by decomposing generation and verification into separate steps, enabling automatic rejection of label-inconsistent outputs.

We use an automated experimental pipeline to systematically control training configurations, including data source (original, generated, mixed), augmentation ratio, and target subtask, ensuring consistent and fair comparison across all settings. Synthetic training examples are generated with the Qwen2.5-14B model using the Ollama framework². Both augmentation strategies—raw prompting and our proposed agentic workflow—use the same base model, the same prompts, and the same sampling procedure for aspect–polarity pairs. We evaluate three data configurations: (1) training on the original SemEval data only, (2) training on synthetic data only, and (3) training on a mixture of original and synthetic data. We Also test augmentation scales of x1 (equal to the size of the original training set) and x2 (twice the size).

5 Results and Discussions

Table 2 provides an overall summary of model performance across all tasks, datasets, and augmentation strategies. Several clear patterns emerge. First, training exclusively on synthetic data leads to a substantial drop in performance for both models, with prompting-based data performing worst and

agentic data offering a noticeably higher baseline. Second, when synthetic data is combined with real training examples, agentic augmentation consistently improves or closely matches the original-data-only performance, particularly for T5-Base, which shows gains across all three ABSA subtasks. Tk-Instruct benefits more modestly from augmentation, reflecting its stronger instruction-tuned prior, but still shows stable or slightly improved performance with agentic data. In contrast, prompting-based augmentation is less reliable. While it occasionally matches the original baseline, it often fails to provide meaningful improvements and sometimes degrades performance, especially in the more structurally complex tasks, such as aspect sentiment pair extraction and aspect term extraction.

A further observation is that increasing the amount of synthetic data beyond a 1:1 ratio generally does not yield additional gains. For both models and both augmentation strategies, the x2 setting either plateaus or slightly reduces performance relative to x1, indicating that larger quantities of synthetic data amplify noise rather than add useful training signal. Overall, these results show that high-quality, agenticly generated data can enhance ABSA performance when used in moderation, whereas naive prompting and large synthetic

²<https://ollama.com>

Model	Data Source	Training Strategy	Ratio	ASPE	ATE	ATSC
T5-Base	Original	Original	–	78.84	88.15	86.65
T5-Base	Agentic	Gen	1.0	39.16	54.56	81.72
T5-Base	Agentic	Gen	2.0	39.45	54.10	82.26
T5-Base	Agentic	Mixed	1.0	79.6	88.66	88.10
T5-Base	Agentic	Mixed	2.0	78.26	88.58	87.37
T5-Base	Prompting	Gen	1.0	31.29	49.97	80.53
T5-Base	Prompting	Gen	2.0	30.60	47.90	79.387
T5-Base	Prompting	Mixed	1.0	78.26	87.90	87.27
T5-Base	Prompting	Mixed	2.0	76.88	87.67	87.08
Tk-Instruct	Original	Original	–	80.84	89.45	88.25
Tk-Instruct	Agentic	Gen	1.0	42.17	54.69	80.79
Tk-Instruct	Agentic	Gen	2.0	42.54	53.05	81.22
Tk-Instruct	Agentic	Mixed	1.0	80.31	89.84	87.81
Tk-Instruct	Agentic	Mixed	2.0	80.02	89.07	87.54
Tk-Instruct	Prompting	Gen	1.0	31.98	46.40	79.34
Tk-Instruct	Prompting	Gen	2.0	30.62	44.67	78.39
Tk-Instruct	Prompting	Mixed	1.0	78.40	88.88	87.24
Tk-Instruct	Prompting	Mixed	2.0	79.15	87.98	87.21

Table 2: F₁ scores of training strategies averaged over data sets and runs.

datasets offer limited benefit.

5.1 Label Consistency of Generated Data

Label accuracy is critical for synthetic ABSA data, as LLM-generated sentences may alter aspect terms or express incorrect polarities, introducing noise that harms downstream training. To measure this, we use Tk-Instruct model (fine-tuned on SemEval-2016 Restaurants) as an automatic judge and compute how many generated examples preserve their intended aspect–polarity labels. We evaluate 1,800 synthetic samples from both the agentic and prompting methods, matching the size and conditions of the original training set.

It is important to note that the Tk-Instruct model used here serves only as an independent judge and was not involved in the data generation process. All the synthetic data was produced using Qwen2.5. We chose a fine-tuned Tk-Instruct for this measurement because of its high accuracy. Its judgments are used purely for analysis and have no influence on the synthetic data used for training.

Table 3 shows that agentic data has substantially higher label consistency across all tasks, with especially large gains in ATE and ASPE. These results confirm that the verification steps in the agentic workflow produce cleaner, more reliable synthetic labels than raw prompting.

5.2 Training on Generated Data

We evaluate how well models perform when trained exclusively on synthetic data generated through either prompting or the agentic workflow. Figure

Data Type	ATE	ATSC	ASPE
Agentic	78.17%	80.37%	33.89%
Prompting	43.89%	72.94%	18.33%

Table 3: Label consistency of augmented data from Rest16 dataset, measured by Tk-Instruct.

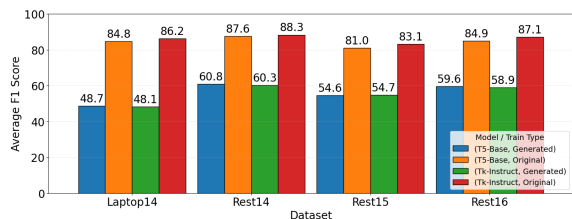


Figure 2: Average F₁ score across all ABSA tasks for each dataset, comparing original vs. generated-only training, signaling the clear gap in quality between real and synthetic data.

2 shows that training on synthetic data alone leads to a large and consistent performance drop across all tasks, datasets, and model architectures. This confirms that current generation methods, despite improvement from agentic workflows, can not yet replace human-annotated ABSA data. Detailed per-task and per-dataset results for ATE, ATSC, and ASPE are provided in Tables 6, 7, and 8 in the Appendix.

A major contributor to this performance gap is the lack of linguistic richness in synthetic reviews. Real SemEval sentences often contain indirect sentiment, idioms, sarcasm, vague references, and domain-specific expressions that are difficult for generation models to reproduce. Even with

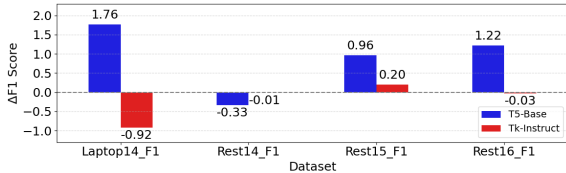


Figure 3: ΔF_1 between baseline and added agentic data, across both T5-Base and Tk-Instruct. Each bar represents mean F_1 scores averaged over ATE, ATSC, and ASPE tasks for each dataset. This plot shows the clear difference of the augmentation effectiveness on these models.

stylistic conditioning, synthetic examples tend to be cleaner, more literal, and less ambiguous, reducing the diversity needed for robust generalization. In addition, as mentioned earlier (Table 3), synthetic data—especially from prompting—still contains label inconsistencies that introduce further noise. These semantic simplifications and residual labeling errors compound most severely in complex tasks like ASPE, widening the gap between synthetic-only and real-data training.

5.3 Training on Mixed Data

While synthetic data on its own does not match the performance of human-annotated training data, it can still have a positive impact when used in combination with real examples (Table 4 for the aspect term extraction task). In this setup, the synthetic data serves as a complementary source of variation. It helps reinforce the core task structure and boosts the diversity of examples the model sees, while the real data provides that essential anchor with linguistically rich and semantically consistent examples. Detailed per-task and per-dataset results for ATSC, and ASPE are provided in Tables 9 and 10 in the Appendix.

5.4 Model Sensitivity to Data Augmentation: T5 vs. Tk-Instruct

Our results show that T5-Base and Tk-Instruct respond very differently to synthetic augmentation. Across tasks and datasets, T5-Base consistently benefits from agentic data in the Mixed x1 setting, with clear gains in ATSC and stable or slightly improved performance in ATE. In contrast, augmentation has only marginal effects on Tk-Instruct, and in some cases—particularly ASPE—slightly lowers performance. This pattern is also reflected in the aggregated task-level comparison (Figure 3), where T5-Base shows noticeably larger improve-

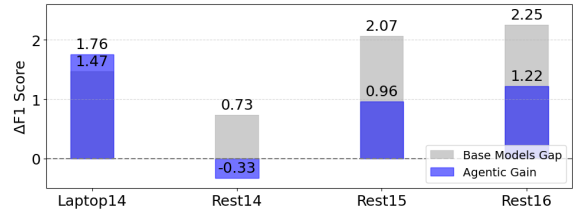


Figure 4: Agentic data augmentation narrows the F_1 gap between T5-Base and Tk-Instruct. Grey bars indicate the original model gap; blue bars show the performance gain of T5-Base after agentic augmentation (Mixed x1).

ments from mixing real and synthetic data.

These differences can be attributed to the models’ pretraining. Tk-Instruct is heavily instruction-tuned on over 1,600 NLP tasks (Wang et al., 2022), giving it broad exposure to task structures and linguistic patterns that already overlap with ABSA. As a result, its baseline performance is strong and additional synthetic examples add little new information, sometimes even introducing redundant or noisy patterns. T5-Base, lacking this instruction-driven prior, benefits much more from the additional variation and task-specific structure introduced by agentic augmentation.

This difference reinforces the idea that the utility of data augmentation is inversely related to the generalization capabilities already embedded in a model’s pretraining. When a model has already been exposed to massive, diverse, and well-labeled task instructions, the benefits from new, generated examples are just not as significant.

Additionally, This difference has an important consequence: targeted synthetic augmentation can significantly narrow, and sometimes eliminate, the performance gap between the two models. Although Tk-Instruct consistently outperforms T5-Base when trained only on human-annotated data, the introduction of agentic synthetic examples (Mixed x1) leads to sharp improvements in T5-Base. As shown in Figure 4, these gains often bring T5-Base close to Tk-Instruct and even allow it to surpass the counterpart.

This shift highlights a key contribution of our work: task-specific agentic augmentation can substitute for large-scale instruction tuning. Whereas Tk-Instruct achieves its performance through massive, manually curated pretraining, our agentic pipeline is fully automatic and tailored directly to ABSA. Despite this simplicity, it provides substantial benefits to models like T5-Base and reduces their dependence on costly human-designed pre-

Model	Data Type	Laptop14	Rest14	Rest15	Rest16	
T5-Base	Original Data	93.32	95.12	80.90	83.28	
	Mixed Data - Agentic					
	Mixed x1	93.82 $\uparrow 0.50$	94.63 $\downarrow 0.49$	80.98 $\uparrow 0.08$	85.21 $\uparrow 1.93$	
	Mixed x2	93.50 $\uparrow 0.18$	94.26 $\downarrow 0.86$	82.22 $\uparrow 1.32$	84.37 $\uparrow 1.09$	
	Mixed Data - Prompting					
	Mixed x1	92.83 $\downarrow 0.49$	94.53 $\downarrow 0.59$	80.43 $\downarrow 0.47$	83.83 $\uparrow 0.55$	
	Mixed x2	92.97 $\downarrow 0.35$	94.14 $\downarrow 0.98$	78.88 $\downarrow 2.02$	84.72 $\uparrow 1.44$	
Tk-Instruct	Original Data	93.91	95.01	83.41	85.47	
	Mixed Data - Agentic					
	Mixed x1	93.93 $\uparrow 0.02$	94.20 $\downarrow 0.81$	84.12 $\uparrow 0.71$	87.14 $\uparrow 1.67$	
	Mixed x2	94.07 $\uparrow 0.16$	93.34 $\downarrow 1.67$	82.50 $\downarrow 0.91$	86.38 $\uparrow 0.91$	
	Mixed Data - Prompting					
	Mixed x1	93.80 $\downarrow 0.11$	94.10 $\downarrow 0.91$	82.30 $\downarrow 1.11$	85.32 $\downarrow 0.15$	
	Mixed x2	93.35 $\downarrow 0.56$	94.20 $\downarrow 0.81$	80.00 $\downarrow 3.41$	84.39 $\downarrow 1.08$	

Table 4: ATE F_1 scores across datasets for each model trained on original data and mixed datasets combining original and synthetic examples. Mixed x1 and Mixed x2 denote synthetic-to-original data ratios of 1:1 and 2:1, respectively. Agentic augmentation provides clearer gains, while prompting-based augmentation yields limited or negative impact.

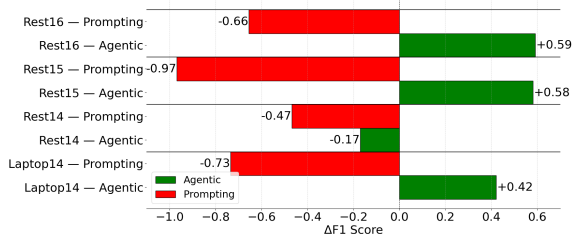


Figure 5: Average F_1 score change (ΔF_1) for agentic and prompting augmentation, computed as the difference between the Mixed x1 setup and original-only baseline. Scores are averaged across three ABSA subtasks and two model architectures for each dataset.

training corpora.

5.5 Agentic vs. Prompting-Based Augmentation

Since both augmentation methods use the same language model, the same prompts, and the same aspect-polarity sampling strategy, any difference in downstream performance is expected to come from the generation process itself. Across all tasks, datasets, and model architectures, agentic augmentation consistently improves or matches baseline performance, whereas prompting-based augmentation rarely helps and often degrades model accuracy. Figure 5 summarizes this trend.

The advantage of the agentic workflow stems from its built-in evaluation and verification steps. Unlike prompting, which accepts every LLM output—even those with incorrect aspect terms, mismatched polarities, or overly simplistic phrasing—the agentic pipeline filters out noisy genera-

tions through explicit policy construction and label checking. As a result, agentic data achieves substantially higher label consistency (Table 3), making it more beneficial when utilized either alone or mixed with real data.

Both methods also employ uniform sampling to counteract label and term-frequency biases in the original datasets, and both help expose models to underrepresented aspect terms and polarities. Figure 6 depicts the distribution in SemEval 16 restaurant data set. However, only the agentic approach provides this benefit without introducing too much additional noise. Prompting-generated data suffers from low label accuracy 3, which cancels out the expected gains of bias correction and ultimately harms downstream training.

Overall, the evidence shows that high-quality augmentation requires more than good prompts: it requires structured generation, verification, and filtering. The agentic framework delivers this control, producing cleaner, more reliable synthetic examples, while naive prompting fails to meet the consistency that ABSA tasks demand.

5.6 Task Sensitivity to Augmentation: ATE vs. ASPE vs. ATSC

The effects of data augmentation are not uniform across ABSA subtasks. As we can see in Figure 7, both ATE and ATSC benefit consistently from agentic synthetic data, showing clear improvements across most datasets and models. In contrast, gains in ASPE are limited and often inconsistent, especially for Tk-Instruct. This pattern aligns with

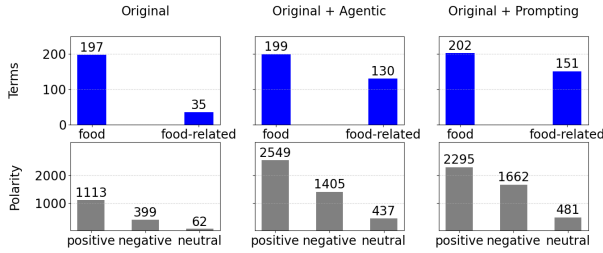


Figure 6: Comparison of food-related term frequency and sentiment polarity in Rest 16 across original and mixed data sets, suggesting that augmentation mitigates the inherent biases.

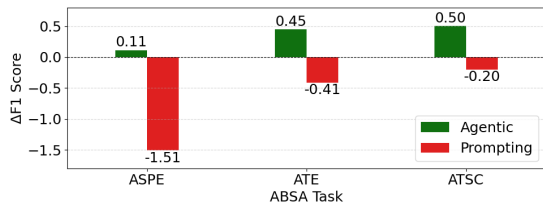


Figure 7: Average change in F1 score across all datasets and both models for each ABSA task (ATE, ATSC, ASPE) under Mixed x1 conditions. This figure points out the effectiveness of augmentation for each task.

the relative difficulty of the tasks: ATSC requires only sentiment classification for a given aspect, making it the easiest to support with synthetic examples. ATE is more challenging because the model must identify aspect spans, and ASPE is the most complex, requiring both extraction and sentiment assignment simultaneously.

The label consistency results in Table 3 reinforce this trend. Synthetic examples are highly reliable for ATSC, where the model only needs to express sentiment toward an already-given aspect. Because this task involves a single, explicit decision, both prompting and agentic generation tend to produce clean and well-aligned labels. For ATE, consistency drops: the model must identify aspect spans on its own, and generated sentences sometimes introduce ambiguity—such as multi-word expressions, coordination, or slight wording mismatches—that make span extraction harder to learn from. ASPE shows the lowest consistency, as it compounds the challenges of both ATE and ATSC. Here, even small deviations (e.g., missing one term, splitting a phrase, misplacing a sentiment) break the aspect–sentiment pairing and introduce noise into training.

This downward progression in label reliability directly mirrors the downstream performance patterns. When the labels are simple and unambigu-

ous, as in ATSC, augmentation provides strong and consistent improvements. When labels become harder to generate and validate, as in ATE and especially ASPE, the benefit diminishes and may even reverse. In essence, the harder the task, the more fragile augmentation becomes, because even subtle inconsistencies in synthetic labels can mislead the model and outweigh the value of increased data diversity.

6 Conclusions and Future Work

In this research, we explored the effectiveness of agentic data augmentation specifically for Aspect-Based Sentiment Analysis (ABSA). By comparing an agentic workflow to an otherwise identical prompting-based pipeline—using the same model, prompts, and sampling—we isolated the impact of structured generation and verification on downstream performance across three ABSA subtasks and four benchmark datasets. Our results show that, while synthetic data does not contain as much information as human annotations, high-quality agentic data improves performance when mixed with real examples. These gains are strongest for T5-Base, enabling it to approach or even surpass the baseline performance of the more extensively pretrained Tk-Instruct model. This demonstrates that task-specific, high-quality augmentation can partially close the gap created by large-scale instruction tuning.

We also find that agentic augmentation consistently outperforms naive prompting due to its higher label consistency and lower noise, despite using the same underlying model. Benefits are task-dependent: ATSC and ATE show clear improvements, while ASPE sees limited gains due to its higher structural complexity. Finally, augmentation is effective only in moderation—1:1 mixing yields improvements, but larger synthetic datasets dilute quality and reduce performance. Overall, these findings highlight that controlled, agentic generation provides a practical and scalable way to enhance ABSA models, offering meaningful gains without requiring extensive pretraining or additional human annotation.

Future work may extend this agentic augmentation framework to additional languages and domains, leveraging multilingual LLMs to broaden applicability with minimal structural changes. Another promising direction is to decompose the pipeline into task-specialized components—for ex-

ample, using a strong text generator for sample creation and a dedicated ABSA model for validation—to further improve data quality. Incorporating adaptive sampling or confidence-based filtering could help prioritize informative and reliable synthetic examples, reducing noise during training. Finally, integrating model-in-the-loop feedback into the agentic workflow would allow the system to generate data that directly targets a model’s weaknesses, enabling more efficient, responsive, and context-aware augmentation.

7 Limitations

This work is constrained by the capabilities of the open-source LLMs available to us. Although Qwen2.5 provides solid tool-calling performance, it remains less capable than proprietary state-of-the-art systems, limiting the complexity of the agent architecture we could implement. In addition, the selected model might be less capable than other larger open-source models, which can impact the quality of the decisions it makes during the process of data generation.

8 Ethical Considerations

This work uses publicly available SemEval datasets and open-source models such as T5, Tk-Instruct, and Qwen2.5. While widely adopted, these resources may contain biases or inaccuracies that can influence model behavior; a detailed analysis of such issues is beyond the scope of this study. We also acknowledge that sentiment analysis systems can be misused if deployed irresponsibly. Although our focus is strictly on technical aspects of data augmentation, we emphasize the need for careful, ethical application of these methods and encourage future work to examine fairness, bias, and potential downstream risks more closely.

Acknowledgments

This work was supported by the BMFTR and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence and the AISafety Project, funded under the BMFTR grant proposal No. 05D23PD1.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

David Z Chen, Adam Faulkner, and Sahil Badyal. 2022. Unsupervised data augmentation for aspect based sentiment analysis. In *Proceedings of the 29th international conference on computational linguistics*, pages 6746–6751.

Chance DeSmet and Diane Cook. 2024. Hydragan: A cooperative agent model for multi-objective data generation. *ACM transactions on intelligent systems and technology*, 15(3):1–21.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4417–4422.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021b. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4417–4422.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021a. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International conference on pattern recognition (ICPR)*, pages 8797–8803. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021b. Aeda: An easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.

Z. Li, H. Zhu, Z. Lu, M. Yin, R. Bonatti, S. Abdali, J. Wagle, and K. Koishida. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, and 1 others. 2024. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülsen Eryigit. 2016. **SemEval-2016 Task 5: Aspect Based Sentiment Analysis**. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California.

- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. [SemEval-2015 Task 12: Aspect Based Sentiment Analysis](#). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. [Semeval-2014 task 4: Aspect based sentiment analysis](#). In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland.
- Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Awalgaoonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, and 1 others. 2025. [Apigen-mt: Agentic pipeline for multi-turn data generation via simulated agent-human interplay](#). *arXiv preprint arXiv:2504.03601*.
- 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.
- Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. [Instructabsa: Instruction learning for aspect based sentiment analysis](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736.
- Saptarshi Sengupta, Harsh Vashistha, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. [Mag-v: A multi-agent framework for synthetic data generation and verification](#). *arXiv preprint arXiv:2412.04494*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, and 1 others. 2022. [Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. [React: Synergizing reasoning and acting in language models](#). In *The eleventh international conference on learning representations*.
- Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. [Zerogen: Efficient zero-shot learning via dataset generation](#). *arXiv preprint arXiv:2202.07922*.
- Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. [Progen: Progressive zero-shot dataset generation via in-context feedback](#). In *Findings of the association for computational linguistics: EMNLP 2022*, pages 3671–3683.
- Qihuang Zhong, Haiyun Li, Luyao Zhuang, Juhua Liu, and Bo Du. 2024. [Iterative data generation with large language models for aspect-based sentiment analysis](#). *arXiv preprint arXiv:2407.00341*.

9 Appendix

9.1 Agentic Generation and Evaluation Traces

The following examples illustrate the agentic workflow for data generation and evaluation. The generation trace shows how sentences are produced under explicit stylistic and label constraints, while the evaluation trace demonstrates how incorrect or misaligned outputs are automatically filtered. Together, these examples highlight how the agentic pipeline enforces label correctness before synthetic data is accepted.

Agentic Generation Trace Example

Agent Input:

generate a sentence with dataset information in mind.

Agent Reasoning Trace:

- **Tool Call:** `get_info()`
- **Tool Response:** `{"writing_style": "informal", "grammar_structure": "simple sentences", "length": "medium"}, "Terms": ['Ginger House', "Polarity": ['positive']}`
- **Tool Call:**
`generate_sentences(style_info= {"writing_style": "informal", "grammar_structure": "simple sentences", "length": "medium"})`
- **Tool Response:**
The Ginger House is a cozy spot that really warms the heart!
Terms= ['Ginger House']
Polarity= ['positive']

Agentic Evaluation Trace Example

Input:

The balcony was cramped and had limited tables, but I loved the view from it.
Terms= ['balcony'], Polarity= ['negative']

Agent Trace:

- **Tool Call:** `label_inclusion`
- **Tool Response:** OK
- **Tool Call:** `evaluate_sentence`
- **Tool Response:** NOT_OK
- **Final Decision:** NOT_OK
The generated sentence failed the second check

9.2 Additional Results

ID	raw_text	aspectTerms	aspectCategories
2383	Bottom line: B+ for the food, F for the service.	[[term: <i>food</i> , polarity: <i>positive</i>], {term: <i>service</i> , polarity: <i>negative</i> }]	[[category: <i>food</i> , polarity: <i>positive</i>], {category: <i>service</i> , polarity: <i>negative</i> }]
766	We always go there on the weekends and leave extremely full and satisfied.	[[term: <i>noaspectterm</i> , polarity: <i>none</i>]]	[[category: <i>anecdotes</i> , polarity: <i>positive</i>]]
1419	We actually gave 10% tip (which we have never done despite mediocre food and service), because we felt totally ripped off.	[[term: <i>food</i> , polarity: <i>neutral</i>], {term: <i>service</i> , polarity: <i>neutral</i> }, {term: <i>tip</i> , polarity: <i>negative</i> }]	[[category: <i>food</i> , polarity: <i>neutral</i>], {category: <i>service</i> , polarity: <i>neutral</i> }, {category: <i>price</i> , polarity: <i>negative</i> }]
1700	The food was good.	[[term: <i>food</i> , polarity: <i>positive</i>]]	[[category: <i>food</i> , polarity: <i>positive</i>]]
1892	Service is great, takeout is good too.	[[term: <i>Service</i> , polarity: <i>positive</i>], {term: <i>takeout</i> , polarity: <i>positive</i> }]	[[category: <i>food</i> , polarity: <i>positive</i>], {category: <i>service</i> , polarity: <i>positive</i> }]

Table 5: Sample entries from SemEval-2014 Restaurant dataset

Model	Data Type	Laptop14	Rest14	Rest15	Rest16
T5-Base	Original Data	93.32	95.12	80.90	83.28
	Generated Data - Agentic				
	Gen. x1	46.18	66.53	50.82	54.73
	Gen. x2	46.83	65.09	50.19	54.30
	Generated Data - Prompting				
	Gen. x1	41.46	58.85	48.45	51.13
	Gen. x2	40.35	55.73	46.49	49.05
Tk-Instruct	Original Data	93.91	95.01	83.41	85.47
	Generated Data - Agentic				
	Gen. x1	47.23	65.23	50.63	55.69
	Gen. x2	45.39	64.00	50.82	52.02
	Generated Data - Prompting				
	Gen. x1	39.69	55.38	44.67	45.86
	Gen. x2	39.26	51.71	43.89	43.83

Table 6: ATE F_1 scores across datasets for each model trained on original data and on synthetic data only. Gen. x1 and Gen. x2 indicate synthetic data generated at 1:1 and 2:1 ratios relative to the original training set. Agentic generation consistently outperforms prompting, while synthetic-only training shows lower performance.

Model	Data Type	Laptop14	Rest14	Rest15	Rest16
T5-Base	Original Data	79.78	86.25	88.74	91.86
	Generated Data - Agentic				
	Gen. x1	77.42	79.64	81.54	88.29
	Gen. x2	78.05	80.00	81.73	89.26
	Generated Data - Prompting				
	Gen. x1	75.86	76.85	81.84	87.57
	Gen. x2	74.92	75.53	80.44	86.66
Tk-Instruct	Original Data	82.75	87.41	88.56	94.30
	Generated Data - Agentic				
	Gen. x1	74.60	79.37	80.44	88.78
	Gen. x2	76.64	79.10	79.88	89.26
	Generated Data - Prompting				
	Gen. x1	73.04	76.25	82.28	86.17
	Gen. x2	72.57	76.25	79.70	85.04

Table 7: ATSC F_1 scores across datasets for each model trained on original data and on synthetic data only. Gen. x1 and Gen. x2 indicate synthetic data generated at 1:1 and 2:1 ratios relative to the original training set. Agentic generation consistently outperforms prompting, while synthetic-only training shows lower performance.

Model	Data Type	Laptop14	Rest14	Rest15	Rest16	
T5-Base	Original Data	81.19	81.34	73.38	79.47	
	Generated Data - Agentic					
	Gen. x1	28.42	48.79	37.94	41.49	
	Gen. x2	28.81	48.88	37.58	42.56	
	Generated Data - Prompting					
	Gen. x1	23.13	38.15	28.58	35.30	
	Gen. x2	22.80	36.00	29.09	34.51	
Tk-Instruct	Original Data	82.03	82.49	77.25	81.60	
	Generated Data - Agentic					
	Gen. x1	31.98	50.07	40.02	46.63	
	Gen. x2	31.72	50.60	41.04	46.83	
	Generated Data - Prompting					
	Gen. x1	23.47	39.47	31.87	33.12	
	Gen. x2	22.04	36.03	31.32	33.12	

Table 8: ASPE F_1 scores across datasets for each model when trained on original data and on synthetic data only. Gen. x1 and Gen. x2 denote synthetic datasets generated at ratios of 1:1 and 2:1 relative to the original training set size. The performance gap between synthetic-only and original data is most pronounced for ASPE, highlighting the difficulty of learning joint aspect-sentiment extraction from generated data alone, even with agentic generation.

Model	Data Type	Laptop14	Rest14	Rest15	Rest16	
T5-Base	Original Data	79.78	86.25	88.74	91.86	
	Mixed Data - Agentic					
	Mixed x1	83.38 $\uparrow 3.60$	86.78 $\uparrow 0.53$	88.92 $\uparrow 0.18$	93.33 $\uparrow 1.47$	
	Mixed x2	81.34 $\uparrow 1.56$	86.42 $\uparrow 0.17$	88.56 $\downarrow 0.18$	93.17 $\uparrow 1.31$	
	Mixed Data - Prompting					
	Mixed x1	79.93 $\uparrow 0.15$	86.96 $\uparrow 0.71$	88.37 $\downarrow 0.37$	93.82 $\uparrow 1.96$	
	Mixed x2	81.34 $\uparrow 1.56$	85.62 $\downarrow 0.63$	88.37 $\downarrow 0.37$	93.00 $\uparrow 1.14$	
Tk-Instruct	Original Data	82.75	87.41	88.56	94.30	
	Mixed Data - Agentic					
	Mixed x1	80.72 $\downarrow 2.03$	88.30 $\uparrow 0.89$	88.92 $\uparrow 0.36$	93.33 $\downarrow 0.97$	
	Mixed x2	81.34 $\downarrow 1.41$	87.41 $\downarrow 0.00$	88.74 $\uparrow 0.18$	92.68 $\downarrow 1.62$	
	Mixed Data - Prompting					
	Mixed x1	80.87 $\downarrow 1.88$	87.58 $\uparrow 0.17$	88.00 $\downarrow 0.56$	92.52 $\downarrow 1.78$	
	Mixed x2	81.19 $\downarrow 1.56$	87.23 $\downarrow 0.18$	87.60 $\downarrow 0.96$	92.84 $\downarrow 1.46$	

Table 9: ATSC F_1 scores across datasets for each model trained on original data and mixed datasets combining original and synthetic examples. Mixed x1 and Mixed x2 denote synthetic-to-original data ratios of 1:1 and 2:1; Agentic augmentation yields the largest and most consistent gains for T5-Base, whereas Tk-Instruct shows mixed or negative changes under both augmentation methods.

Model	Data Type	Laptop14	Rest14	Rest15	Rest16	
T5-Base	Original Data	81.19	81.34	73.38	79.47	
	Mixed Data - Agentic					
	Mixed x1	82.37 $\uparrow 1.18$	80.31 $\downarrow 1.03$	76.00 $\uparrow 2.62$	79.72 $\uparrow 0.25$	
	Mixed x2	78.84 $\downarrow 2.35$	81.52 $\uparrow 0.18$	73.89 $\uparrow 0.51$	78.79 $\downarrow 0.68$	
	Mixed Data - Prompting					
	Mixed x1	80.68 $\downarrow 0.51$	80.42 $\downarrow 0.92$	72.92 $\downarrow 0.46$	79.02 $\downarrow 0.45$	
	Mixed x2	78.71 $\downarrow 2.48$	80.42 $\downarrow 0.92$	72.34 $\downarrow 1.04$	76.06 $\downarrow 3.41$	
Tk-Instruct	Original Data	82.03	82.49	77.25	81.60	
	Mixed Data - Agentic					
	Mixed x1	81.29 $\downarrow 0.74$	82.38 $\downarrow 0.11$	76.79 $\downarrow 0.46$	80.80 $\downarrow 0.80$	
	Mixed x2	80.62 $\downarrow 1.41$	81.66 $\downarrow 0.83$	75.60 $\downarrow 1.65$	82.21 $\uparrow 0.61$	
	Mixed Data - Prompting					
	Mixed x1	80.46 $\downarrow 1.57$	81.22 $\downarrow 1.27$	74.41 $\downarrow 2.84$	77.54 $\downarrow 4.06$	
	Mixed x2	79.70 $\downarrow 2.33$	81.52 $\downarrow 0.97$	74.10 $\downarrow 3.15$	81.30 $\downarrow 0.30$	

Table 10: ASPE F_1 scores across datasets for each model trained on original data and mixed datasets combining original and synthetic examples. Mixed x1 and Mixed x2 denote synthetic-to-original data ratios of 1:1 and 2:1, respectively; Agentic augmentation yields modest and dataset-dependent improvements. Prompting-based augmentation generally leads to performance drops for both models.

Antisocial Behavior Prediction: A Survey and Practical Guide

Anaïs Ollagnier

Université Côte d’Azur, Inria, CNRS, I3S / 930 route des Colles, BP 145,
Sophia Antipolis Cedex, 06903, France
ollagnier@i3s.unice.fr

Abstract

Antisocial behavior (ASB) on social media encompasses online behaviors that harm individuals, groups, or platform ecosystems, including hate speech, harassment, cyberbullying, trolling, and coordinated abuse. While most prior work has focused on detecting harm after it occurs, a growing body of research on ASB prediction seeks to forecast future harmful outcomes before they materialize, including—but not limited to—hate-speech diffusion, conversational derailment, and user recidivism. However, this emerging field remains fragmented, with limited conceptual grounding and few integrative frameworks. This paper establishes a foundation for ASB prediction by introducing a structured taxonomy spanning temporal, structural, and behavioral dimensions. Drawing on 49 machine learning studies identified through a literature review, we map predictive goals to datasets, modeling choices, and evaluation practices, and identify key challenges, including the lack of standardized benchmarks, the dominance of text-centric representations, and trade-offs between accuracy and interpretability. We conclude by outlining actionable directions toward more robust, generalizable, and responsible ASB prediction systems.

1 Introduction

Social media platforms have reshaped how people consume information and engage in public discourse (Kahn and Kellner, 2004; Brown et al., 2007; Quattrociocchi et al., 2014). While these systems enable large-scale participation (Criado et al., 2013; Etta et al., 2022), they also facilitate harmful dynamics such as misinformation (Vicario et al., 2016), echo chambers (Cinelli et al., 2021), and hostility (Cheng et al., 2017; Saveski et al., 2021). These conditions often lead to antisocial behavior (ASB)—understood here as online behaviors that disregard or violate the rights and well-being of others, including hate speech, harassment, trolling,

and related forms of harmful interaction (Ollagnier et al., 2023a,b; Chowdhury et al., 2019)—with serious individual and societal consequences (Parent et al., 2019; Saha et al., 2019).

While harmful content detection has been extensively studied, ASB prediction—defined as the computational modeling of future harmful behaviors—remains underexplored, as evidenced by our systematic review identifying only 49 machine-learning studies since 2014 (Appendix A). The field has nonetheless accelerated sharply: from isolated contributions in 2014–2017 (one paper per year), to modest growth in 2018–2019, and a marked surge from 2020 onward. This expansion reflects heightened awareness of online harms, broader access to social media data, rapid advances in NLP and machine learning (ML), and growing regulatory and societal pressures—including the EU Digital Services Act¹ (2025), the UN’s International Day for Countering Hate Speech², and grassroots counter-speech movements such as #IamHere³. Importantly, this evolution is not only quantitative but also qualitative: since 2020, ASB prediction has gained visibility in top-tier venues such as EMNLP, NAACL, ICWSM, KDD, and IEEE Transactions on Network Science and Engineering, reflecting growing methodological sophistication and interdisciplinary engagement. Yet the field remains fragmented, with limited conceptual grounding and few integrative frameworks.

To address this gap, we draw on our systematic literature review of machine-learning-oriented studies to propose a structured synthesis of progress to date, identify best practices, and highlight current limitations alongside open challenges. After introducing background on ASB (§2), the survey is structured around three dimensions: task design (§3), data selection (§4), and evaluation

¹EU Digital Services Act

²UN Strategy and Plan of Action on Hate Speech

³#IamHere International

(§5). We then discuss key challenges and promising research directions (§6), with full details of the review methodology provided in Appendix A.

2 Background

2.1 Definitions

Hate speech has traditionally occupied a central place in academic and legal debates, yet it represents only one subset of abusive language, typically defined by its *function* (e.g., incitement) and *target* (identity-based groups) (Poletto et al., 2021). More broadly, ASB originates in social and clinical psychology, where it denotes a persistent pattern of disregard for or violation of the rights and well-being of others, as formalized in the DSM-5 (American Psychiatric Association, 2013). In online contexts, we use ASB not as a clinical diagnosis but as a behavioral construct capturing observable actions that harm individuals, groups, or platform ecosystems. Adopting ASB as an umbrella term reduces the conceptual fragmentation introduced by overlapping notions such as *online abuse*, *toxic language*, and *cyberaggression* (Poletto et al., 2021; Alkomah and Ma, 2022; Jahan and Oussalah, 2023). Drawing on social psychology, platform governance frameworks (Meta, 2022), and computational social science (Gruzd et al., 2020; Haythornthwaite, 2023), we define online ASB along three dimensions: **personal harms** (e.g., harassment, cyberbullying), **group-directed harms** (e.g., identity-based hate, stereotyping), and **environmental disruptions** (e.g., trolling, misinformation, coordinated harassment). This definition integrates both linguistic content and behavioral patterns, providing a clear foundation for ASB prediction.

2.2 Taxonomy of ASB computational tasks

To address the absence of a structured framework for ASB prediction, we propose a taxonomy grounded in the collected literature. Adapting the thematic analysis framework of Zhou et al. (2022) to our goals, we organize predictive tasks along two complementary dimensions: their *temporal orientation*, which specifies when harm is anticipated (e.g., emergence, escalation, or diffusion), and their *operational purpose*, which clarifies why it is modeled (e.g., moderation, risk assessment, or intervention). This two-dimensional taxonomy yields five task categories, illustrated below with scenarios (see Appendix B for a detailed overview).

Harm Emergence. This task forecasts whether an initially civil interaction—such as a thread, comment exchange, or reply chain—will develop into harmful behavior. Models look for subtle precursors like escalating rhetoric, sentiment shifts, or power dynamics to anticipate toxicity before explicit harmful content appears.

User A: I don't think the article should call the policy "controversial."

User B: Reliable sources use that term, so it seems justified.

User A: **You're just pushing your agenda.**

User B: That's unfair—please stick to the evidence.

User A: People like you ruin every discussion here!

Observed pattern: The bolded remark introduces an adversarial tone in an otherwise civil exchange, signaling a possible shift toward personal attacks.

Harm Propagation. Rather than predicting if harm will occur, this task models how existing harmful content will spread across networks. It estimates the reach, speed, and diffusion trajectory of toxic material, identifying which hateful posts are likely to go viral or which communities will amplify them.

User A: Those people are ruining our country. Share this so everyone knows the truth!

User B: Exactly! Everyone in my group needs to see this. (shares to a community group)

User C: This is going viral—just shared it with 10 friends.

User D: Spreading fast—check out this hateful meme. (reshare with image)

Observed pattern: Rapid resharing across groups indicates accelerating virality, pointing to high exposure risk if left unchecked.

Early Harm Detection. This task emphasizes rapid intervention, predicting harmful behavior from minimal early signals—often just the first few messages in a thread. The aim is to enable triage before harm escalates.

User A: "Does anyone have tips for first-time job seekers?"

User B: "Maybe start by not being so lazy."

User C: "Yeah, people like you just want handouts."

Observed pattern: Even at the outset, the exchange shows dismissive and stereotyping remarks, making it a likely candidate for escalation into bullying.

Behavioral Risk. Here, the focus shifts from content to individuals. Models assess whether a user is likely to engage in, or become a target of, ASB. They build risk profiles from prior behavior, lin-

guistic cues, and engagement patterns.

Perpetration risk (user likely to engage in ASB).
User X (Mon): “Debate’s heated, but let’s keep it civil.”
User X (Wed): “Some people just don’t belong here...”
User X (Thu): joins @HardLineForum; likes several posts with coded slurs.
User X (Fri): “They’re parasites—open your eyes.”
Observed pattern: A progression from neutral to exclusionary to dehumanizing language, coupled with extremist ties, reveals growing perpetration risk.

Proactive Moderation. Models in this category support interventions before harm materializes. They predict harmfulness at the moment of posting, suggest safer edits, or prioritize items for review.

User A (drafting a comment): “This journalist is a lying idiot—don’t trust anything she writes.”
Observed pattern: At the drafting stage, the system flags personal insults and polarizing phrasing, offering a constructive rewording before publication.

3 Step 1: Design your task

The first step in building an ASB prediction system is to determine which task(s) you aim to address. Tasks vary depending on whether you predict categories or continuous outcomes, and at which level—actors or content—they occur.

3.1 Classifying ASB

Classification tasks assign a discrete label to an observation to characterize the phenomenon. Schemes range from binary (e.g., toxic vs. non-toxic content (Kennedy et al., 2020a; Saveski et al., 2021) or banned vs. not banned users (Cheng et al., 2015)) to multi-class setups, such as high/medium/low incivility (Yu et al., 2024).

Actor classification. Actor-level tasks predict user-level risks and behaviors. Models estimate whether a user will be banned (Cheng et al., 2015; Hickey et al., 2025), join or escalate activity in harmful communities, spread toxic narratives (Irani et al., 2021), or generate abusive replies (Tsantarliotis et al., 2017). Some studies also forecast longitudinal trends, such as users’ contributions to future harm or their recidivism risk (Song et al., 2025).

Content classification. Content-level tasks focus on posts, threads, or conversations, predicting whether an interaction will lead to or amplify harm. Examples include forecasting whether a comment

thread will derail into personal attacks (Nonaka and Yoshida, 2025; Altarawneh et al., 2023), whether a post will trigger cyberbullying (Hosseinmardi et al., 2015), or whether a tweet will spark widespread toxic replies (Al-Merekhi et al., 2020).

3.2 Regressing ASB

Regression tasks predict continuous measures of harm offering finer-grained signals that reflect intensity or likelihood.

Actor regression. At the actor level, models forecast continuous user outcomes, such as how a user’s harmful behavior may evolve, their probability of returning after moderation, or their trajectory of engagement with toxic communities (Chelmiss and Yao, 2019; Levy et al., 2022).

Content regression. At the content level, models estimate harm scores for posts or conversations, such as the predicted intensity of harassment in a thread (Dahiya et al., 2021; Meng et al., 2023), expected hate triggered by a new post, or diffusion strength of harmful narratives (Han et al., 2021).

Classification often over-simplifies behaviors, producing false positives on strong but benign language or false negatives on subtle harms. Mitigation: adopt multi-label or hierarchical schemes, add contextual metadata, and treat annotator disagreement as informative (Lambert et al., 2022; Kim et al., 2025). Regression suffers from label subjectivity and drift: harm scores vary with annotators, time, and platform norms, leading to uncalibrated outputs. Mitigation: apply calibration techniques, use multi-rater judgments, and train on longitudinal data to capture evolving behaviors (Meng et al., 2023; Gajo et al., 2023; Alharthi et al., 2025).

3.3 Timing of Prediction: Ex-ante vs. Peeking

Beyond deciding what to predict, you must also decide **when** to predict. A detailed overview of feature-based model strategies and formulations is provided in Appendix C.

Ex-ante prediction Ex-ante models predict harm *at or before* content is posted, using only static features such as text, images, user metadata, or network structure. For example, some models predict whether an Instagram post will trigger cyberbullying using only the image, caption, and poster’s social graph (Hosseinmardi et al., 2015, 2016). Context-aware variants add structural or personality features while remaining anticipatory (Han et al., 2021; Liu et al., 2024a; Solovev and Pröllochs, 2023). Ex-ante strategies are especially com-

mon in classification tasks because they enable immediate moderation.

Peeking prediction. Peeking strategies allow models to observe early interactions—such as initial replies—before predicting. This often improves accuracy but introduces delay, making them ideal for early warning systems. Examples include forecasting hate intensity in threads or detecting cyberbullying after initial comment patterns (López-Vizcaíno et al., 2023a; Solovev and Pröllochs, 2023). Progressive approaches update predictions as data arrives and can stop once confidence is sufficient (Al-Merekhi et al., 2020; Meng et al., 2023).

Ex-ante and peeking strategies present distinct trade-offs for ASB prediction. Ex-ante approaches enable instant action but rely only on static features, often reducing accuracy and increasing false positives. In contrast, peeking strategies achieve higher accuracy by incorporating early conversational signals but delay intervention, risking harmful escalation. A practical solution, inspired by principles of information diffusion, is an a priori strategy: treat ex-ante predictions as low-confidence hypotheses (e.g., soft warnings or reduced reach) that are progressively refined as interaction evidence accumulates (Zhou et al., 2022). This hybrid approach ensures moderation that is both timely and adaptive.

4 Step 2: Select the data

After defining the task, the next step is to decide whether to build a new dataset or rely on existing resources. This section synthesizes, based on our systematic review of ASB prediction studies, the main data collection procedures reported in the literature and the characteristics of publicly available datasets employed.

4.1 Collecting your own data

The following analysis framework builds on the collection strategies outlined by Bonaldi et al. (2024).

Crawling. Crawling refers to the automated collection of real content from online platforms such as social networking sites (e.g., YouTube (Kennedy et al., 2020a), Twitter/X (Saveski et al., 2021; Alharthi et al., 2025)), discussion forums (e.g., Wikipedia Talk Pages (Altarawneh et al., 2023; Nonaka and Yoshida, 2025), Reddit (Yu et al., 2024; Hickey et al., 2025)), or news platforms (e.g., CNN, NYT (Han et al., 2021; Liu et al., 2024a)). This approach is particularly attractive because it can yield large-scale and diverse data samples, capturing a wide variety of contexts and

interaction styles. However, the way content is retrieved imposes important constraints. Most crawling procedures either (i) rely on selected “representative” keywords to filter and retrieve relevant posts, or (ii) prioritize content already highlighted by platform-specific popularity metrics (e.g., top-level comments on Reddit or trending videos on YouTube). While the first method helps to collect data that directly reflects known forms of ASB, it tends to suffer from limited lexical coverage and reduced sensitivity to alternative rhetorical strategies. Moreover, in the context of ASB prediction, even non-aggressive or neutral behaviors can evolve into toxic interactions (Al-Merekhi et al., 2020; Solovev and Pröllochs, 2023), making keyword-based sampling insufficient. Conversely, using more generic content without pre-filtering can expand lexical diversity and contextual coverage, but it introduces a new layer of complexity. Such content must undergo a meticulous curation process that depends heavily on each platform’s structure and interaction dynamics—for instance, retweet cascades on Twitter or multi-turn, threaded exchanges on Reddit. This complexity introduces additional limitations: platform-specific interaction patterns can bias the data, the resulting dataset may present imbalanced or sparse evidence of the targeted phenomena, and meaningful cross-platform comparisons become substantially more challenging. Finally, crawled data is often ephemeral—posts and accounts can be deleted or modified—hindering long-term reproducibility and making it difficult for future researchers to replicate or validate findings (Klubicka and Fernández, 2018; Florio et al., 2020).

Hybrid and fully automated collection. Hybrid approaches combine automated collection methods (such as web crawling, platform APIs, or bulk dataset ingestion) with manual interventions in the data collection pipeline (such as language adaptation or dataset post-editing). This allows researchers to scale up data acquisition while still ensuring that the collected material matches the domain and task requirements. In Kim et al. (2025), for example, an existing dataset was automatically ingested and supplemented with additional Korean SNS conversations. To increase coverage, corpora from AI Hub were integrated, with only the general and sexual conversation subsets retained.

Crowdsourcing and **nichesourcing** are established data collection strategies in related domains dealing with abusive language (Bonaldi et al., 2024;

Ollagnier et al., 2024), but, to the best of our knowledge, they have not yet been applied to ASB prediction. Crowdsourcing relies on non-expert annotators to generate, rephrase, or simulate ASB content and practices. This approach enables the collection of large quantities of data at relatively low cost. However, the resulting material often reflects simpler argument patterns and limited rhetorical diversity (Qian et al., 2019), which restricts its usefulness for modeling nuanced precursors of ASB. Nichesourcing, by contrast, relies on expert contributors—such as specialists, moderators, or trained practitioners—to create or simulate ASB content. This strategy yields data of higher quality, with richer syntactic variety and greater semantic complexity, particularly valuable for capturing subtle cues or early warning signs of harmful behavior. Its main drawback is high cost and limited scalability, which make it difficult to assemble large datasets (Chung et al., 2019).

Crawling remains the most common procedure for collecting ASB data in the wild. However, crowdsourcing and nichesourcing, while not widely explored for this purpose, offer promising alternatives. Well-designed crowdsourcing tasks, such as scenario-based role-playing exercises (Ollagnier, 2024), could elicit a wide range of interpersonal behaviors that may escalate into toxicity, thereby enriching the triggers considered in ASB prediction models. Likewise, nichesourcing—drawing on domain experts—could capture rare or highly specific interaction patterns that precede harmful outcomes, providing high-quality examples to strengthen computational modeling.

4.2 Choosing from existing datasets

Rather than collecting new data, a practical alternative is to leverage existing datasets. We review the publicly available resources identified in our survey of ASB prediction, summarized in Table 1.

Shape of the interactions. A crucial dimension to consider is the structure of interactions captured in a dataset, as this directly determines the types of ASB tasks it can support. Some datasets focus on *isolated content*, such as individual tweets or toxic news comments, which are treated as independent units without any interaction history; these are commonly used to address tasks such as harm emergence, behavioral risk, and proactive moderation (Kennedy et al., 2020a; Han et al., 2021; Irani et al., 2021). Others capture *local interaction* patterns, for instance short-lived exchanges like immediate replies or audience responses, which are often leveraged in behavioral risk or early harm

detection tasks within platforms such as Reddit (Yu et al., 2024; Song et al., 2025; Hickey et al., 2025). A third group encompasses *conversational flow* datasets that preserve multi-turn, threaded discussions, as seen in Reddit threads (Bao et al., 2021) or Wikipedia Talk Pages (Altarawneh et al., 2023; Nonaka and Yoshida, 2025), making them particularly suitable for modeling discourse evolution and for tasks such as early harm detection or proactive moderation. Finally, some datasets reflect *networked spread*, modeling how harmful content propagates through user networks—for example, via retweet cascades on Twitter—thereby enabling studies of diffusion processes and influence dynamics relevant to harm emergence or harm propagation (Saveski et al., 2021; Dahiya et al., 2021).

Targets of hate. As in related domains dealing with abusive language, ASB prediction datasets often focus on phenomena directed toward specific, and often multiple, targeted minorities. For example, Kennedy et al. (2020a) propose a rich annotation scheme with eight broad target identity groups (race/ethnicity, religion, national origin or citizenship, gender, sexual orientation, age, disability, and political ideology) and further refines these into 42 specific subgroups. In a complementary direction, Lambert et al. (2022) classify deleted comments based on normative violations, including misogynistic content, hate speech targeting racial or sexual minorities, attacks over opposing political views (e.g., pro- or anti-Trump), as well as abusive comments directed at moderators or individuals, such as name-calling or accusations that others are “too sensitive”. Other datasets capture more niche contexts: Meng et al. (2023) focus on anti-Asian hate, reflecting culture- and ethnicity-related hostility, while Gajo et al. (2023) examine incel forums, a setting where hate speech manifests through misogynistic and gendered discourse.

Types of hate addressed. Beyond specifying targets, many datasets classify hate speech by the *form* or *type* of negative reference expressed toward those groups. Building on taxonomies such as Chung et al. (2019), these categories capture whether the utterance manifests as insults, stereotyping, dehumanization, threats, or calls to violence. For instance, Kennedy et al. (2020a) annotate ten fine-grained dimensions, including (dis)respect, humiliation, inferior status, violence, genocide, and attack/defense. In contrast, broader schemes like Saveski et al. (2021) emphasize

Dataset	Size	Source(s)	Lang.	Target	Add. Info
Isolated Content					
Kennedy et al. (2020a)	135,556	YouTube, Twitter, Reddit	EN	✓	✓
Han et al. (2021)	5,571	Patch website	EN	×	✓
Irani et al. (2021)	40,000	Twitter	EN/ES	×	×
Lambert et al. (2022)	5,059	Reddit	EN	✓	✓
Solovev and Pröllochs (2023)	691,237	Twitter	EN	×	✓
Liu et al. (2024a)	5,571	Patch website	EN	×	✓
Conversational Flow					
Bao et al. (2021)	2,388	Reddit	EN	×	×
Meng et al. (2023)	890,372	Twitter	EN	✓	✓
Altarawneh et al. (2023)	11,030	WTP/CMV	EN	×	✓
Gajo et al. (2023)	5,503	Incels.is / Il forum dei brutti	EN/IT	✓	✓
Alharthi et al. (2025)	2,369	Twitter	EN	×	✓
Kim et al. (2025)	15,000	Chat platforms	KO	✓	×
Nonaka and Yoshida (2025)	11,030	WTP/CMV	EN	×	✓
Local Interaction					
Yu et al. (2024)	34,115	Reddit	EN	×	✓
Song et al. (2025)	5,724	Reddit	EN	×	✓
Hickey et al. (2025)	50,003	Reddit	EN	×	✓
Networked Spread					
Saveski et al. (2021)	1.18M	Twitter	EN	×	✓
Dahiya et al. (2021)	4,533	Twitter	EN	×	×

Table 1: Overview of publicly available datasets organized by interaction type.

generic toxicity markers such as rudeness or disrespect. Domain-specific datasets also highlight particular forms: Liu et al. (2024a) and Han et al. (2021) classify hate-related news comments into subtypes such as homicide or kidnapping, while Alharthi et al. (2025) provide aggregate indicators including counts of insults, profanity, threats, and identity attacks. Other resources focus on structural consequences of hate, for example examining how misogyny reinforces systemic oppression (Gajo et al., 2023; Hickey et al., 2025).

Languages. Most reported datasets are based on English-language corpora, with only a few extending to other languages, such as Korean (Kim et al., 2025). Others take a bilingual perspective—e.g., English–Spanish (Irani et al., 2021) and English–Italian (Gajo et al., 2023)—enabling cross-lingual and comparative analyses.

Additional information. Several datasets incorporate rich auxiliary signals that characterize conversations and users in greater detail. For example, Nonaka and Yoshida (2025) provide pairs of hate speech and counterspeech, further annotated according to the hater’s subsequent behavior (no reentry, hateful reentry, or non-hateful reentry). Other resources, such as Solovev and Pröllochs (2023), enrich each instance with a broad set of metadata, including text complexity, temporal information,

engagement indicators, and lexical features (e.g., moral and emotional content). Likewise, Yu et al. (2024) supply conversation-level attributes such as hate scores, subreddit tags, and reply-level labels. Some datasets also document annotator characteristics and target groups: for instance, Kennedy et al. (2020a) report annotator demographics and specify 40 fine-grained target subgroups.

Dataset selection is primarily guided by the target task, as different tasks require specific features (e.g., conversational flow for harm propagation). However, selection should also consider methodological factors such as construct validity, biases, and evidential support. To aid in this process, Liu et al. (2024b) offer guiding questions that clarify what capabilities a dataset measures. Practical aspects like dataset size and coverage also matter: few-shot prompting may suffice for small datasets, while fine-tuning typically demands larger corpora. Rich metadata—such as multiple annotator perspectives—can be especially valuable for subjective tasks, as explored in Frenda et al. (2025). Finally, transparent annotation procedures are critical, as they directly influence reliability and validity (see Appendix D). Beyond task-specific corpora, datasets from related domains (e.g., toxic language detection, cyberbullying, misinformation) should be systematically scrutinized, as their reuse can expand coverage and enhance comparability across ASB prediction research.

5 Step 3: Evaluate

In this step, we review how the literature evaluates ASB prediction models, focusing on the task

types outlined in 2.2. Since these tasks are most often formulated as either classification or regression problems, the choice of evaluation strategies naturally varies with the formulation.

5.1 Evaluating classification

Most ASB prediction studies formulate tasks such as early harm detection, harm emergence prediction, or proactive moderation as classification problems (Saveski et al., 2021; Lambert et al., 2022; Hickey et al., 2025). When ground-truth labels are available, models are commonly assessed using *accuracy*, *precision*, *recall*, and *F1-score*, often complemented by *ROC curves* and the *area under the ROC curve* to capture performance across decision thresholds. Several works report *macro-averaged metrics* (e.g., macro F1 or macro precision) to handle the class imbalance typical of ASB datasets, while others employ *latency-aware measures*, such as latency-weighted F1 (López-Vizcaíno et al., 2021) or time-aware precision (López-Vizcaíno et al., 2023a), to better capture the timing sensitivity required in early detection tasks. In ranking-based evaluation settings, measures such as *MAP@20* or *HITS@20* are used to assess whether models correctly prioritize high-risk content or users (Makkar and Chakraborty, 2020; Masud et al., 2021). *Complementary cumulative distribution functions* have also been applied to visualize performance across thresholds (Solovev and Pröllochs, 2023). Finally, some studies benchmark models against human judgment to verify performance; for example, Kim et al. (2025) introduce the *Human-to-Model Ratio*, which evaluates how closely a model’s detection speed aligns with that of human annotators in early detection scenarios. Complementing these quantitative measures, qualitative error analysis is often employed to reveal systematic weaknesses or biases in model predictions (Yu et al., 2024; Song et al., 2025).

5.2 Evaluating Regression

When ASB prediction tasks are framed to produce continuous outputs evaluation typically relies on metrics that capture both correlation and error. Correlation-based measures, such as the Pearson or Spearman correlation coefficients (Dahiya et al., 2021; Meng et al., 2023), assess how well the predicted scores align with the ground truth trends, providing insight into the model’s ability to rank or follow underlying dynamics. Complementarily, error-based measures such as Root Mean Square

Error, Mean Absolute Percentage Error, Symmetric Mean Absolute Percentage Error, or L2 loss quantify the magnitude of prediction errors (Wu et al., 2022; Hebert et al., 2023), thus revealing how close the predictions are in absolute terms. Some studies additionally report coefficients of determination (R-squared) to capture the proportion of variance explained by the model (Han et al., 2021; Liu et al., 2024a), or information-theoretic criteria such as log-likelihood scores and the Bayesian Information Criterion to balance goodness of fit with model complexity (Levy et al., 2022).

ASB prediction research remains constrained by an overreliance on standard classification and regression metrics and by the lack of context-sensitive evaluation protocols. While such metrics are useful, they are insufficient in high-stakes settings where error type, fairness, and user trust matter. Building on principles for systematic benchmark design (Liu et al., 2024b), best practices for evaluating ASB prediction should include cost-sensitive and asymmetric error measures, calibration and subgroup analyses, and human-centered criteria such as perceived harm and legitimacy (Olteanu et al., 2017). Metrics drawn from adjacent domains—such as diffusion or reach-based measures—can further capture the temporal and social impact of predicted harms (Zhou et al., 2022).

6 Open Challenges in ASB Prediction

Standardized Benchmarking. ASB prediction lacks standardized benchmarks and evaluation protocols, both globally and within predictive settings. Current tasks are underspecified, rely on heterogeneous datasets, and neglect cultural and pragmatic variation, which undermines reproducibility and cross-platform generalization. To address this, the field must move beyond ad hoc evaluations toward linguistically and culturally grounded frameworks that account for the temporal, structural, and ethical dimensions of ASB. Promising directions include meta-collection of existing resources (Piot et al., 2024) and the adaptation of Evidence-Centered Benchmark Design (Liu et al., 2024b), enabling systematic evidence collection about model capabilities. Incorporating insights from linguistics (e.g., speech act theory, relevance theory, dynamic and epistemic logic) and cultural psychology (e.g., the GLOBE framework (Karinshak et al., 2024)) would enhance both the modeling and the evaluation of a model’s ability to capture communicative intentions, conversational dynamics, and cultural variation. Developing such benchmarks is critical for protocols that are rigorous, comparable, and

sensitive to the diversity of online harms. Lessons from related shared tasks—such as *eRisk Lab*⁴ and *PAN Author Profiling*⁵—show how community-driven evaluations can foster standardization, comparability, and faster progress toward deployable solutions in ASB prediction.

Multimodal and Pragmatic Representation Learning.

Although our review considered ML studies broadly, ASB prediction remains dominated by text-centric features such as lexical cues or sentiment markers, often overlooking the multimodal and pragmatic dimensions of online interactions. As noted in Section 4, even neutral or non-aggressive behaviors can escalate into toxicity, underscoring the need to model evolving interaction dynamics (Ollagnier, 2024). Harmful behavior also frequently manifests through images, videos, memes, or subtle pragmatic cues such as politeness strategies, idioms, and conversational flow (Fried et al., 2023). Addressing these phenomena requires advances beyond current text-only pipelines, including the encoding of pragmatic knowledge, the integration of heterogeneous information networks linking users, posts, and contexts, and the development of graph-based architectures capable of combining multimodal and discourse-level signals (Wang et al., 2024). In parallel, recent progress in LLMs offers an additional research avenue, as such models provide richer contextual representations that may support predictive modeling when appropriately grounded (Albladi et al., 2025). Progress across these complementary directions is essential for developing ASB prediction models that are not only accurate but also sensitive to the complex and gradual ways in which harmful behavior develops and escalates.

Interpretability and Human Oversight. Modern ASB prediction models—such as transformers, GNNs, or multimodal systems—achieve strong performance but remain opaque (Mathew et al., 2021). In predictive settings, where errors directly impact user rights and content visibility, such opacity poses serious challenges to accountability. Regulations like the GDPR stress a “right to explanation” (Council, 2016), making transparency essential to avoid unfair sanctions or missed harms. Existing explainability methods provide limited value: they often reduce to feature importance, ig-

⁴<https://erisk.irlab.org/>

⁵<https://pan.webis.de/clef21/pan21-web/author-profiling.html>

nore cultural and linguistic nuance, and lack consistent evaluation (Nguyen et al., 2021). A key challenge is to design explanation frameworks that incorporate pragmatic reasoning, cultural grounding, and user feedback. Human-in-the-loop oversight can improve transparency and reduce bias. Progress requires interactive workflows, metrics for human–AI alignment, and governance models that ensure fairness and trust in high-stakes moderation (Kotarctic et al., 2022).

Global Challenges in ASB Prediction. Across the literature, several recurring challenges echo broader concerns in tackling ASB. First, **language and cultural diversity** remains limited, with most datasets English-centric and poorly generalizing across regions and communities (Alkomah and Ma, 2022). Second, the **granularity of labeling** is often too coarse, obscuring subtle forms of abuse such as implicit hate or trolling; richer annotation schemes (e.g., multi-label taxonomies, continuous ratings, or annotator demographics) would enable more precise modeling (Vidgen et al., 2019; Kennedy et al., 2020a; Kirk et al., 2022). Third, **data scope and representativeness** are constrained by small, biased datasets that underrepresent marginalized voices and multimodal contexts, as supported in Table 1. Finally, the problem of **evolving behavior and concept drift** means that models trained on static corpora struggle against rapidly changing discourse, slang, and evasion tactics (Florio et al., 2020). Together, these challenges highlight the need for multilingual, fine-grained, cross-platform, representative, and adaptive approaches if ASB prediction is to be both robust and socially responsible.

7 Conclusion

We presented an in-depth review of the emerging field of ASB prediction, introducing a taxonomy that structures existing ML research into five core task types, together with guidance on task design, dataset curation, and evaluation. We further identified key open challenges, including the creation of standardized benchmarks, advances in multimodal and pragmatic representation learning, and the development of explainable, human-centered systems. Our aim is to provide both a consolidated overview of current progress and a foundation for future work, offering a structured entry point for researchers, developers, and practitioners seeking to advance the predictive understanding and mitigation of harmful online behavior.

8 Limitations

To an external observer, the number of papers reviewed in this study may seem limited. However, this reflects both the emerging status of ASB prediction as a distinct research subfield and our focused inclusion criteria. We considered only machine-learning-oriented studies that define a clear prediction task—such as classification or regression—supported by a dataset, modeling framework, or empirical evaluation. This focus ensures that the review highlights work advancing predictive methodologies for ASB. Our search covered peer-reviewed conferences, journals, and preprints indexed in major databases (Google Scholar, DBLP, Scopus, IEEE Xplore, SpringerLink, and ScienceDirect), following established practices in abusive language research (e.g., (Tontodimamma et al., 2021; Jahan and Oussalah, 2023)).

A further limitation arises from restricted access to closed-access venues, leading to possible underrepresentation of research outside open-access repositories or preprint archives. Despite extensive keyword-driven searches, some relevant studies using alternative terms (e.g., “harmful content forecasting”, “online risk modeling”) may have been missed. To reduce omissions, the author supplemented automated searches with a curated archive of domain-relevant publications. As the field grows, future surveys would benefit from broader interdisciplinary coverage and more inclusive search strategies to fully capture the diversity of predictive work on ASB online.

9 Ethical Considerations

Research on ASB prediction inevitably involves the collection, annotation, and modeling of harmful or offensive content, which poses ethical risks for both researchers and affected users. Annotation tasks, in particular, can expose annotators and development teams to distressing material, potentially impacting mental health. In line with prior recommendations (Vidgen et al., 2019; Kirk et al., 2022), projects involving ASB data should incorporate protective measures, including rotating exposure schedules, mental health support resources, and the use of clear content warnings.

Privacy and data ethics are equally critical in this domain, especially when models are trained on user-generated content from social media platforms. Even with anonymization, re-identification risks remain—particularly for marginalized or vulnerable

communities whose linguistic patterns or behavioral markers may be unique, underrepresented, or stigmatized. These concerns are amplified in predictive tasks such as user risk profiling, where models infer the likelihood of individuals engaging in or becoming targets of ASB based on prior behavior, linguistic cues, and engagement patterns. While such approaches may support early intervention or moderation, they also introduce significant ethical risks, including surveillance-like monitoring, reputational bias, and opaque classification processes. In proactive moderation and early-warning systems, false positives or disproportionate targeting may further marginalize at-risk populations, undermining the very goals of safety and fairness. To mitigate these risks, ASB prediction systems must adhere to core principles of fairness, transparency, and accountability, as emphasized in regulatory frameworks such as the GDPR (Council, 2016), and should align with ongoing efforts in explainable AI (Mehta and Passi, 2022) to ensure that model decisions are interpretable and contestable.

Finally, as generative and predictive models become increasingly integrated into real-time moderation workflows, issues of responsibility and accountability become paramount. These systems should not operate autonomously in high-stakes settings. Instead, human-in-the-loop oversight, transparent decision-making processes, and strong governance mechanisms are essential to prevent overreach, hallucinated inferences, or unintended harm amplification. Aligning technical performance with ethical and social responsibility must remain a central goal in the development and deployment of ASB prediction systems.

References

- Hind A. Al-Merekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. [Are these comments triggering? predicting triggers of toxicity in online discussions](#). In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 3033–3040. ACM / IW3C2.
- Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl D. Seals. 2025. [Hate speech detection using large language models: A comprehensive review](#). *IEEE Access*, 13:20871–20892.
- Raneem Alharthi, Rajwa Alharthi, Ravi Shekhar, Aiqi Jiang, and Arkaitz Zubiaga. 2025. [Will I get hate](#)

- speech predicting the volume of abusive replies before posting in social media. *CoRR*, abs/2503.03005.
- Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Inf.*, 13(6):273.
- Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation derailment forecasting with graph convolutional networks. *CoRR*, abs/2306.12982.
- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, 5 edition. American Psychiatric Publishing.
- Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-asian hateful users on twitter during COVID-19. pages 4655–4666.
- Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1134–1145. ACM / IW3C2.
- Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3480–3499. Association for Computational Linguistics.
- Jo Brown, Amanda J Broderick, and Nick Lee. 2007. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing*, 21(3):2–20.
- Charalampos Chelmiss and Mengfan Yao. 2019. **Minority report: Cyberbullying prediction on instagram**. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*, pages 37–45. ACM.
- Justin Cheng, Michael S. Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. **Anyone can become a troll: Causes of trolling behavior in online discussions**. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 1217–1230. ACM.
- Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. **Antisocial behavior in online discussion communities**. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 61–70. AAAI Press.
- Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. 2019. **#youtoo? detection of personal recollections of sexual harassment on social media**. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2527–2537. Association for Computational Linguistics.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. **CONAN - COUNTER NARRATIVES THROUGH NICHE-SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.
- Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. **Italian counter narrative generation to fight online hate speech**. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. **The echo chamber effect on social media**. *Proc. Natl. Acad. Sci. USA*, 118(9):e2023301118.
- European Council. 2016. EU regulation 2016/679 general data protection regulation (gdpr). *Official Journal of the European Union*, 59(6):1–88.
- J. Ignacio Criado, Rodrigo Sandoval-Almazán, and J. Ramón Gil-García. 2013. **Government innovation through social media**. *Gov. Inf. Q.*, 30(4):319–326.
- Snehil Dahiya, Shalini Sharma, Dhruv Sahnun, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. **Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter**. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2732–2742. ACM.
- Gabriele Etta, Matteo Cinelli, Niccolò Di Marco, Michele Avalle, Alessandro Panconesi, and Walter Quattrociocchi. 2024. **A topology-based approach for predicting toxic outcomes on twitter and youtube**. *IEEE Trans. Netw. Sci. Eng.*, 11(5):4875–4885.
- Gabriele Etta, Emanuele Sangiorgio, Niccolò Di Marco, Michele Avalle, Antonio Scala, Matteo Cinelli, and Walter Quattrociocchi. 2022. **Characterizing engagement dynamics across topics on facebook**. *CoRR*, abs/2211.15988.
- Tope Christopher Falade, Niloofar Yousefi, and Nitin Agarwal. 2024. **Toxicity prediction in reddit**. In *30th Americas Conference on Information Systems: Elevating Life through Digital Social*

- Entrepreneurship, AMCIS 2024, Salt Lake City, UT, USA, August 15-17, 2024*. Association for Information Systems.
- Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. [Time of your hate: The challenge of time in hate speech detection on social media](#). *Applied Sciences*, 10(12).
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. [Perspectivist approaches to natural language processing: a survey](#). *Lang. Resour. Evaluation*, 59(2):1719–1746.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. [Pragmatics in language grounding: Phenomena, tasks, and modeling approaches](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12619–12640. Association for Computational Linguistics.
- Paolo Gajo, Arianna Muti, Katerina Korre, Silvia Bernardini, and Alberto Barrón-Cedeño. 2023. [On the identification and forecasting of hate speech in in-celdom](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023, Varna, Bulgaria, 4-6 September 2023*, pages 373–384. INCOMA Ltd., Shoumen, Bulgaria.
- Anatoliy Gruzd, Philip Mai, and Zahra Vahedi. 2020. Studying anti-social behaviour on reddit with communalytic. *The SAGE Handbook of Social Media Research Methods*, pages 503–520.
- Songqiao Han, Hailiang Huang, Jiangwei Liu, and Shengsheng Xiao. 2021. [American hate crime trends prediction with event extraction](#). *arXiv preprint arXiv:2111.04951*.
- Caroline Haythornthwaite. 2023. [Moderation, networks, and anti-social behavior online](#). *Social Media + Society*, 9(3):20563051231196874.
- Liam Hebert, Lukasz Golab, and Robin Cohen. 2023. [Predicting hateful discussions on reddit using graph transformer networks and communal context](#). volume abs/2301.04248.
- Daniel Hickey, Daniel M. T. Fessler, Matheus Schmitz, Kristina Lerman, and Keith Burghardt. 2025. [The peripatetic hater: Predicting movement among hate subreddits](#). In *Proceedings of the Nineteenth International AAI Conference on Web and Social Media, June 23-26, 2025, Copenhagen, Denmark*, pages 786–803. AAAI Press.
- Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. [Prediction of cyberbullying incidents on the instagram social network](#). *CoRR*, abs/1508.06257.
- Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. [Prediction of cyberbullying incidents in a media-based social network](#). In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 186–192. IEEE Computer Society.
- Mia Mohammad Imran, Robert Zita, Rebekah Copeland, Preetha Chatterjee, Rahat Rizvi Rahman, and Kostadin Damevski. 2025. [Understanding and predicting derailment in toxic conversations on github](#). *CoRR*, abs/2503.02191.
- Darius Irani, Avyakta Wrat, and Silvio Amir. 2021. [Early detection of online hate speech spreaders with learned user representations](#). In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2004–2010. CEUR-WS.org.
- Md Saroar Jahan and Mourad Oussalah. 2023. [A systematic review of hate speech automatic detection using natural language processing](#). *Neurocomputing*, 546:126232.
- Richard Kahn and Douglas Kellner. 2004. [New media and internet activism: From the 'battle of seattle' to blogging](#). *New Media Soc.*, 6(1):87–95.
- Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. [LLM-GLOBE: A benchmark evaluating the cultural values embedded in LLM output](#). *CoRR*, abs/2411.06032.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020a. [Constructing interval variables via faceted rasch measurement and multi-task deep learning: a hate speech application](#). *CoRR*, abs/2009.10277.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. [Constructing interval variables via faceted rasch measurement and multi-task deep learning: a hate speech application](#). *CoRR*, abs/2009.10277.
- Sahrish Khan, Rabeeh Ayaz Abbasi, Muddassar Azam Sindhu, Sachi Arafat, Akmal Saeed Khattak, Ali Daud, and Mubashar Mushtaq. 2024. [Predicting the victims of hate speech on microblogging platforms](#). *Heliyon*, 10(23).
- Dohyeon Kim, Taehoon Kim, and Jihoon Yang. 2025. [Early detection of online grooming with language models](#). In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC 2025, Catania International Airport, Catania, Italy, 31 March 2025 - 4 April 2025*, pages 963–970. ACM.
- Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. [Handling and presenting harmful](#)

- text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 497–510. Association for Computational Linguistics.
- Barbara Kitchenham. 2004. Procedures for performing systematic reviews. Technical report, Keele University, Department of Computer Science.
- Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. *CoRR*, abs/1805.04661.
- Ana Kotarcic, Dominik Hangartner, Fabrizio Gilardi, Selina Kurer, and Karsten Donnay. 2022. Human-in-the-loop hate speech classification in a multilingual context. *CoRR*, abs/2212.02108.
- Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational resilience: Quantifying and predicting conversational outcomes following adverse events. In *Proceedings of the Sixteenth International AAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 548–559. AAAI Press.
- Sharon Levy, Robert E. Kraut, Jane A. Yu, Kristen M. Altenburger, and Yi-Chia Wang. 2022. Understanding conflicts in online conversations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2592–2602. ACM.
- Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining, ICDM 2021 - Workshops, Auckland, New Zealand, December 7-10, 2021*, pages 967–974. IEEE.
- Jiangwei Liu, Xiangzhen Jia, You Wu, Jingshu Zhang, and Xiaohong Huang. 2024a. From news to knowledge: Predicting hate crime trends through event extraction from media content. *IAENG International Journal of Applied Mathematics*, 54(4).
- Yu Lu Liu, Su Lin Blodgett, Jackie C. K. Cheung, Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024b. ECBD: evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16349–16365. Association for Computational Linguistics.
- Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Thierry Artières, and Fidel Casheda. 2023a. Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, 23(10):4788.
- Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Thierry Artières, and Fidel Casheda. 2023b. Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, 23(10):4788.
- Manuel F. López-Vizcaíno, Francisco Javier Nóvoa, Victor Carneiro, and Fidel Casheda. 2021. Early detection of cyberbullying on social media networks. *Future Gener. Comput. Syst.*, 118:219–229.
- Sakshi Makkar and Tanmoy Chakraborty. 2020. *Hate speech diffusion in twitter social media*. Ph.D. thesis, IIT-Delhi.
- Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pages 504–515. IEEE.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.
- Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8).
- Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *Knowl. Based Syst.*, 275:110644.
- Meta. 2022. Facebook community standards. <https://transparency.fb.com/policies/community-standards/>. (2022c).
- David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Bmj*, 339.
- Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, and Ibrahim Kamel. 2019. Real-time detection of cyberbullying in arabic twitter streams. In *10th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2019, Canary Islands, Spain, June 24-26, 2019*, pages 1–5. IEEE.
- Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *CoRR*, abs/2305.04927.
- Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. 2021. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6.

- Kenya Nonaka and Mitsuo Yoshida. 2025. [Zero-shot prediction of conversational derailment with large language models](#). *IEEE Access*, 13:55081–55093.
- Anaïs Ollagnier. 2024. [Cyberagressionado-v2: Leveraging pragmatic-level information to decipher online hate in french multiparty chats](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4287–4298. ELRA and ICCL.
- Anaïs Ollagnier, Elena Cabrio, and Serena Villata. 2023a. [Harnessing bullying traces to enhance bullying participant role identification in multi-party chats](#). In *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*. Florida Online Journals.
- Anaïs Ollagnier, Elena Cabrio, and Serena Villata. 2023b. [Unsupervised fine-grained hate speech target community detection and characterisation on social media](#). *Soc. Netw. Anal. Min.*, 13(1):58.
- Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Valerio Basile. 2024. [Cyberagressionado-large: French multiparty chat dataset to address online hate](#). *Revue TAL : traitement automatique des langues*, 65(3):21–44.
- Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. [The limits of abstract evaluation metrics: The case of hate speech detection](#). In *Proceedings of the 2017 ACM on Web Science Conference, WebSci '17*, page 405–406, New York, NY, USA. Association for Computing Machinery.
- Mark C. Parent, Taylor D. Gobble, and Aaron Rochlen. 2019. [Social media behavior, toxic masculinity, and depression](#). *Psychology of Men & Masculinity*, 20(3):277–287. Epub 2018 Apr 23.
- Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. [Metahate: A dataset for unifying efforts on hate speech detection](#). In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 2025–2039. AAAI Press.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. [Resources and benchmark corpora for hate speech detection: a systematic review](#). *Lang. Resour. Evaluation*, 55(2):477–523.
- Nektaria Potha and Manolis Maragoudakis. 2014. [Cyberbullying detection using time series modeling](#). In *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, pages 373–382. IEEE Computer Society.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding, and William Yang Wang. 2019. [A benchmark dataset for learning to intervene in online hate speech](#). 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4754–4763. Association for Computational Linguistics.
- Walter Quattrocchi, Guido Caldarelli, and Antonio Scala. 2014. [Opinion dynamics on interacting networks: media competition and social influence](#). *Scientific Reports*, 4:4938.
- Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. [Prevalence and psychological effects of hateful speech in online college communities](#). In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*, pages 255–264. ACM.
- Martin Saveski, Brandon Roy, and Deb Roy. 2021. [The structure of toxic conversations on twitter](#). In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1086–1097. ACM / IW3C2.
- Kirill Solovev and Nicolas Pröllochs. 2023. [Moralized language predicts hate speech on social media](#). *PNAS nexus*, 2(1):pgac281.
- Xiaoying Song, Sharon Lisseth Perez, Xinchun Yu, Eduardo Blanco, and Lingzi Hong. 2025. [Echoes of discord: Forecasting hater reactions to counterspeech](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4892–4905. Association for Computational Linguistics.
- Billy Spann and Nitin Agarwal. 2023. [Predicting toxicity in reddit discussion threads](#). In *Proceedings of the 16th International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRIMS 2023)*.
- Sajedul Rahim Talukder and Bogdan Carbutar. 2018. [Abusniff: Automatic detection and defenses against abusive facebook friends](#). In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 385–394. AAAI Press.
- Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2024. [Beyond trial-and-error: Predicting user abandonment after a moderation intervention](#). *CoRR*, abs/2404.14846.
- Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. [Thirty years of research into hate speech: topics of interest and their evolution](#). *Scientometrics*, 126(1):157–179.
- Paraskevas Tsantaliotis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. [Defining and predicting troll vulnerability in online social media](#). *Soc. Netw. Anal. Min.*, 7(1):26:1–26:15.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. [Echo chambers: Emotional contagion and group polarization on facebook](#). *CoRR*, abs/1607.01032.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Ying Wang, Yingji Li, Yue Wu, and Xin Wang. 2024. [Exploring multiple hypergraphs for heterogeneous graph neural networks](#). *Expert Syst. Appl.*, 236:121230.

Xiao-Kun Wu, Tian-Fang Zhao, Lu Lu, and Wei-Neng Chen. 2022. Predicting the hate: A gstm model based on covid-19 hate speech datasets. *Information Processing & Management*, 59(4):102998.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2024. [Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech](#). In *Proceedings of the Eighteenth International AAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 1740–1752. AAAI Press.

Jiaqing Yuan and Munindar P. Singh. 2023. [Conversation modeling to predict derailment](#). In *Proceedings of the Seventeenth International AAI Conference on Web and Social Media, ICWSM 2023, Limassol, Cyprus, June 5-8, 2023*, pages 926–935. AAAI Press.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025. [Forecasting communication derailments through conversation generation](#). *CoRR*, abs/2504.08905.

Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2022. [A survey of information cascade analysis: Models, predictions, and recent advances](#). *ACM Comput. Surv.*, 54(2):27:1–27:36.

A Methodology of the Review

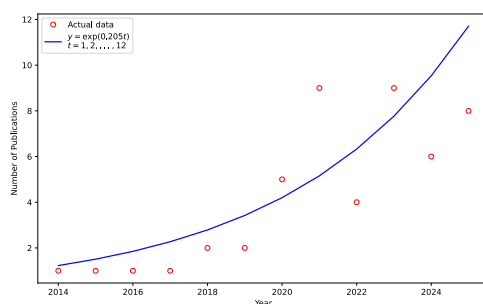


Figure 1: Number of publications on ASB prediction per year: observed and expected distribution.

Figure 1 illustrates the steady growth of research on ASB prediction, reflecting the increasing interest in proactive and forward-looking approaches to online harm. To ensure methodological rigor, transparency, and reproducibility, we conducted this survey following established guidelines for systematic literature reviews in computer science and engineering (Kitchenham, 2004), in conjunction with the PRISMA framework (Moher et al., 2009). The study selection process is summarized in Fig. 2.

The objective of this review is to provide a structured synthesis of progress to date in ASB prediction tasks within the ML literature, clearly distinguishing prediction-oriented approaches from the more mature body of work on detection and classification. We operationalize prediction as tasks that forecast future outcomes or risks using supervised ML techniques, rather than merely identifying existing harmful content. Accordingly, we include only peer-reviewed, openly accessible studies that adhere to this definition, while systematically excluding work limited to detection, identification, profiling, or post-hoc analysis without a forward-looking objective.

Following prior surveys of harmful language and online abuse (Poletto et al., 2021; Jahan and Oussalah, 2023), we retrieved candidate papers from major scholarly databases, including Google Scholar, DBLP, Scopus, IEEE Xplore, Springer-Link, and ScienceDirect. To streamline and centralize retrieval, we used the SciLEX aggregator⁶. Search queries combined prediction-oriented keywords (e.g., prediction, forecasting, early detection, propagation, temporal modeling, real-time prediction) with ASB-related terms (e.g., antisocial behaviour, cyberaggression, hate speech, cyberbullying, online abuse, toxicity, online grooming, hate crime). Searches were applied to titles and abstracts, restricted to English-language publications between 2010 and 2025, with the final search conducted on June 17, 2025.

The screening process proceeded in multiple stages. First, titles and abstracts were reviewed to remove irrelevant or tangential works. Second, full-text screening was conducted to verify that remaining studies met all inclusion criteria, particularly with respect to their predictive framing and ML-based methodology. Duplicate records retrieved across databases were removed. Each included paper was then annotated using a classifi-

⁶<https://github.com/Wimmics/SciLEX/>

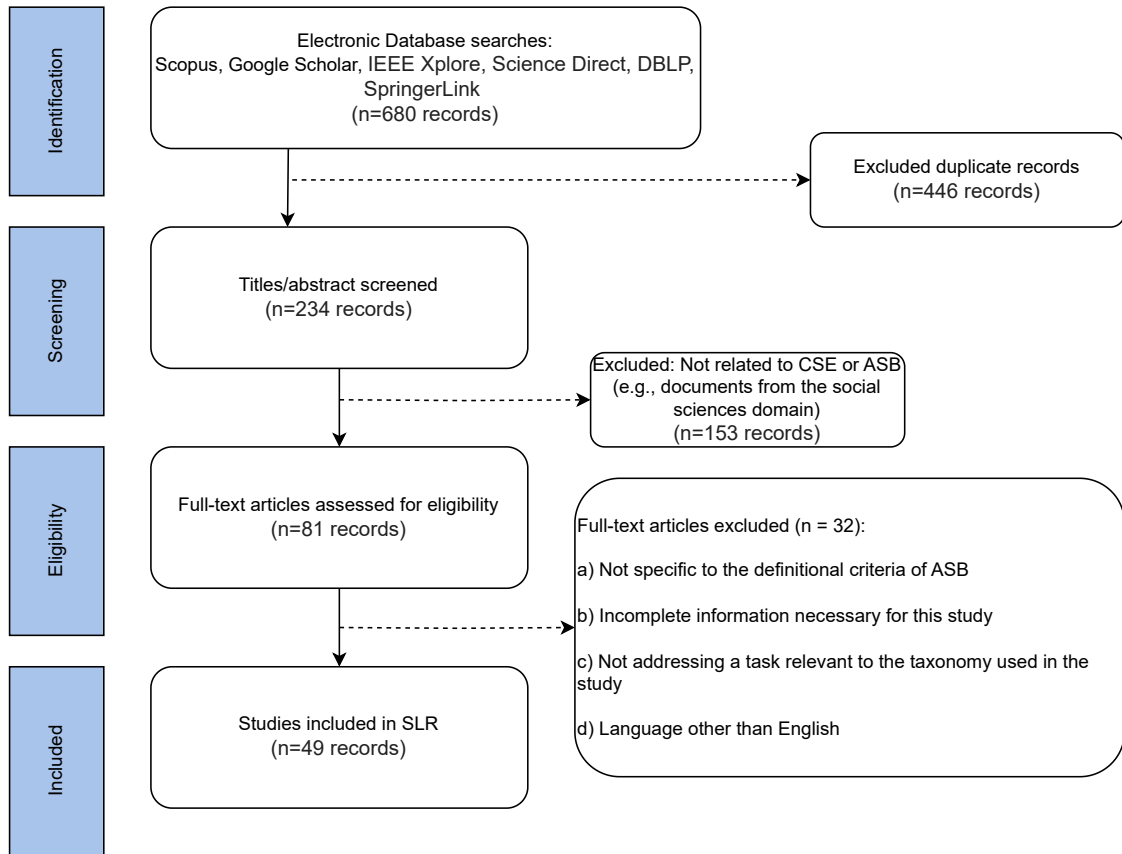


Figure 2: PRISMA flowchart for selection of antisocial behaviour prediction studies.

cation scheme capturing its disciplinary grounding (e.g., NLP, computational social science), primary research contribution (e.g., predictive modeling, resource creation, intervention design), methodological approach (e.g., neural, traditional ML, or hybrid), and the specific ASB prediction task addressed, following the taxonomy in Section 2.2.

This systematic and multi-stage process resulted in a final corpus of 49 papers, which form the empirical basis for our synthesis of the state of the art and the identification of open challenges and future directions in ASB prediction research.

B Task Category Distribution in ASB Prediction

The distribution of task categories in current ASB prediction research, illustrated in Figure 3, reveals a clear concentration around early harm detection and harm emergence prediction, which together represent the majority of studied scenarios. Early detection tasks are designed to anticipate escalation risks in conversations, flag toxic replies at their onset, or detect early signs of con-

flict based on minimal interaction cues (e.g., (Al-Merekhi et al., 2020; Mubarak et al., 2023; Kim et al., 2025)). Harm emergence tasks focus on forecasting whether harmful behavior will surface in the course of an interaction, such as predicting conversational derailment or the likelihood of toxic replies (e.g., (Nonaka and Yoshida, 2025; Altarawneh et al., 2023)). Harm Propagation Prediction forms a substantial secondary cluster, encompassing tasks like modeling the spread and virality of hate speech, estimating future hate intensity, or forecasting how toxic content diffuses across social networks (Makkar and Chakraborty, 2020; Meng et al., 2023; Alharthi et al., 2025). In parallel, behavioral risk prediction—though less frequent—covers user-centric forecasting, such as predicting recidivism, moderation outcomes, or the likelihood of becoming a hate speech amplifier (Potha and Maragoudakis, 2014; An et al., 2021; Khan et al., 2024). Lastly, proactive moderation support includes predictive tasks aimed at assisting moderation efforts in advance, such as estimating abuse likelihood at posting time, predicting

conversation resilience, or evaluating harassment severity through multimodal features (Talukder and Carburnar, 2018; Bao et al., 2021; Lambert et al., 2022). Overall, this landscape reflects a multi-dimensional field, with strong emphasis on early content-level interventions and growing attention toward user trajectories and longitudinal harm forecasting.

C Ex-ante Prediction versus Peeking Strategy

As summarized in Table 2, ex-ante strategies are more prevalent in the literature, particularly for classification tasks (Klubicka and Fernández, 2018; Masud et al., 2021), due to their appeal in proactive moderation. However, peeking and progressive strategies are gaining traction, especially in multi-turn conversation or cascade-based tasks where the temporal evolution of content is central.

D Annotators Training Procedure

Engaging with hate speech and counterspeech is known to have significant social and psychological consequences for those involved. This has led to the development of specific guidelines and best practices to ensure that such work is conducted responsibly (Chung et al., 2020; Ollagnier, 2024; Bonaldi et al., 2024). In the same way, research on ASB should adopt precautions similar to those recommended in other domains dealing with abusive language. Recognizing, labeling, and interpreting ASB in online content—whether for escalation forecasting, propagation modeling, or proactive moderation—requires not only clear technical guidelines but also domain-specific sensitivity and practical experience. Annotators are frequently exposed to distressing material and are tasked with making fine-grained judgments about nuanced patterns of harm. When they lack prior expertise in ASB labeling, targeted training becomes essential to help them develop the necessary skills and ensure consistent, high-quality annotations, as highlighted in prior work on abusive language annotation and harm labeling (Vidgen et al., 2019).

The most commonly reported procedure for training annotators includes the following steps:

- a) **Reviewing guidelines and documentation:** annotators first study platform or research-specific annotation guidelines and public documentation on harmful behaviors (e.g., definitions of harassment, trolling, or hate speech).

- b) **Learning from examples:** they are then exposed to curated examples of ASB labeling (including edge cases such as implicit hate or sarcasm) and, where relevant, examples of expert annotations for specific tasks (e.g., labeling early harm signals).
- c) **Practice sessions:** annotators perform trial labeling on a subsample of posts or conversation threads, receiving iterative feedback.
- d) **Discussion and calibration:** disagreements and difficult cases are discussed in regular meetings with an expert or lead annotator to achieve shared understanding and improve consistency.

Table 3 summarises the training steps reported in studies that explicitly describe how their annotators were trained, among those studies whose datasets are publicly available and documented. Importantly, because annotators are frequently exposed to harmful or distressing material, their well-being must be prioritized. Precautions include explaining the prosocial purpose of the work, limiting annotation sessions to a few hours per day, encouraging regular breaks, and providing structured opportunities to raise concerns or seek support (Vidgen et al., 2019; Kirk et al., 2022). These measures help ensure reliable annotations while safeguarding the mental health of those involved.

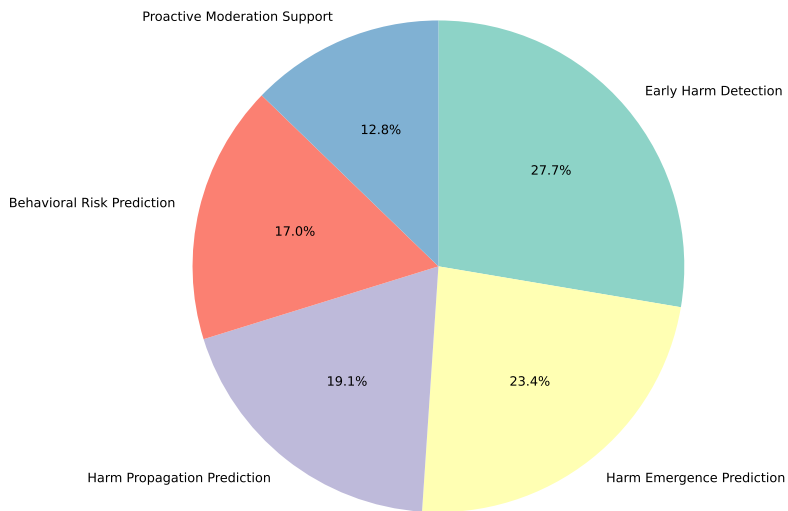


Figure 3: Distribution of ASB prediction task categories across reviewed papers.

Strategy	Formulation	Reference
Ex-ante	Classification	(Hosseinmardi et al., 2015, 2016; Klubicka and Fernández, 2018; Talukder and Carbunar, 2018; Mouheb et al., 2019; Makkar and Chakraborty, 2020; Masud et al., 2021; Bao et al., 2021; An et al., 2021; Irani et al., 2021; Mubarak et al., 2023; López-Vizcaíno et al., 2023b; Yuan and Singh, 2023; Spann and Agarwal, 2023; Altarawneh et al., 2023; Etta et al., 2024; Khan et al., 2024; Falade et al., 2024; Tessa et al., 2024; Kim et al., 2025; Imran et al., 2025; Nonaka and Yoshida, 2025; Zhang et al., 2025; Song et al., 2025; Hickey et al., 2025)
	Regression	(Kennedy et al., 2020b; Han et al., 2021; Wu et al., 2022; Levy et al., 2022; Hebert et al., 2023; Meng et al., 2023; Gajo et al., 2023; Yu et al., 2024; Liu et al., 2024a; Alharthi et al., 2025)
Peeking	Classification	(Potha and Maragoudakis, 2014; Cheng et al., 2015; Saveski et al., 2021; López-Vizcaíno et al., 2023a; Lin et al., 2021; Lambert et al., 2022; Solovev and Pröllochs, 2023)
	Regression	(Chelmis and Yao, 2019; Dahiya et al., 2021)

Table 2: Strategies and formulations of feature-based models.

Study	a	b	c	d
Kennedy et al. (2020a)	–	–	✓	–
Saveski et al. (2021)	✓	–	–	–
Bao et al. (2021)	✓	✓	✓	✓
Dahiya et al. (2021)	✓	–	–	–
Gajo et al. (2023)	✓	✓	✓	✓
Alharthi et al. (2025)	✓	✓	✓	✓
Kim et al. (2025)	✓	✓	–	–
Nonaka and Yoshida (2025)	✓	–	–	–
Song et al. (2025)	✓	✓	–	–
Hickey et al. (2025)	✓	✓	–	–

Table 3: Steps for annotators’ training in studies that explicitly describe them, as detailed in D.

Real-Time Mitigation of Negative Emotion in Customer Care Calls

Surupendu Gangopadhyay *

SoundHound AI

sgangopadhyay@soundhound.com

Mahnoosh Mehrabani *

SoundHound AI

mmehrabani@soundhound.com

Abstract

Speech emotion recognition (SER) is a compelling yet challenging research area with substantial practical relevance, particularly in enhancing human-machine interaction. Despite considerable progress in the field, the scarcity of realistic datasets that reflect real-world conditions makes it difficult to analyze system behavior in practice and can lead to degraded performance in industrial applications. In this study, we propose a system that detects negative emotions at each turn in a conversation by leveraging both linguistic and acoustic features. The approach is evaluated on real-world data, with a particular focus on identifying and responding to negative emotion in customer support scenarios. Designed for real-time application, the system is suitable for live deployment in call center environments. Furthermore, we propose an effective prompting strategy for using large language models (LLMs) as annotators, generating labeled data used to fine-tune small language models that achieve performance on par with the LLM used for annotation, while remaining suitable for real-time deployment.

1 Introduction

Automatic emotion recognition using modalities such as text, speech, image, and video has been an active research area for several decades, with significant advancements achieved through the application of increasingly sophisticated models. Recognizing emotions from speech is particularly valuable in diverse human-machine interactions, including conversational agents, virtual assistants, and interactive systems. Although numerous studies report promising evaluation outcomes on public benchmark datasets (Caralt et al., 2025; Shou et al., 2024; Xue et al., 2024; Fu, 2024; Lei et al., 2023; Stricker and Paroubek, 2024; Ma et al., 2023; Lin et al., 2023), the field continues to face a shortage

of realistic data and practical methods suitable for real-world deployment.

This study investigates speech emotion recognition (SER) in spoken dialogue systems, with a particular focus on enterprise customer support. In this domain, users interact with automated or virtual agents to resolve service or product-related issues. Accurate detection of emotions in such interactions can enhance the user experience, support timely intervention, and improve the overall effectiveness of automated support systems. The principal aim is to identify and appropriately respond to negative emotional states like anger or frustration by employing mitigation strategies, such as escalating the call to a human agent or incorporating more empathetic system responses.

A critical aspect of this work is the requirement for real-time processing, which significantly increases the complexity of the task. In this setting, "real-time" refers to the ability to detect emotions with low CPU latency as the conversation progresses, rather than after it concludes. This imposes strict constraints on system responsiveness, as predictions must be made quickly enough to support timely interventions during the interaction. Consequently, the deployment of highly complex models becomes impractical, as they often introduce delays that are incompatible with the demands of live dialogue systems. The rapid advancement and application of large language models (LLMs) have significantly influenced a wide range of research domains, including SER. While LLMs offer substantial benefits for classification tasks particularly in low-resource scenarios where limited or no labeled data is available their use is associated with considerable computational cost and latency. These limitations pose significant challenges for practical, large-scale SER model deployment. Therefore, in this work, we leverage the capabilities of LLMs in an offline setting to generate and curate training data for SER, thereby avoiding the latency and cost

*This work was done as part of Interactions LLC

constraints associated with real-time inference.

A core challenge in speech emotion recognition is the scarcity of comprehensive, high-quality datasets. Most publicly available corpora are limited in size and primarily consist of acted speech recorded in controlled settings, which lack the natural variability and spontaneity characteristic of real-world emotional expression. Furthermore, imbalanced emotion classes and insufficient demographic diversity impede the robustness and generalizability of trained models (Garg et al., 2024; Upadhyay et al., 2025; Lin et al., 2024). Although real-world, in-domain data is often available in industry, for example, through customer service interactions or voice-enabled platforms, such data are typically unlabeled and proprietary, and annotation at scale is both costly and time-consuming. To address this limitation, we leverage LLMs to generate labeled training data by guiding them with an annotation protocol originally developed for human labeling of evaluation data.

To identify negative emotions such as caller frustration, we integrate both linguistic and acoustic cues, as each provides complementary information. Linguistic indicators include semantic signals such as negative sentiment or intent, as well as explicit expressions such as angry language, complaints about the system, repeated requests, and swearing. Acoustic features capture prosodic and paralinguistic signals, including elevated pitch, increased volume, vocal strain, abrupt changes in loudness, and variations in speaking rate. By combining these modalities, the system is better equipped to detect subtle and context-dependent expressions of negative affect. The contributions of this paper are summarized as follows:

1. We develop a low CPU latency, efficient, production-ready solution for real-time detection of negative customer emotions, integrating both linguistic and acoustic cues in speech.
2. We leverage prompt tuning to utilize large language models (LLMs) as offline annotators for generating labeled training data, which is then used to fine-tune lightweight LLMs suitable for low-latency deployment.
3. We conduct a comparative evaluation of multiple modeling approaches on publicly available datasets and real-world industry data to assess performance under different conditions.

2 Related Work

Recent advances in large language models (LLMs) have shifted emotion recognition from traditional discriminative models to generative paradigms across modalities (Ma et al., 2025; Zhang et al., 2024; Broekens et al., 2023). Techniques such as chain-of-thought reasoning, contextual modeling, in-context learning, and fine-tuning have further improved automatic emotion recognition performance (Bhaumik and Strzalkowski, 2024; Zhang et al., 2023; Peng et al., 2024). While some studies explore multi-modal fusion of acoustic and textual features (Dutta and Ganapathy, 2025; Wu et al., 2025), most work remains text-centric. A growing body of research applies LLMs to speech emotion recognition (SER) (Xu et al., 2024; Gong et al., 2023; Chen et al., 2025), often relying on speaker identity, context, or auxiliary tasks like speaker prediction (Lei et al., 2023; Xue et al., 2024), which limits their suitability for real-time applications with unknown participants and limited context.

In task-oriented dialogues, emotion recognition typically prioritizes detecting negative states—such as frustration, anger, or dissatisfaction—rather than covering a broad emotional spectrum. Prior work has used LLMs with in-context learning to identify frustration from transcripts (Caralt et al., 2025), or relied on conversational patterns and agent behaviors to infer dissatisfaction. In this work, we focus on real-time SER for task-oriented conversational systems in commercial settings. Our goal is to detect negative emotional states using only the audio of the current customer utterance at inference time. To ensure practical deployment, we avoid reliance on dialogue history, speaker identity, or other contextual metadata that may be unavailable or introduce system complexity. The following sections detail our proposed method, dataset construction process, results, and discussion.

3 Methodology

We propose a low-cost, low-latency SER approach that is developed using spontaneous, in-domain data from real customer care calls. Large language models are employed as annotators to label production data, which is then used to fine-tune small language models (SLMs) for efficient inference.

Approach Name	Prompting Strategy	Context	Input	Input format
Simple	Simple prompt	None	Agent prompt + Current utterance	Audio
ZCOT_Simple	Zero-shot COT			
ZCOT_Context		Last 3 agent-customer interactions	Context + Agent prompt +	Text +
ZCOT_Detailed_Context		Last 3 agent-customer interactions, DSIDs and utterance sequence no.		
PAS_Detailed_Context	Plan and Solve			
RP_Detailed_Context	Roleplay			

Table 1: Summary of prompting approaches for using Gemini 2.0 Flash for annotation

3.1 Leveraging Large Language Models for Offline Data Annotation

Large language models (LLMs) have shown strong performance in various natural language and speech tasks, including SER. A key challenge in SER is the limited availability of high-quality, in-domain labeled data for training robust models. Traditional annotation methods, which rely on manual labeling, are time-consuming and costly, particularly for large datasets needed to capture the variability in natural speech.

To address this challenge, we adopt an offline approach in which a large language model, more precisely Gemini 2.0 Flash, is employed to automatically annotate training data for emotion recognition. The resulting annotated dataset is then used to fine-tune SLMs designed for efficient real-time prediction. This strategy enables us to benefit from the language understanding capabilities of LLMs while avoiding their high inference latency and computational requirements during deployment.

Initially, we compare the performance of different zero-shot prompting approaches to improve the performance of the model as annotator. These strategies are optimized over the ProEval test set discussed in Section 4.1. Table 1 shows the different prompting approaches that we used for our experiments. These approaches rely on using the current customer utterance audio to predict the emotion label. In the Simple approach, we prompt the model with an annotation guide that was originally developed for human annotators to ensure consistency and alignment between machine-generated and human-generated labels. However, the previous context of the utterance is not provided. In ZCOT_Simple, we use the zero-shot chain-of-thought (COT) approach (Kojima et al., 2022) where the model first generates a reasoning before arriving at the result, thus improving the prediction performance of the model. In ZCOT_Context, along with the current customer ut-

User: You are now an AI agent tasked with analyzing audio from callers to detect their emotional state. Your analysis should categorize the caller’s emotion as either Non-negative or Negative.

Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-negative Emotion if the caller does not seem to have a negative emotion in that utterance.

Since we are doing a binary labeling here, positive emotions will also be marked as non-negative. Note that sometimes negative emotion can be more subtle and does not always involve negative language. It is crucial to actively monitor for indications of negative emotion throughout the call, including at the beginning of the interaction. Please provide a concise determination.

Assistant: Understood. I will analyze the audio, focusing on both linguistic and acoustic cues to determine the caller’s emotional state. I will categorize the emotion as either "Neutral" or "Negative," with "Negative" encompassing anger, irritation, dissatisfaction, and other negative expressions. I will pay close attention to:

****Linguistic Cues:**** Negative/angry words, swearing, complaints about the system or suggestions, repeated requests.

****Acoustic Cues:**** Changes in tone, higher pitch, louder volume, tension/strain in voice, sudden changes in loudness or speed.

I will classify any utterance without clear negative indicators, including those with positive emotions, as "Non-negative." I will be particularly vigilant for subtle signs of negative emotion throughout the entire interaction. My output will be a concise determination of "Non-negative" or "Negative."

Figure 1: Prompt used for RP_Detailed_Context prompting approach

terance, we provide a context window that contains previous agent-customer interactions in text format and this context is not available at inference

time, while the rest of the prompt remains the same as ZCOT_Simple. In ZCOT_Detailed_Context, the context window consists of utterance sequence numbers, and dialog state ids (DSIDs) along with the previous agent-user interactions.

In PAS_Detailed_Context, we use the Plan and Solve prompting strategy (Wang et al., 2023), which improves the zeroshot COT reasoning, by specifically prompting the model to first understand the problem, then plan a strategy to solve the problem and finally follow the strategy to solve the problem. The prompts of these approaches are shown in the Appendix A. In RP_Detailed_Context, we use the roleplay prompting strategy (Kong et al., 2024), wherein the model is prompted to act as an intent analyst following which the model will play the role of an intent analyst to recognize the emotion from the utterance. In both the prior and latter approaches, the context window is same as ZCOT_Detailed_Context approach. The prompt for RP_Detailed_Context is shown in Figure 1.

We empirically determined that annotation performance is optimal when the context window consists of a maximum of three previous agent-customer interactions in text format along with the current customer utterance audio. The results of the annotation performance of Gemini 2.0 Flash using Simple approach for different context window sizes are shared in Appendix B.

3.2 Fine-tuning SLMs

Running a foundational model for real-time SER is not feasible, as it will involve high cost and latency. Instead, we need to focus on small language models (SLMs) that are finetuned only for SER. To this end, we use the train dataset annotated using the foundational model to finetune a smaller audio-based language model, more specifically, Qwen 2 Audio 7B Instruct (Chu et al., 2024). We use LoRA (Hu et al.) to finetune the model. Since, our existing legacy system is built to process only the current user’s utterance in real time, so we use only the audio of the customer utterance as input to the model and do not add any other context to the input.

3.3 Reduction of Inference Time CPU Latency

SLMs having 7B paramters have low inference time latency on GPU. However, on CPU the inference time latency increases considerably. Although model compression methods such as bitsandbytes

(Detmers et al., 2021) are available, they are optimized to reduce the inference time latency in GPUs. Since our solution focuses on reducing CPU inference time latency, we need to use SLMs that are close to 1B-2B parameters. Thus, we use the knowledge distillation method using the fine-tuned Qwen 2 Audio 7B Instruct as the teacher model and Aero 1 Audio ¹, which is an audio-based SLM with 1.5B parameters as the student model. Since the vocabulary size of both the teacher and student models is different, we use universal logit distillation (ULD) (Boizard et al.) which calculates the Wasserstein distance to minimize the distance between the soft probabilities of the teacher and student models.

4 Datasets

Our goal in this study is to present evaluation results that accurately reflect the production environment and real-world data. To this end, we report SER accuracies, ROC-AUC, class-wise precision and recall on two internal datasets collected through distinct methodologies and provide an analysis of the observed differences. Additionally, we provide results on two widely used publicly available datasets for comparative analysis.

4.1 Internal Test Sets

A key challenge in SER is the strong class imbalance present in both training and real-world evaluation data, where emotionally neutral or low-arousal states dominate, and emotionally expressive states are relatively rare. This imbalance hinders model learning and reliable performance assessment, making it crucial to construct evaluation sets that fairly represent both classes, especially in production settings where detecting negative affect is a primary concern.

Throughout this paper, we refer to emotion classes as negative and non-negative, rather than the conventional neutral vs. emotional split. This reflects our focus on detecting negative states (e.g., anger, frustration), which are most relevant in customer support. In the customer care domain, other emotions such as happiness or surprise occur too infrequently to support consistent labeling and are grouped under the non-negative class.

ProdEval This is our most realistic evaluation set, constructed from real production conversations. As noted earlier, a key challenge in emotion data

¹<https://huggingface.co/lmms-lab/Aero-1-Audio>

selection is the natural class imbalance. In our application, analysis of randomly sampled calls shows that fewer than 1% of conversations exhibit clear negative emotions. To ensure meaningful evaluation, we applied class balancing during the selection process to include sufficient samples of both negative and non-negative classes.

In order to make expert annotation more efficient, we utilized a data selection strategy that increased the proportion of negative emotion calls without increasing the overall number of calls to be labeled. To achieve this, we applied a rule-based filtering approach informed by domain knowledge. Specifically, we constructed queries over production data to identify calls likely to contain negative emotions, focusing on conversational patterns associated with system misunderstandings, one of the most common sources of caller frustration in our application. To reduce potential bias from targeted sampling, we supplemented these with randomly selected calls, resulting in a distribution with a higher proportion of negative emotion compared to the natural distribution observed in production data.

The selected evaluation data was annotated by domain experts familiar with the specific characteristics and user behaviors of the target application. Because the system is designed to detect negative emotion at the level of individual user turns, annotation was performed at the utterance level, enabling fine-grained supervision aligned with the model’s inference granularity. To support accurate labeling, each utterance was presented along with the full audio of the call and all preceding turns, allowing annotators to consider both acoustic cues and conversational context when making judgments.

CueEval This evaluation set was also derived from production conversations, but it differs from ProdEval in both data selection and annotation methodology. As with ProdEval, random sampling was avoided due to the highly skewed distribution of emotional expressions in natural data. Instead, the selection strategy focused on leveraging existing production-side labels generated through a combination of human-in-the-loop processes and model-based inference that indicated instances of swearing or anger. These segments were combined with interactions labeled neutral or positive based on customer satisfaction cues, such as expressions of gratitude toward the system. The resulting subset was then re-annotated by expert raters in an offline setting to ensure consistent, high-quality emotion

labels.

Although CueEval reflects real production utterances, it does not capture the full realism of ProdEval. This limitation stems from its selection criteria, which primarily targeted linguistic cues such as swearing or overt expressions of anger—based on existing production labels. In contrast, ProdEval includes more subtle and context-dependent expressions of negative emotion, such as very short utterances (e.g., “yes” or “no”) that carry emotional charge but are challenging to detect automatically. These instances more accurately reflect the nuanced nature of real-world interactions. Furthermore, unlike ProdEval, where annotators had access to the full call audio and interaction history, CueEval annotations were performed at the utterance level without full conversational context, potentially limiting labeling accuracy for ambiguous cases.

4.2 Public Test Sets

We also evaluate our prediction models on two publicly available SER datasets: MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008). We only perform evaluation on the test split and do not use the train split to train our models specifically to perform prediction on these datasets. MELD dataset consists of utterances from the Friends TV series, it consists of the following labels: anger, disgust, sadness, joy, neutral, surprise and fear. IEMOCAP dataset consists of dyadic sessions where the emotional utterances of an actor are based on hypothetical scenarios. It consists of the following labels: happiness, anger, sadness, frustration and neutral. Similar to Caralt et al. (2025), in MELD, we consider the emotion labels: anger and disgust as negative emotion and thus relabel these emotions as negative emotion while the rest of the emotion labels are relabeled as non-negative emotion. So also, in IEMOCAP, we consider the emotion labels: anger, disgust and frustration as negative emotion while the rest of the emotion labels are relabeled as non-negative emotion

4.3 Train Dataset

The train dataset is created using the same strategy that we used to create ProdEval. We use the role-play prompting strategy to annotate the utterances in the train dataset. After performing the annotation, we further reduced the class imbalance of negative and non-negative emotions by using 1:K negative sampling.

Table 2 shows the number of samples for negative and non-negative classes across all the datasets.

Datasets	Non-Negative	Negative
ProdEval	1030	566
CueEval	114	92
MELD	1256	413
IEMOCAP	551	384
Train Dataset	25593	6447

Table 2: Statistics of negative and non-negative labels in the test and train datasets

5 Results and Discussion

5.1 Experiment Settings

We use Gemini 2.0 Flash as our annotator, leveraging its multi-modal capabilities to enhance annotation quality. For model comparison, we evaluate our fine-tuned Qwen 2 Audio 7B Instruct (Qwen 2 Audio) and distilled Aero 1 Audio (Aero) models against both text- and audio-based baselines. In the text-based setting, we use BERT base uncased (BERT) (Devlin et al., 2019), Llama 3.1 8B Instruct (Llama 3.1) (Dubey et al., 2024), Qwen 2 7B Instruct (Qwen 2 Text) (Team et al., 2024), and Qwen 3 4B Instruct (Qwen 3 Text) (Yang et al., 2025) as our baseline models. These models are trained on our dataset and evaluated via ASR-transcribed utterances. The text-based SLMs: Llama 3.1, Qwen 2 Text and Qwen 3 Text are finetuned using LoRA.

Prompt for text-based SLMs
You are a helpful assistant.
User: Detect the emotion from the text. Classify as Negative or Non-Negative emotion
Text: <Customer utterance text>
Assistant: <Emotion>

Table 3: Prompt template used for Llama 3.1, Qwen 2 Text, Qwen 3 Text models

Prompt for audio-based SLMs
You are a helpful assistant.
User: Detect the emotion from the audio. Classify as Negative or Non-Negative emotion
Audio: example_audio.wav
Assistant: <Emotion>

Table 4: Prompt template used for Aero, Qwen 2 Audio, Kimi Audio, Gemini models

Among audio models, we include SER-Odyssey-Baseline-WavLM-Arousal (WavLM-Arousal) (Goncalves et al., 2024), classifying utterances with arousal scores >0.5 as negative. We also test Kimi Audio 7B Instruct (Kimi Audio) (Team, 2024)—a Qwen 2.5-based model via zero-shot inference due to LoRA incompatibility. Finally, we include Gemini 2.0 Flash (Gemini) as a zero-shot baseline using roleplay prompting without dialogue context. All audio models are evaluated using only the current customer utterance. Prompts used for text-based SLMs are detailed in Table 3. Prompts used for Gemini and audio-based SLMs are detailed in Table 4. Given the class imbalance in SER, we report precision and recall per class, along with ROC-AUC and accuracy for overall comparison. Recall is particularly important in our use case, as the goal is to detect negative customer emotions in customer care calls. Failing to identify angry or frustrated customers can result in unresolved issues or escalations, making high recall in negative emotion classes a critical priority.

Approach Name	ROC-AUC	Accuracy
Simple	65.16	72.27
ZCOT_Simple	65.61	73.52
ZCOT_Context	69.57	72.33
ZCOT_Detailed_Context	69.99	73.27
PAS_Detailed_Context	67.24	72.96
RP_Detailed_Context	69.18	74.34
Negative Emotion		
Approach Name	Precision	Recall
Simple	68	41
ZCOT_Simple	75	38
ZCOT_Context	61	60
ZCOT_Detailed_Context	63	59
PAS_Detailed_Context	67	48
RP_Detailed_Context	68	51
Non-Negative Emotion		
Approach Name	Precision	Recall
Simple	73	90
ZCOT_Simple	73	93
ZCOT_Context	78	79
ZCOT_Detailed_Context	78	81
PAS_Detailed_Context	75	87
RP_Detailed_Context	76	87

Table 5: The performance of Gemini 2.0 Flash as an annotator when using different approaches to prompt the LLM.

5.2 Experimental Results

5.2.1 Performance of Gemini 2.0 Flash as annotator

The performance of Gemini 2.0 Flash as an annotator on ProdEval data is shown in Table 5. Note that when evaluating an LLM as an annotator, unlike during inference, where recall is often more important than precision, the goal is to achieve high performance on both precision and recall. This is because low precision will lead to mislabeled training samples. We observe that RP_Detailed_Context approach gives the highest accuracy and the ROC-AUC score is close to ZCOT_Detailed_Context approach. The recall of RP_Detailed_Context approach on negative emotion is also slightly lower than ZCOT_Detailed_Context approach. The precision is better than ZCOT_Detailed_Context approach and only lower than ZCOT_Simple approach. However, the ZCOT_Simple approach has the lowest recall on negative emotion. The performance of all the approaches on non-negative emotion is comparable. ZCOT_Simple, ZCOT_Context, and ZCOT_Detailed_Context use the zershot COT prompting, wherein the model first generates the reasoning based on the prompt, after which it uses the prompt and reasoning to generate the emotion, effectively being a two-pass approach (Kojima et al., 2022) and in the process also increasing the cost of output generation when used in scale. In contrast, RP_Detailed_Context uses the roleplay approach where assigning the role of intent analyst leads to implicit COT and also provides better reasoning than zero-shot COT (Kong et al., 2024), thus effectively being a single-pass approach. Thus, in case of data annotation at scale, RP_Detailed_Context approach is suitable as its ability to detect negative emotion is comparable to that of ZCOT_Detailed_Context while also giving highest overall accuracy and lowering the cost of output generation. Furthermore, in ProdEval the Cohen’s Kappa score between predictions using RP_Detailed_Context and human annotator is 0.4 indicating fair agreement between Gemini as an annotator and human annotator. Thus, showing the effectiveness of our prompting method for large scale annotation. The prediction performance of the text- and audio-based models are shown in Table 6 which we will discuss in detail in the following sections.

5.2.2 Performance of text-based models

We observe that among the text-based models Llama 3.1 has the highest avg. Accuracy and ROC-AUC, whereas in terms of recall of negative emotion, Qwen 2 Text has the highest avg. recall. Qwen 3 Text has avg. accuracy, avg. ROC-AUC and avg. recall of negative emotion, which is closer to Llama 3.1. However, interestingly Qwen 3 Text has only 4B parameters compared to Llama 3.1 which has 8B parameters. Since, Qwen 3 models are trained on a combination of COT generated reasoning data and curated datasets, so the implicit thinking ability is ingrained in the model that may contribute to improved performance even when the model has less parameters. Finally, we observe that BERT is the worst model in terms of recall of negative emotion. However, in terms of avg. precision, the performance of BERT is comparable to that of other text-based models. In the case of non-negative emotion, the avg. precision and recall of all the text-based models is comparable, which is expected since non-negative emotion is the majority class and is easier to detect compared to negative emotion labels.

5.2.3 Performance of audio-based models

We first discuss the results of our finetuned Qwen 2 Audio model and other audio models that we use as baselines. We observe that our finetuned Qwen 2 Audio model, surpasses the text-based models by considerable margins in terms of avg. accuracy, avg. ROC-AUC and avg. recall on negative emotion. WavLM Arousal model gives the best performance in ProdEval but does not perform well on CueEval as the dataset is based on linguistic cues, while the model is more inclined towards detecting acoustic cues. Kimi Audio performs better than Qwen 2 Audio even at zero-shot because Kimi Audio improves upon the Qwen 2.5 7B architecture by using Whisper and GLM tokenizer to tokenize the audio which may play a role in improving the performance. We also observe that Gemini gives performance that is comparable to Qwen 2 Audio in ProdEval, while it gives a higher performance in CueEval. Thus, showing that when using acoustic and linguistic cues, Gemini will require more context to give better results, but it will perform well when the dataset has only linguistic cues.

In MELD, the performance of Qwen 2 Audio degrades in terms of accuracy, ROC-AUC and precision. This could be because MELD is based on TV series and the emotions tend to be exaggerated and dramatic which may have negatively affected

Models	ROC-AUC				Avg. ROC-AUC	Accuracy				Avg. Accuracy
	ProdEval	CueEval	MELD	IEMOCAP		ProdEval	CueEval	MELD	IEMOCAP	
Text Models										
BERT	61.87	68.44	55.08	60.64	61.51	69.57	68.76	68.48	57.75	66.14
Llama 3.1	62.8	69.7	69.6	66.19	67.07	69.63	68.44	61.47	69.09	67.16
Qwen 2 Text	61.91	64.79	67.9	68.02	65.66	68.69	63.59	58.65	64.15	63.77
Qwen 3 Text	62.03	68.09	69.66	65.02	66.20	69.57	67.47	63.39	67.37	66.95
Audio Models										
WavLM-Arousal	70.47	53	58.7	77.51	64.92	75.25	54.36	41.4	72.19	60.80
Kimi Audio	71.29	83.78	73.52	79.52	77.03	76.28	84.95	81.3	76.14	79.67
Gemini	64.84	88.19	71.3	74.79	74.78	72.02	87.86	73.81	73.68	76.84
Qwen 2 Audio	66.42	81.49	65.92	69.95	70.95	72.2	80.09	57.27	72.4	70.49
Aero	67.83	77.68	62.12	69.54	69.29	73.21	76.69	47.51	72.62	67.51
Precision and Recall for Negative Emotion										
Models	Precision				Avg. Precision	Recall				Avg. Recall
	ProdEval	CueEval	MELD	IEMOCAP		ProdEval	CueEval	MELD	IEMOCAP	
Text Models										
BERT	63	63	34	73	58.25	35	72	29	44	45
Llama 3.1	61	80	38	70	62.25	39	58	86	82	66.25
Qwen 2 Text	59	57	36	68	55	39	76	86	86	71.75
Qwen 3 Text	62	61	39	70	58	36	74	82	78	67.5
Audio Models										
WavLM-Arousal	72	49	29	38	47	54	40	93	86	68.25
Kimi Audio	69	92	63	64	72	54	73	58	98	70.75
Gemini	68	83	48	84	70.75	40	91	66	69	66.5
Qwen 2 Audio	65	71	35	73	61	46	95	83	84	77
Aero	67	69	31	72	59.75	49	87	91	87	78.50
Precision and Recall for Non-Negative Emotion										
Models	Precision				Avg. Precision	Recall				Avg. Recall
	ProdEval	CueEval	MELD	IEMOCAP		ProdEval	CueEval	MELD	IEMOCAP	
Text Models										
BERT	71	74	78	49	68	88	66	82	77	78.25
Llama 3.1	72	61	92	66	72.75	86	82	54	50	68
Qwen 2 Text	72	73	92	67	76	85	54	50	43	58
Qwen 3 Text	71	75	91	62	74.75	88	62	57	52	64.75
Audio Models										
WavLM-Arousal	78	58	91	96	80.75	89	66	24	69	62
Kimi Audio	77	81	87	98	85.75	87	95	89	61	83
Gemini	73	92	87	64	79	90	85	76	81	83
Qwen 2 Audio	75	94	90	71	82.5	86	68	49	56	64.75
Aero	76	87	92	73	82.00	86	68	33	52	59.75

Table 6: The prediction performance of the audio- and text-based baseline models and our finetuned audio-based SLM and distilled audio-based SLM on different test sets.

the model. In case of IEMOCAP, the accuracy and ROC-AUC of Qwen 2 Audio is comparable to Gemini, while the recall on negative emotion outperforms Gemini. Since, IEMOCAP comprises of conversation between multiple speakers and consists of genuine emotions, it is closer to our ProdEval and CueEval datasets. Thus, the results on ProdEval, CueEval and IEMOCAP datasets show that SLMs that are fine-tuned for domain-specific tasks can outperform generalized larger LLMs. Similar, to text-based models, in the case of non-negative emotion, the avg. precision and recall of all the audio-based models is comparable. Overall, we can conclude that the audio-based models outperform the text-based models as they factor in the linguistic and acoustic cues present in the utterance to predict the emotion.

5.2.4 Performance of distilled audio-based model

The Aero model is distilled from our finetuned Qwen 2 Audio model. The results show that the distilled Aero model has the highest recall on negative emotion among all the text- and audio based models. Since Aero is distilled from our finetuned Qwen 2 Audio, it follows the similar trends as Qwen 2 Audio across all the datasets. The avg. accuracy of Aero is slightly lower than that of Qwen 2 Audio. At dataset level, the recall of negative emotion, is consistently better than Qwen 2 Audio and also exceeds or remains on par with Kimi Audio and Gemini. Thus, we show that distillation from teacher model to student model can improve classification performance even if both models belong to different language model families. Overall, in production, we need a high-recall model for detecting negative emotion and Aero fulfills that requirement.

Models	CPU Latency (in seconds)
BERT	0.04
Qwen 2 Text	5.56
WavLM Arousal	0.66
Qwen 2 Audio	7.36
Aero	2.56

Table 7: Inference time CPU latency for text-based models and audio-based models

5.3 Latency Analysis

As we need our model to have a high-recall as well as low latency during inference time, so we analyze the inference time CPU latency of Qwen 2 Audio and Aero. We compare these models with Qwen 2 Text, which is the best performing text-based model in terms of recall of negative emotion, BERT and WavLM-Arousal, which are light-weight models and have only a few million parameters. The average latency of the models is shown in Table 7. We observe that avg. latency of BERT and WavLM-Arousal is less than 1 second because both models have low parameters. However, the classification performance is much lower than the other SLMs discussed earlier. Qwen 2 Text and Qwen 2 Audio have avg. latency of 5.56s and 7.36s, respectively, which make them unsuitable for real-time deployment. However, Aero has an avg. latency of 2.56 secs while maintaining a classification performance that is almost on par with Qwen 2 Audio and outperforms other models in terms of recall of negative emotion. Thus, it is suitable for real-time deployment while maintaining quality performance.

6 Conclusions

We developed low-latency, high-recall audio-based SLM for real-time detection of negative emotions in customer care, with consistent performance across internal and public test sets. Using LLMs for large-scale annotation, we created realistic training data with quality close to human labels. Audio-based models outperformed text-only approaches, effectively capturing emotion from short utterances without context. Notably, smaller fine-tuned models matched or exceeded the performance of larger LLMs. For future work, we aim to further improve latency and recall, and to monitor system performance post-deployment in real-world settings.

Limitations

Despite the comprehensive evaluation, several limitations should be noted. First, while we focus on internal datasets that closely represent real-world data, their scope may still be limited in diversity, which could affect the generalizability of our results to all production environments. Additionally, our approach primarily targets detection of negative emotion and its performance in recognizing other emotional states remains an area for further exploration. Furthermore, our work does not cover the detection and mitigation of data drift that can occur in production environment due to shift in distribution of production data compared to training data. Finally, while we aim for real-time applicability, factors such as computational resources and model scalability in dynamic environments were not fully addressed in this study.

References

- Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a generative approach for emotion detection and reasoning. *arXiv preprint arXiv:2408.04906*.
- Nicolas Boizard, Kevin El Haddad, CELINE HUDELOT, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *Transactions on Machine Learning Research*.
- Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, pages 1–8, Cambridge, MA, USA. IEEE.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- Mireia Hernandez Caralt, Ivan Sekulić, Filip Carevic, Nghia Khau, Diana Nicoleta Popa, Bruna Guedes, Victor Guimarães, Zeyu Yang, Andre Manso, Meghana Reddy, and 1 others. 2025. “stupid robot, i want to speak to a human!” user frustration detection in task-oriented dialog systems. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 276–285.
- Youjun Chen, Xurong Xie, Haoning Xu, Mengzhe Geng, Guinan Li, Chengxi Deng, Huimeng Wang, Shujie Hu, and Xunying Liu. 2025. Towards llm-empowered fine-grained speech descriptors for explainable emotion recognition. *arXiv preprint arXiv:2505.23236*.

- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.
- 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.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Soumya Dutta and Sriram Ganapathy. 2025. Llm supervised pre-training for multimodal emotion recognition in conversations. *arXiv preprint arXiv:2501.11468*.
- Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.
- Apoorva Garg, Sagnik Ghosh, Poonam Ramu, Naren Vempaty, and Harsh Shah. 2024. Deep learning approaches for speech emotion recognition: Challenges and opportunities. *ICT Express*.
- L. Goncalves, A. N. Salman, A. Reddy Naini, L. Morovelazquez, T. Thebaud, L. Paola Garcia, N. Dehak, B. Sisman, and C. Busso. 2024. Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results. In *Odyssey 2024: The Speaker and Language Recognition Workshop*, volume To appear, Quebec, Canada.
- Taesik Gong, Josh Belanich, Krishna Somandepalli, Arsha Nagrani, Brian Eoff, and Brendan Jou. 2023. Lanser: Language-model supported speech emotion recognition. In *Proceedings of Interspeech 2023*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *NAACL-HLT*.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.
- Hsien-Chin Lin, Shutong Feng, Christian Geischauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić. 2023. Emous: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2526–2531.
- Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung-yi Lee. 2024. Emo-bias: A large scale evaluation of social bias on speech emotion recognition. *arXiv preprint arXiv:2406.05065*.
- Fei Ma, Yucheng Yuan, Yifan Xie, Hongwei Ren, Ivan Liu, Ying He, Fuji Ren, Fei Richard Yu, and Shiguang Ni. 2025. Generative technology for human emotion recognition: A scoping review. *Information Fusion*, 115:102753.
- Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*.
- Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W. Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330, Seoul, South Korea. IEEE.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.
- Yuntao Shou, Wei Ai, Jiayi Du, Tao Meng, Haiyan Liu, and Nan Yin. 2024. Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*.
- Armand Stricker and Patrick Paroubek. 2024. A unified approach to emotion detection and task-oriented dialogue modeling. *arXiv preprint arXiv:2401.13789*.
- Kimi Team. 2024. [Kimi-audio technical report. Preprint](#), arXiv:arXiv:placeholder.
- Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3.

Shreya G Upadhyay, Woan-Shiuan Chien, and Chi-Chun Lee. 2025. Is it still fair? investigating gender fairness in cross-corpus speech emotion recognition. *arXiv preprint arXiv:2501.00995*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634.

Yuezhou Wu, Siling Zhang, and Pengfei Li. 2025. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*, 15(1):5584.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, pages 19323–19331.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W. Schuller. 2024. Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems*, 11(5):6690–6704.

A Prompts used for using Gemini 2.0 Flash for annotation

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:

Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.

Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.
2. System misunderstands the caller and repeats the same options or reprompts multiple times
3. Caller does not understand or does not have the information that virtual assistant is asking
4. Long, unresolved conversations where the user's task remains incomplete.

Return only "Negative Emotion" or "Non-Negative Emotion" based on the audio. Do not add any additional comments.

Figure 2: Prompt used for Simple prompting approach

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:
Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.
 Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.
2. System misunderstands the caller and repeats the same options or reprompts multiple times
3. Caller does not understand or does not have the information that virtual assistant is asking
4. Long, unresolved conversations where the user's task remains incomplete.

Let's think step by step to understand the emotion of the utterance of the user. Provide some explanation.

Figure 3: Prompt used for ZCOT_Simple, ZCOT_Context, and ZCOT_Detailed_Context prompting approaches

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:
Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.
 Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.
2. System misunderstands the caller and repeats the same options or reprompts multiple times
3. Caller does not understand or does not have the information that virtual assistant is asking
4. Long, unresolved conversations where the user's task remains incomplete.

The sequence number of the utterances in the conversation is given in brackets as Utterance sequence no.
 The dialog state id indicates the dialog state of the user utterance is provided for each user utterance in brackets as Dialog state id.
 Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step.
 Pay attention to the context of the conversation (if available), use the text of the context, utterance sequence numbers and dialog state ids to understand the current utterance. Use this to refine your plan. Pay attention to commonsense and logical reasoning.

Figure 4: Prompt used for PAS_Detailed_Context prompting approach

B Impact of Context Window Size on Annotation Performance

Context Window Size	ROC-AUC	Accuracy
1	64.31	69.38
2	61.77	70.51
3	64.48	70.82
4	64.32	70.01
5	63.64	69.69
Negative Emotion		
Context Window Size	Precision	Recall
1	59	47
2	68	32
3	63	43
4	61	45
5	60	43
Non-Negative Emotion		
Context Window Size	Precision	Recall
1	74	82
2	71	92
3	73	86
4	73	84
5	73	85

Table 8: Classification performance of Simple prompting approach for different context window sizes. The context window size indicates the number of previous agent-customer interactions that are provided as context.

Says Who? Argument Convincingness and Reader Stance Are Correlated with Perceived Author Personality

Sabine Weber, Lynn Greschner, and Roman Klinger

Fundamentals of Natural Language Processing, University of Bamberg, Germany
{sabine.weber, lynn.greschner, roman.klinger}@uni-bamberg.de

Abstract

Alongside its literal meaning, text also carries implicit social signals: information that is used by the reader to assign the author of the text a specific identity or make assumptions about the author's character. The reader creates a mental image of the author which influences the interpretation of the presented information. This is especially relevant for argumentative text, where the credibility of the information might depend on who provides it. We therefore focus on the question: How do readers of an argument imagine its author? Using the ContArgA corpus, we study arguments annotated for convincingness and perceived author properties (level of education and Big Five personality traits). We find that annotators perceive an author to be similar to themselves when they agree with the stance of the argument. We also find that the envisioned personality traits and education level of the author are statistically significantly correlated with the argument's convincingness. We conduct experiments with four generative LLMs and a RoBERTa-based regression model showing that LLMs do not replicate the annotators judgments. Argument convincingness can however provide a useful signal for modeling perceived author personality when it is explicitly used during training.

1 Introduction

When interpreting a text, social clues about the author (also called *social meaning*, Nguyen et al., 2021) and referential meaning are often intrinsically linked. Reading a social media post by a disliked politician might lead to a different interpretation than reading the same text but assuming it was written by a friend. Similarly, properties of the text like word choice might give clues to an author's educational level or personality traits. This interplay is especially relevant with regards to persuasive text, where a judgment of the argument's convincingness is derived from both the argument

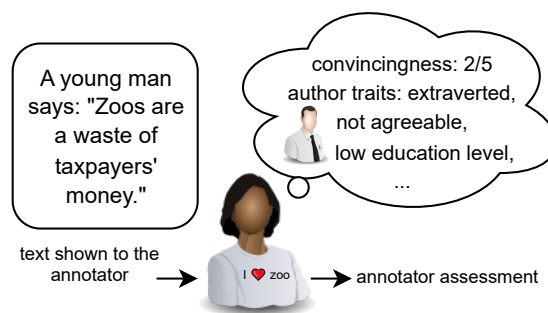


Figure 1: Work flow example. The annotator's stance (pro zoos, as shown on the t-shirt) clashes with the stance of the argument. The argument is seen as unconvincing and leads to an unfavorable assessment of the author.

itself and the source of the argument (Petty and Cacioppo, 1986). It is our hypothesis that, when only minimal contextual grounding is available, readers develop their mental image of the author mostly from the argumentative text, allowing us a glimpse into the construction of social meaning.

Previous work has found the assessment of argument convincingness to be a subjective task with low inter-annotator agreement (Quensel et al., 2025). While convincingness is derived from the argument text itself, it is also influenced by factors that are dependent on the reader of the argument: their familiarity with the topic and their stance towards it (Greschner et al., 2025). In this work, we examine both the argument-focused variable of convincingness as well as the annotator-focused variable of topic stance with regards to their role in the creation of a mental image of an argument's author. This way we move from an abstract view of arguments towards a perspectivist approach that integrates human differences.

To examine the phenomenon of author perception we investigate the connections between the convincingness of an argument and traits of its en-

visioned author (Big Five personality traits (Costa and McCrae, 1999) and education level). We find that convincingness is positively correlated with perceived education and the personality traits Agreeableness and Conscientiousness. We also investigate how properties of the annotator shape the envisioned author, specifically the annotators' agreement with the stance of the argument. We find that when the stance of the argument aligns with the stance of the annotator, they perceive the envisioned author to be more similar to themselves than when stances do not align.

We study the role of convincingness in modeling author personality by testing four generative LLMs, either with no convincingness signal, the annotator-assigned score, or a random value. Results show that the tested LLMs do not benefit from the convincingness signal. In contrast, a RoBERTa-based regression model trained with the convincingness signal better aligns with human annotations than the same model without it. This highlights convincingness as a valuable cue for modeling perceived author personality.

Our main research questions are:

RQ 1: Do annotators build an internal representation of an author when presented with an argument? (*Yes*)

RQ 2: What perceived author properties are associated with the individual assessment of convincingness in arguments? (*All personality traits show correlation with convincingness, with Agreeableness having the strongest correlation*)

RQ 3: How does annotator stance influence the similarity between annotator and envisioned author? (*When stances agree, there is a statistically significant correlation of all personality traits of annotator and envisioned author*)

RQ 4: Does the correlation in of convincingness and envisioned author personality traits (established in RQ2) carry over to computational modeling, helping to predict perceived author personality? (*Yes*)

Understanding how readers envision an argument's author is important because social inferences play a role in credibility judgments, yet they remain underexplored in computational argumentation. By examining how readers envision authors within a controlled setting, our work offers an empirical basis for understanding these social inferences. Our findings highlight that subjective factors like assessment of convincingness and reader stance are a non-negligible part of how people pro-

cess argumentative text. We make all data and code publicly available.¹

2 Related Work

2.1 Social Meaning

Understanding and modeling the information that a text conveys about its author has been the focus of computational sociolinguistic research, looking especially at vernacular and dialect (Nguyen et al., 2016). Recent work has pointed out that the dimension of social meaning remains underexplored in the context of NLP, especially because modern NLP systems train on large data sets, where text is removed from the situational context of its creation, capturing the abstract patterns of language rather than its situation dependent use (Yang et al., 2025).

While social meaning might be embedded along representational meaning in language models, these models do not actively draw on this knowledge (Lauscher et al., 2022). Nguyen et al. (2021) argue that linguistic forms with different social meaning should not receive the same representation if social meaning is relevant for the task at hand. This would require disentangling the two types of meaning with regards to social context and properties of the interaction participants.

Recent work in perspectivism shares in this criticism of socially unaware models (Frenda et al., 2024). Perspectivist authors point out that variations between annotators should not be seen as noise or a product of insufficient annotator training, but as a source of information about the task at hand (Kanclerz et al., 2022; Casola et al., 2025; Weber-Genzel et al., 2024).

Our work follows this advice by considering not only argument text, but also individual properties of the annotator like their personality traits, education level, stance towards the argument topic, assessment of argument convincingness and the mental image that annotators create of the author.

2.2 Personality Traits

The task of deducing an authors personality traits from text is examined in the domain of author profiling (Verhoeven et al., 2016; Kreuter et al., 2022). In contrast to this work, we do not aim to deduce the ground truth personality traits of the author of an argument, but rather what a reader of the argument thinks they are – assessing an imagined au-

¹<https://www.uni-bamberg.de/en/nlproc/projects/emcona>

thor, that might be different for each reader, rather than an actually existing person.

Another line of work examines the way readers construct mental representations of fictional characters from text. Pizzolli and Strapparava (2019) use the Big Five personality trait model to create character profiles from dialogues in theater plays and Tiuleneva et al. (2024) release a data set of character utterances annotated with Big Five personality traits. While these works are more similar to our in that they are also concerned with the mental representation of personality from text alone, they do not integrate subjective assessments of the annotators into modeling, thus ignoring differences between readers.

2.3 Perceived Author Identity in Arguments

While there is some research on perceived author properties in other domains, there is only little research about implicit information conveyed about the author in textual arguments, one instance being Bender et al. (2011). The paper studies claims to authority and agreement in Wikipedia forum discussions. While Bender et al. (2011) do not study the assumptions that conversation participants make about one another, they examine how speaker identity and authority is constructed in text, calling this *identity work* in reference to sociolinguistic research (Bucholtz and Hall, 2010).

Another notable work is the ContArgA corpus (Greschner et al., 2025), which allows us to examine this identity work at play in one concentrated snapshot. Unlike a lengthy forum discussion that allows for opinions to develop slowly, the corpus offers an opportunity to see the mental model that annotators develop of the author, based only on a text, minimal demographic information and their own prior belief.

3 Experimental Settings

To answer our research questions, we take a two-step approach. We first conduct a detailed statistical analysis of the ContArgA corpus to answer RQs 1, 2, and 3 and then use the gained insights to design our modeling experiments, thereby answering RQ 4.

3.1 The ContArgA Corpus

To conduct our research, we require a corpus that combines short textual arguments with annotator assessments of the text’s convincingness and properties of the envisioned author of the text, e.g., the

author’s education level and Big Five personality traits. We also need the corpus to contain the same information (education level and Big Five Personality traits) about the annotator. The ContArgA corpus (Greschner et al., 2025) satisfies these requirements which is why we select it for our study.

The ContArgA corpus contains 800 arguments that were sampled from two existing argument corpora and re-annotated for a variety of different variables. Each argument was annotated by 5 annotators resulting in a total of 4000 annotations. Each annotator annotated at least 2 arguments, but multiple participation was allowed. Annotators were recruited via the annotation platform Prolific and they represent an even distribution across ages and genders. For further details, refer to Greschner et al. (2025).

During the annotation process, annotators were instructed to imagine themselves as participants in a town hall discussion on a contentious issue, watching a speaker approaching the podium and presenting an argument in favor or against the issue (an example of this can be seen in Figure 1). They then provided judgments of the convincingness of the argument, emotions they experienced and an assessment of the person saying the argument. While the corpus contains a variety of annotations, we are specifically interested in variables pertaining to the annotator and the envisioned author of the argument.

Input. After being introduced to the scenario annotators see the textual argument along a minimal description of the person uttering the argument. The description presents the annotator either with an old or a young person and a man or a woman, e.g.: “An old woman approaches the microphone and makes a statement: ...”. The input variables are age and gender of the author and the argument they present. Textual arguments that explicitly refer to the speakers age or gender were removed during corpus creation to avoid conflicting inputs.

Annotations. The annotators provide three kinds of information: First, they provide information about themselves, specifically by disclosing their level of education and their stance towards the discussion topic and by filling a Big Five personality test (Gosling et al., 2003). Second, they provide their assessment of the argument, specifically by annotating its convincingness on a scale of 1 to 5, with 1 being least and 5 being most convincing. Third, they provide information about the en-

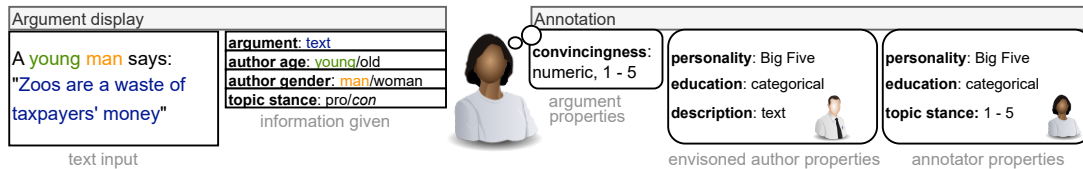


Figure 2: Overview of relevant variables collected in the ContArgA corpus.

visioned author, assigning them an education level, filling the same Big Five personality test they filled earlier for themselves for the author and lastly filling an optional free text field with further details about the envisioned author. An overview of the examined variables can be found in Figure 2.

3.2 Modeling

To examine the role of convincingness in modeling perceived author personality traits, we use the full ContArgA corpus as a test set. All LLMs receive the ContArgA corpus as input, while for the RoBERTa-based model we perform 5-fold cross-validation. This way we can report results for the RoBERTa model for all instances of the ContArgA corpus. In all test cases we compute the root mean squared error in comparison with the gold labels provided by the annotators for each of the Big Five personality traits. We release all models, training code and LLM outputs.²

LLM. To examine the capabilities of light-weight LLMs in modeling implicit assumptions about the author of a text we choose four recent open-weight LLMs (Gemma 8.5B (Mesnard et al., 2024), Mistral 7.2B (Jiang et al., 2023), Mixtral 46.7B (Jiang et al., 2024), 3.2B (Grattafiori et al., 2024)). We use a zero-shot prompting setup in which we put the full text that the annotators see as the user prompt. As system prompt we set the task of filling the Big Five personality test for the author of the argument. We also provide the not yet filled-out questionnaire itself, the same way as the human annotators were shown during the creation of the ContArgA corpus (the full prompt can be found in Appendix A). We set three experimental conditions: One providing the LLM with the convincingness value annotated by the human annotators (as an additional textual line in the prompt), one providing a random value, and one without this information. Building on the results of our earlier statistical analysis, we hypothesize that having the annotator-assigned convincingness values (in com-

parison to a random value or no value at all) should help the LLM to make author assessments that are more similar to human judgments.

All models are hosted on our own infrastructure via Ollama³. We use Ollama default parameters for all models, namely a temperature of 0.7, token limit of 1024 and Top P sampling set to 0.95.

RoBERTa-based Model. In addition to LLM prompting, we compare against a RoBERTa-based (Liu et al., 2019) regression model which is inspired by the FiLM model used for learning from input signals with different sized vector encodings (Perez et al., 2018). Each argument text is combined with the demographic data of the speaker (as in the example in Figure 2), tokenized and encoded by RoBERTa⁴, and the representation of the [CLS] token is extracted as a fixed-length vector. To account for the role of perceived convincingness, we implement two model variants:

In the Baseline Model, the [CLS] embedding is passed through a two-layer feed-forward regression head (512 \rightarrow 256 hidden units, ReLU activations, dropout rate 0.2), producing continuous predictions for the Big Five personality dimensions. In the Convincingness-Augmented Model, in addition to the argument text, the annotator-provided convincingness score is supplied as an auxiliary scalar input. The score is projected into the same dimensionality as the [CLS] embedding via a two-layer projection network, and used to generate feature-wise scaling and shifting parameters (FiLM modulation). These parameters are applied to the [CLS] embedding to yield a convincingness-conditioned representation, which is then passed through the same feed-forward regression head as the baseline. This architecture ensures that the augmented model has equivalent capacity to the baseline, with the only difference being the inclusion of convincingness information.

RoBERTa-based models were trained with early stopping, learning rate of 2×10^{-5} and a weight

²<https://www.uni-bamberg.de/en/nlproc/projects/emcona>

³<https://github.com/ollama/ollama>

⁴<https://huggingface.co/FacebookAI/RoBERTa-base>

decay of 0.01. The best performing model was trained for 9 epochs. All models were trained on a single Nvidia L40 GPU with one training run taking on average 1.5 minutes.

4 Results

In this following section we will outline the results of the analysis of the ContArgA data set and the modeling experiments.

4.1 RQ 1: Do annotators imagine an author when presented with arguments?

To answer this question we look at several statistical properties of the data presented in the ContArgA corpus.

Free Text Input. We investigate whether annotators use the free-text description field offered to give additional details about the author to attribute fully developed characteristics to them, and whether these descriptions differ depending on the argument or between annotators.

To address these questions, we analyze the free-text responses provided by annotators. We first examine the frequency and content of the entries, and then quantify variation using the Jaccard index, a measure of word overlap. Variation is assessed both within annotators (comparing their descriptions across different arguments) and across annotators (comparing descriptions for the same argument).

Overall, 69% of annotators completed the free-text field. Among the 100 most frequently used adjectives are “passionate” (22 mentions), “strong” (20 mentions), and “religious” (14 mentions), while common nouns include identity terms such as “father” (18 mentions), “student” (15 mentions), and “parent” (7 mentions). These patterns suggest that annotators ascribe specific roles and characteristics to the imagined authors.

To determine whether different arguments elicit different author descriptions, we calculated the Jaccard index for all entries from the same annotator. The resulting low average of 0.074 indicates that individual annotators provide distinct descriptions for different arguments. Exceptions exist, such as an annotator who consistently questioned whether each envisioned author was “of foreign descent.” Some annotators also maintain a repeated sentence structure (e.g., “She seemed . . .” or “The person is . . .”) while varying the descriptors.

We then assessed whether different annotators envision the same author differently by computing the Jaccard index across annotators for the same argument. The very low average of 0.035 confirms substantial variation between annotators. Although occasional overlap occurs (e.g., two annotators describing an author as “right-wing”), more often, annotators use divergent descriptors, such as one calling the author “well-educated” while another describes them as “naive”.

Demographic Variables of the Author. We hypothesize that if an annotator envisages an author, they might apply the same demographic biases to them as they would to an existing person. Specifically, we want to examine whether perceptions of an author’s age or gender influence how annotators envisage them, e.g. whether older authors are assumed to be wiser than younger authors, or whether female authors are envisioned as more emotional than male authors. In the ConArgA annotation process demographic information about the author was presented alongside the argument text (see Figure 1), which allows us to answer these questions.

To answer them, we assign numerical values to the successive categorical education levels and perform a T-Test between the respective groups. Because the personality traits are valued with numeric values, e.g., Extraversion = 2, we perform a T-Test between respective groups here, too. We do not find any statistically significant differences in envisioned education level between female and male authors, and only slight differences in assumed personality traits: Women are rated slightly (but statistically significantly) higher in Agreeableness and lower in Emotional Stability, which aligns with gender stereotypes in the region where the corpus was collected (Plant et al., 2000).

Demographic bias is more pronounced along the age axis than along the gender axis: Old authors are assumed to have a lower education level than young ones and are perceived as statistically significantly different in all personality traits, being assumed to be less extraverted and open and more agreeable, conscientious and emotionally stable than young authors. These findings point towards the annotators forming a complex mental image of the author when confronted with the arguments, rather than envisioning an average or random person.

Trait	
Extraversion	-0.08***
Agreeableness	0.39***
Conscientiousness	0.22***
Emotional Stability	0.18***
Openness	0.06**

Table 1: Pearson correlation (r) between assumed author traits and argument convincingness. Significance levels ** $p < .01$, *** $p < .001$

4.2 RQ 2: What perceived author properties are associated with the individual assessment of convincingness in arguments?

We aim to investigate whether the author of a convincing argument is imagined as having specific personality traits. To address this question, we compute Pearson correlations between argument convincingness and the envisioned education level and personality traits of the author.

We observe a statistically significant positive correlation between convincingness and the author’s perceived education level (Pearson’s $r = 0.20$). Additionally, all personality traits show statistically significant correlations with convincingness, most of them positive, except for Extraversion. The correlation results can be found in Table 1.

This leads us to the following conclusion: The more convincing the argument a speaker is presenting, the more they are perceived to have a high education, to be agreeable, conscientious and emotionally stable. Openness and Extraversion play a less important role, with less convincing arguments associated with higher Extraversion.

4.3 RQ 3: How does annotator stance influence the similarity of annotator and envisioned author?

Overall similarity. To answer the question if the annotators imagine the author to be similar to themselves we compare the education level and personality traits of the annotators with the values they assigned to the author, by looking at average values and Pearson correlation. We find that the mean difference between the education levels of annotator and speaker is -0.86 , corresponding to roughly one education level, and a Pearson correlation of 0.23 .

We find small but statistically significant correlations between all personality traits of the annotator

Trait	Agree	Disagree
Extrav.	-0.06*	-0.13***
Agreeabl.	0.20***	0.07*
Conscient.	0.13***	0.05
E. Stab.	0.10***	0.02
Open.	0.06*	0.02

Table 2: Pearson correlation (r) between participant traits and assumed author traits, by stance agreement. Significance levels ** $p < .01$, *** $p < .001$

and the author, ranging from weakest -0.09 for Extraversion to strongest 0.13 for Agreeableness. This shows that while there are big differences between single annotators, the properties of the imagined author are not entirely determined by the argument text but also to a significant degree by the annotators themselves.

Stance Alignment. The ContArgA corpus provides both the stance of each argument (pro/con) and the annotator’s own stance, enabling a comparison of aligned vs. opposing stance conditions. We partition annotations accordingly and compute correlations between annotators’ own education level and personality traits and those attributed to imagined authors.

We find that the assumed author education correlates statistically significantly with the annotator education in both cases, but is stronger when annotator and argument agree ($r = .28$) then when they disagree ($r = .21$). When looking at personality traits (see Table 2) we see statistically significant correlations for all traits when stances agree, with the strongest correlation for Agreeableness. For non-agreeing stances only Extraversion and Agreeableness show statistically significant correlations. This shows that when a speaker agrees with the presented argument they assume the author to be similar to themselves, except for the trait of Extraversion, where there is a negative correlation.

4.4 RQ 4: Does the variable of convincingness help in modeling perceived author personality?

The data set analysis shows that argument convincingness has a strong statistically significant influence on perceived author personality traits, with higher correlations than stance alignment. For this reason we examine the role of argument convincingness in modeling perceived author personality.

Trait	Gemma		LLaMA 3.2		Mistral		Mixtral		RoBERTa Reg.
	Rand	Gold	Rand	Gold	Rand	Gold	Rand	Gold	Gold
Extr.	+0.13	+0.13	+0.33	+0.36	-0.27	-0.31	+0.36	+0.22	+0.008
Agree.	+0.50	+0.50	+0.20	+0.36	-0.27	-0.29	+0.08	-0.55	+0.10
Consc.	+0.02	+0.05	+0.34	+0.20	-0.08	-0.10	+0.13	-0.24	+0.04
E. Stab.	+0.04	+0.04	-0.11	+0.04	+0.14	+0.14	+0.18	-0.28	+0.03
Open.	+0.48	+0.45	-0.09	-0.11	+0.18	+0.17	-0.02	-0.32	+0.002

Table 3: **Difference in Personality Trait Prediction Performance when Convincingness Signal is Added (avg RMSE)** Positive values (green) indicate improvement with the convincingness signal; negative values (red) indicate worse performance. For LLMs we compare improvements with a random convincingness signal and the gold convincingness signal. The RoBERTa model was trained and evaluated with the gold convincingness signal using 5-fold cross-validation. Absolute RMSE values are reported in Appendix C.

LLMs. We hypothesize that if LLMs mimic human assessments of an argument’s author, then their predictions of the author’s Big Five personality traits should improve when the annotator-assigned convincingness value is provided. To test this hypothesis, we use four different LLMs to annotate the arguments, using three prompting setups: One providing the human-annotated convincingness value with the prompt, one providing a random number as convincingness score and one only displaying the argument text without any convincingness information. We compute the root mean square error (RMSE) of the predicted personality trait values with the gold label provided by the annotator. Results can be seen in Table 3.

We can see that adding the annotator-assigned convincingness value to the prompt does not lead to a consistent improvement in prediction quality for Llama3.2, Mistral and Mixtral, worsening prediction quality for some traits. While prediction quality in Gemma is improved when the annotator-assigned convincingness value is provided, the same is also the case with a random value, suggesting that maybe the mention of convincingness as a keyword in the prompt leads the model to a different performance rather than the value itself.

To determine whether the correlation between convincingness values and specific personality trait is present in the LLM predictions, we run the same analyses as used for RQ3, computing the Pearson correlation between the convincingness value assigned to the argument by a human annotator and the different LLM-predicted personality traits of the author. We find that the LLMs do not replicate the connection between higher convincingness and higher Agreeableness, Conscientiousness and Emotional Stability, showing no statistically sig-

nificant correlation between convincingness and personality traits.

This can be seen as evidence that the tested LLMs are not intrinsically capable of reproducing this specific aspect of social meaning via zero-shot prompting.

RoBERTa-based Model. In the RoBERTa-based experiments, we evaluate whether incorporating annotator-assigned convincingness as an input signal improves performance relative to a model that omits this information, testing our hypothesis that convincingness aids in modeling perceived author personality. To do so we train both models and perform 5-fold cross-validation. We compute the root mean square error (RMSE) of the predicted personality trait values with the gold label provided by the annotator.

To test whether the differences between models’ performance are statistically significant we use paired, instance-level resampling (Dror et al., 2018). For each predicted personality trait value, we compute RMSE for both models and form paired differences. We then perform a non-parametric paired bootstrap with 10,000 resamples to estimate the sampling distribution of the mean difference. The observed mean difference is -0.0221 , with a 95 percent confidence interval of $[-0.0301, -0.0141]$, indicating a reliable overall advantage for the model that uses the subjective convincingness information. On the level of specific personality traits, the convincingness-informed model show statistically significant improvement in the prediction of Extraversion, Agreeableness and Openness. All results can be seen in Appendix B.

We find that unlike in the LLM prompting experiment, the RoBERTa model using the convincing-

ness signal performs better on the prediction of all personality traits. While LLMs struggle to integrate the convincingness signal when given in a prompt, explicit integration of the signal in the RoBERTa-based model architecture helps the model to make use of the signal.

When computing the Pearson correlation between the convincingness value assigned to the argument by a human annotator and the personality traits predicted by the RoBERTa-based model, we find that the correlation is much higher than the correlation in the human-annotated data, indicating that while the model uses the signal, it overly relies on this signal rather than learning other clues from text.

5 Discussion

Our results show that when annotators encounter an argument with little context, they still form an impression of the author of the argument. While these impressions are not universal across all annotators, there are nevertheless consistent trends: The argument’s convincingness and the perceived author personality traits are correlated, with more convincing arguments being associated with a higher score in personality traits like Agreeableness, Conscientiousness and Emotional Stability.

This could be seen as a textual expression of the *Halo Effect* known to social psychology (Thorndike, 1920). The Halo Effect describes how a single salient positive trait such as physical attractiveness can create a favorable impression of a person, which then shapes the overall perception of their other qualities, such as assumed intelligence or trustworthiness. In our case, instead of attractiveness, the convincingness of the argument could influence the perception of the personality traits. Previous work shows this effect in multi-modal LLMs used for making hiring decisions (Kim et al., 2025) and in the reproduction of body image stereotypes by LLMs (Asad et al., 2025). While this phenomenon is well known in psychology and taken into account when designing studies, this is to our knowledge the first study to show this phenomenon in textual arguments and to use it in the design of a computational model for author perception.

We also show that people perceive an author to be more similar to themselves when they agree with the stance of the argument. We find this to be in line with social projection theory (Machunsky et al., 2014). When faced with an argument and very little other information about the author, the

annotator might place the imagined author as an in-group or an out-group member based on stance alignment, and therefore assign them more similarity if stances match. This in turn can lead to a different assessment of the information conveyed in the text or in future interactions with the author, e.g. finding statements more trustworthy because they are perceived to originate from ones in-group. While there is some previous work investigating social projection in LLMs outputs (Sumita et al., 2025), this topic seems less explored in NLP and can offer an avenue for future work.

Lastly, we find that LLMs, at least in a zero-shot prompting approach, do not necessarily mimic these human behaviors. This points to the possibility that the tested light-weight models lack the implicit social reasoning or contextual inference abilities required to reconstruct perceived author characteristics from argumentative text alone, which is in line with previous works about the shortcomings of LLMs with regards to social reasoning (Lauscher et al., 2022). We also show that this social knowledge can be learned, using a lightweight approach that does not need adaptation of language models but relies on a regression head on top of fixed RoBERTa embeddings.

Our work may raise the question of desired model behavior. Should LLMs or other models that humans interact with represent envisioned author personality in line with their users? Previous work shows that humans integrate author-specific information when judging the convincingness of arguments (Petty and Cacioppo, 1986), which suggests that models approximating such judgments may require mechanisms for representing this information.

In this work, the LLM experiments are not intended to encourage anthropomorphic interpretations or to propose that models ought to construct a stereotypical author profile. Instead, they serve to probe whether current systems can use author-related cues in a way that is informative for modeling human assessments. This perspective frames LLMs not as stand-ins for human annotators, but as instruments for exploring how particular contextual factors may or may not be captured computationally. Understanding the boundaries of these capabilities is essential for designing methods that reflect human argument evaluation.

Our findings have implications with regards to the study of persuasion and misinformation spread on social media, where readers encounter argu-

ments on contentious topics without the grounding of a personal relationship with an author or a longer discourse to contextualize statements. Readers are likely to construct their own mental models of authors based solely on textual cues, which can influence how persuasive they find an argument. This may reinforce existing biases or in-group preferences – a dynamic that plays a critical role in the amplification of polarizing content online. We therefore encourage the explicit modeling of social phenomena like the Halo Effect and social projection theory in future work.

6 Conclusion and Future Work

In this paper we examine argumentative texts with regards to the creation of social meaning, asking whether the readers of arguments imagine an author of the argument and what factors influence the properties of this envisioned author. Using statistical analysis we find that readers do imagine an author, showing that readers assign social roles and qualities based on the text and that their judgments in part reflect predominant demographic stereotypes. We also show that the convincingness of an argument is correlated with the envisioned author's education level and their personality traits, linking more convincing arguments with more educated and more agreeable and conscientious envisioned authors. We also find that annotators envision authors to be more similar to themselves if the stance of the presented argument (for or against a certain topic) aligns with the stance of the annotator.

Our work builds a connection to work in social psychology, where the Halo Effect is a well described phenomenon. The correlation between the convincingness of an argument and personality traits assigned to the envisioned author of the argument can be seen as one expression of the Halo Effect, where one perceived positive trait of a person (in our case high convincingness) influences the assessment of other unrelated traits. As to our knowledge this is the first work to show this effect in connection with argumentative text.

Future work should build on these findings by integrating them into models of argument and social interaction, addressing the need for representations of social meaning that are disentangled from denotational meaning. Ultimately, modeling social meaning can enable language representations that move beyond surface forms of text to capture the nuances of different usage contexts.

7 Limitations

Our work is limited to the data presented in the ContArgA corpus, which was created by annotators from a relatively constrained geographic area (the UK and Ireland). This limits the strength of the deductions based on it. It also contains only English language arguments. Despite these limitations the ContArgA corpus is to our knowledge the only corpus that examines envisioned author properties in the context of arguments, which makes it the best option to answer our research questions.

Work in perspectivism calls for modeling of single annotators, or a distribution of annotator judgments rather than one gold label. When using convincingness as a signal during modeling, we do so in an unaggregated manner, using the convincingness judgment of a single annotator as input and evaluating model performance against that same annotator's gold labels. Nevertheless, we report model performance averaged over all data points, which could be seen as a break from perspectivist modeling principles. Future work could be dedicated to examining model performance for different groups of annotators or as a distribution over annotators.

8 Ethical Considerations

Central to this work is the ContArgA corpus, which was collected prior to this work and is publicly available data. The collection of the ContArgA corpus was approved by the ethics boards, and conducted via online crowdsourcing for which the annotators were payed and provided consent for the usage of the data. The corpus does not contain any data that would allow for personal identification.

A guiding question for ethical consideration is who profits from our work and who is likely to get harmed by intended or unintended uses of it. We aim for this work to help in the modeling of social interactions online, specifically when arguments are encountered with little contextual information about the author, e.g., in social media contexts. Our findings can help to shine a light on the spread of misinformation or the mechanics of radicalization in online spaces. This can help to make these spaces safer for all participants.

We do not attempt to predict ground truth personality traits of real people, which would be a violation of privacy if used without consent. We are modeling what author readers imagine when they read an argument. This knowledge can be used

not only to study, but also to manipulate author perceptions. When used with malicious intent this knowledge could be used in the creation of convincing chatbots that exploit the connection between stance alignment and perception of personality similarity to manipulate people.

Lastly, AI assistance was used during the creation of this paper. We used GitHub Copilot 0.36.2⁵ for coding and ChatGPT-4⁶ for assistance with L^AT_EX formatting of tables and rephrasing of text for clarity and grammatical correctness.

9 Acknowledgment

This project has been conducted as part of the Emcon (The Interplay of Emotions and Convincingness in Arguments) project, which is funded by the German Research Foundation (DFG, project KL2869/12–1, project number 516512112).

References

- Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, and Pushpak Bhattacharyya. 2025. “You are beautiful, body image stereotypes are ugly!” BISTereo: A benchmark to measure body image stereotypes in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24471–24496, Vienna, Austria. Association for Computational Linguistics.
- Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. *Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages*. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.
- Mary Bucholtz and Kira Hall. 2010. *Locating identity in language*, pages 18–28. Language and Identities, Edinburgh University Press.
- Silvia Casola, Yang Janet Liu, Siyao Peng, Oliver Kraus, Albert Gatt, and Barbara Plank. 2025. *References matter: Investigating the impact of reference set variation on summarization evaluation*. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 274–291, Hanoi, Vietnam. Association for Computational Linguistics.
- Paul Costa and Robert McCrae. 1999. *A five-factor theory of personality*. *Handbook of personality: Theory and research*, 2(01):1999.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. *The hitchhiker’s guide to testing statistical significance in natural language processing*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. *Perspectivist approaches to natural language processing: a survey*. *Lang. Resour. Eval.*, 59(2):1719–1746.
- Samuel Gosling, Peter Rentfrow, and William Swann Jr. 2003. *Ten-item personality inventory*. *Journal of Research in Personality*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and (many others). 2024. *The llama 3 herd of models*. *arXiv preprint*, arXiv:2407.21783.
- Lynn Greschner, Sabine Weber, and Roman Klinger. 2025. *Trust me, I can convince you: The contextualized argument appraisal framework*. *arXiv preprint arXiv:2509.17844*. Accessed: 2025-10-02.
- 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  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*. *arXiv preprint*, arXiv:2310.06825.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. *Mixtral of experts*. ArXiv:2401.04088 [cs].
- Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniec, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. *What if ground truth is subjective? personalized deep neural hate speech detection*. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.
- Kyusik Kim, Jeongwoo Ryu, Hyeonseok Jeon, and Bongwon Suh. 2025. *Blinded by context: Unveiling the halo effect of MLLM in AI hiring*. In *Findings of the Association for Computational Linguistics: ACL*

⁵<https://github.com/microsoft/vscode-copilot-chat>

⁶<https://chatgpt.com>

- 2025, pages 26067–26113, Vienna, Austria. Association for Computational Linguistics.
- Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. [Items from psychometric tests as training data for personality profiling models of Twitter users](#). In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 315–323, Dublin, Ireland. Association for Computational Linguistics.
- Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. [SocioProbe: What, when, and where language models learn about sociodemographics](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. 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](#). *CoRR*, abs/1907.11692.
- Maya Machunsky, Claudia Toma, Vincent Yzerbyt, and Olivier Corneille. 2014. [Social projection increases for positive targets: Ascertaining the effect and exploring its antecedents](#). *Personality and Social Psychology Bulletin*, 40(10):1373–1388.
- Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Faret, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, Kathleen Keane, et al. 2024. [Gemma: Open models based on Gemini research and technology](#). *arXiv preprint*, arXiv:2403.08295.
- Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. [Computational sociolinguistics: A Survey](#). *Computational Linguistics*, 42(3):537–593.
- Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. [On learning and representing social meaning in NLP: a sociolinguistic perspective](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [Film: Visual reasoning with a general conditioning layer](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Richard E. Petty and John T. Cacioppo. 1986. [The elaboration likelihood model of persuasion](#). volume 19 of *Advances in Experimental Social Psychology*, pages 123–205. Academic Press.
- Daniele Pizzolli and Carlo Strapparava. 2019. [Personality traits recognition in literary texts](#). In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.
- E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. 2000. [The gender stereotyping of emotions](#). *Psychology of women quarterly*, 24(1):81–92.
- Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. [Investigating subjective factors of argument strength: Storytelling, emotions, and hedging](#). In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.
- Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2025. [Cognitive biases in large language models: A survey and mitigation experiments](#). In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, SAC '25*, page 1009–1011, New York, NY, USA. Association for Computing Machinery.
- Edward Thorndike. 1920. [A constant error in psychological ratings](#). *Journal of applied psychology*, 4(1):25–29.
- Marina Tiuleneva, Vadim A. Porvatov, and Carlo Strapparava. 2024. [Big-five backstage: A dramatic dataset for characters personality traits & gender analysis](#). In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 114–119, Torino, Italia. ELRA and ICCL.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. *TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637, Portorož, Slovenia. European Language Resources Association (ELRA).

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. *VariErr NLI: Separating annotation error from human label variation*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. *Socially aware language technologies: Perspectives and practices*. *Computational Linguistics*, 51:689–703.

A LLM prompt

Figure 3 shows the full prompt used in our LLM experiments. It uses the same text as the text shown to the human annotators during the creation of the ContArgA corpus.

B Bootstrap Levels

Table 4 shows Bootstrap levels for all personality traits for the RoBERTa-based models.

Trait	95% CI Low.	95% CI Up.
Extraversion	−0.0141	−0.0072
Agreeableness	−0.1060	−0.0617
Conscient.	−0.0135	0.0169
Emot. Stability	−0.0190	0.0057
Openness	−0.0176	−0.0046

Table 4: Bootstrap 95% confidence intervals for the difference in RMSE ($\Delta = \text{RMSE}_A - \text{RMSE}_B$). Negative values indicate better performance by the convincingness-informed model. Intervals excluding zero denote statistically significant differences.

C Full RMSE values for all tested LLMs

Table 5 shows all RMSE values for all tested systems.

```

</system>
You are an expert human annotator. You must fill out the TIPI questionnaire.
Use ONLY numbers 1-7. If unsure, choose the closest valid number.
</user>
You are in a bustling town hall, a bright room with folding chairs and a low hum of conversation.
Posters line the walls, showcasing various perspectives on the issue at hand. A panel of speakers
sits at a long table in front, ready to listen as community members approach the podium. The
crowd is a mix of familiar faces, local officials, advocates, and reporters poised with cameras.
As the moderator calls for public comments, silence falls. Curious and skeptical eyes focus on
the podium as each speaker takes their turn.
{speaker_age} {speaker_gender} approaches the microphone and makes a statement: {argument}
You judge the convincingness of the statement on a scale from 1 (not convincing) to
5 (very convincing) as {convincingness}.

Here are a number of personality traits that may or may not apply to the person.
Please choose a number from 1 (Disagree strongly) to 7 (Agree strongly) to indicate the extent
to which you think the person would agree or disagree with that statement.
You should rate the extent to which the pair of traits applies to the person, even if one
characteristic applies more strongly than the other.

Answer in this format:
1. [rating]
2. [rating]
3. [rating]
4. [rating]
5. [rating]
6. [rating]
7. [rating]
8. [rating]
9. [rating]
10. [rating]

The person is extraverted, enthusiastic.
The person is critical, quarrelsome.
The person is dependable, self-disciplined.
The person is anxious, easily upset.
The person is open to new experiences, complex.
The person is reserved, quiet.
The person is sympathetic, warm.
The person is disorganized, careless.
The person is calm, emotionally stable.
The person is conventional, uncreative.

---
Important:
- Base your ratings only on the information given.
- Do not explain your ratings. Just output the numbers as shown above.

```

Figure 3: LLM prompt used for TIPI annotation.

Model	Setting	Extraversion	Agreeableness	Conscient.	Emotional Stab.	Openness
Gemma	With Conv	1.73	1.75	1.22	1.40	2.64
Gemma	Random Conv	1.73	1.75	1.25	1.40	2.61
Gemma	No Conv	1.86	2.25	1.27	1.44	3.09
LLaMA 3.2	With Conv	1.65	1.98	1.81	1.93	2.42
LLaMA 3.2	Random Conv	1.68	1.87	1.67	2.08	2.40
LLaMA 3.2	No Conv	2.01	2.07	2.01	1.97	2.31
Mistral	No Conv	1.74	2.04	1.42	1.59	2.88
Mistral	Random Conv	2.01	2.31	1.50	1.45	2.70
Mistral	With Conv	2.05	2.33	1.52	1.45	2.71
Mixtral	With Conv	1.25	2.52	1.74	1.97	2.51
Mixtral	Random Conv	1.11	1.89	1.37	1.51	2.21
Mixtral	No Conv	1.47	1.97	1.50	1.69	2.19

Table 5: RMSE by personality trait across models and settings.

A Transformer and Prototype-based Interpretable Model for Contextual Sarcasm Detection

Ximing Wen
Drexel University,
Philadelphia, USA
xw384@drexel.edu

Rezvaneh Rezapour
Drexel University,
Philadelphia, USA
sr3563@drexel.edu

Abstract

Sarcasm detection, with its figurative nature, poses unique challenges for affective systems designed to perform sentiment analysis. While these systems typically perform well at identifying direct expressions of emotion, they struggle with sarcasm’s inherent contradiction between literal and intended sentiment. Since transformer-based language models (LMs) are known for their efficient ability to capture contextual meanings, we propose a method that leverages LMs and prototype-based networks, enhanced by sentiment embeddings, to conduct interpretable sarcasm detection. Our approach is intrinsically interpretable without extra post-hoc interpretability techniques. We test our model on three public benchmark datasets and show that our model outperforms the current state-of-the-art. At the same time, the prototypical layer enhances the model’s inherent interpretability by generating explanations through similar examples in the reference time. Furthermore, we demonstrate the effectiveness of incongruity loss in the ablation study, which we construct using sentiment prototypes.

1 Introduction

The task of automatically detecting sarcasm introduces a complex challenge in natural language processing (NLP). This nuanced task bridges the gap between sentiment analysis and text interpretation, highlighting the complexity of understanding and analyzing sarcasm in written language (Ilavarasan et al., 2020). Sarcasm, characterized by a sharp, often humorous contrast between literal and intended meanings of statements, poses unique difficulties for computational models. These challenges stem from sarcasm’s deep reliance on contextual clues, tone, and common human experiences. This complexity is further amplified in digital communication, where non-verbal cues are largely absent, making it essential to develop advanced models

capable of interpreting such subtleties with high accuracy.

Deep learning models, especially transformer-based language models (LMs), have significantly contributed to advancements in NLP, offering powerful tools for sentiment analysis (Bu et al., 2024), emotion detection (Tu et al., 2024), and, by extension, sarcasm detection (Helal et al., 2024). More specifically, to study sarcasm detection, there is a trend to leverage LMs and a variety of features generated by different models to improve prediction accuracy (Cai et al., 2019; Bedi et al., 2021). However, a persistent critique of deep learning models is their “black-box” nature, which obscures the decision-making process and hinders their interpretability. Current approaches usually adopt post-hoc interpretability methods Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) or attention mechanisms to explain a model’s decision (Ribeiro et al., 2016; Mardaoui and Garreau, 2021; Kumar et al., 2021b). However, those explanations are still word-level and can only tell which part of the input they are looking at. As for sarcasm detection, when the text does not contain words that convey strong sentiment and instead uses other ways, such as analogy to express sarcasm, the word-level explanations could set similar weights to words in the sentence, and humans are ill-equipped to interpret them.

To address this challenge, our paper presents an intrinsically interpretable NLP framework that integrates prototype classification networks (Li et al., 2018) with multi-view of semantic embedding and sentiment embedding from large-scale pre-trained transformer language models. To the best of our knowledge, our study is the first to apply a prototype-based network in sarcasm detection. This is achieved through a unique training regimen that enables the network to learn a collection of pro-

prototype tensors, which encapsulate latent clusters of training samples. At the point of inference, the model makes classification judgments solely based on the similarity to these prototypes, allowing for the model’s decisions to be transparently explained by referencing the training examples most closely aligned with the top-matched prototypes. Together with a sentiment-prototype-based incongruity loss that captures the difference between implicit and explicit sentiment, our approach not only provides clear, human-understandable explanations for its predictions but also achieves state-of-the-art performance. The key contributions of our methods can be summarized as follows:

1. We propose a novel interpretable framework for sarcasm detection. Our framework is built upon a prototype-based network leveraging semantic embedding and sentiment embedding from pre-trained transformer-based language models.
2. Extensive experiments on three public benchmark datasets show that our approach achieves state-of-the-art performance while being interpretable. We also conduct an ablation study to analyze the influence of incongruity loss in our model.
3. We conduct case studies to show that our model can generate human-readable, sentence-level explanations for the model’s reasoning process at the reference time. Our model and training code are available here: <https://github.com/social-nlp-lab/Sarcasm-Detection>.

2 Related Work

2.1 Contextual Sarcasm Detection

Initial research in sarcasm detection primarily relied on simple lexical and syntactic features, and the classifiers are categorized as **Content-based Models** (Carvalho et al., 2009; Davidov et al., 2010; González-Ibáñez et al., 2011), leveraging features like n-grams, and part-of-speech tags (Riloff et al., 2013; Tepperman et al., 2006; Tsur et al., 2010).

With the increase in the usage of sarcasm on online platforms in recent years, the performance of the sarcasm detection model is usually compromised in terms of robustness when faced with texts plagued by grammatical inaccuracies (Švelch,

2015). Moreover, these texts are usually a series of posts and comments that are highly temporal and contextual. As a result, relying solely on linguistic cues has become inadequate, prompting researchers to develop **Context-based Models**. A prominent strand of this research involves mining sentiment incongruity in sarcastic texts with attention-mechanism (Pan et al., 2020; Najafabadi et al., 2024) to improve models’ performance. Diverging from these methods, our approach adopts sentiment prototypes to discern both implicit and explicit sentiments within texts, enhancing the interpretability of the reasoning process. Furthermore, other scholars are exploring the modeling of user interactions via Graph Convolutional Networks (GCN) (Mohan et al., 2023) or employing commonsense knowledge transformers like COMET (Yu et al., 2023).

2.2 Explainability of Transformer Language Models

Post hoc Interpretability: When leveraging LM’s ability to understand context, the model’s complexity prevents people from understanding the model’s reasoning process. In the field of XAI, post-hoc explainable approaches, such as LIME (Ribeiro et al., 2016), and SHAP (Lundberg and Lee, 2017), are used to generate explanations for transformers’ results (Kumar et al., 2021a) through analyzing weights for each input word representation. However, these methods are now facing the challenge of being faithful and robust. Various studies show the sensitivity of the models to small perturbations in the target model’s inputs or parameters (Ivankay et al., 2022; Mardaoui and Garreau, 2021). Other researchers, such as Akula and Garibay (2021) and Rishabh Misra and Prahal Arora (2018), have explored the use of attention mechanisms to interpret prediction outcomes, focusing on how attention scores are allocated across individual words. However, these explanations remain at the word level, indicating only which parts of the input the model attends to. In the context of sarcasm detection, where sarcasm may be conveyed through mechanisms such as analogy rather than explicit sentiment, word-level explanations can assign similar weights to different words in a sentence. This makes it challenging for humans to interpret the underlying sarcastic intent effectively. To address this, we aim for more intuitive and sparse explanations: well-descriptive but short-sequence prototypes.

Prototype-based Reasoning in Deep Neural Networks: Prototype-based methods emphasize that visualizing the reasoning process through prototypes can significantly improve the intuitiveness of interpretation. This approach, leveraging prototype-based reasoning, has been a core aspect of interpretability in classical models for decades, as evidenced by research from Cupello and Mishevich (1988), Fikes and Kehler (1985), and Kim et al. (2015). A pioneering example of incorporating prototypical learning into deep neural networks is the work by Chen et al. (2019), who introduced a novel neural network design for image classification. By inserting a prototype layer following the convolutional layers, the model compares convolution responses across different locations in the predicted image with predefined prototypes. Furthermore, this allows users to grasp why an image is classified in a certain way, such as understanding why a bird is identified as a ‘red-bellied woodpecker’ due to its distinct red belly and head, along with black and white wing stripes. Following this work, researchers explored incorporating prototype layer with transformer-based encoders, such as Universal Sentence Encoder, BERT, BART (Bidirectional and Auto-Regressive Transformers) in fake news detection and hotel review classification (Das et al., 2022; Hong et al., 2024; Wen, 2024). Sarcasm, due to its nature, can benefit from such reasoning provided by prototype-based models. However, this approach is still underexplored.

3 Methodology

Our approach, as shown in Figure 1, first encodes semantic embedding and sentiment embedding with the Semantic View encoder and the Sentiment View encoder separately, and then both the encoded semantic embedding and sentiment embedding are fed into two separate prototype layers. Finally, the output distance vectors from the two prototype layers are concatenated and sent to the output layer to make the prediction.

3.1 Semantic View

The Semantic View focuses on capturing contextual meanings from text input. We use pre-trained language models (LMs) from Hugging Face (Wolf et al., 2020) as encoders to extract sentence-level embeddings. When the input comment has ancestor posts (e.g., SARC Dataset by Khodak et al. (2018)), which help set up the stage for the conversation and

provide more context information, we concatenate the nearest ancestor with the comment as a whole and encode the embeddings with a pre-trained language model (e.g. Sentence-BERT (SBERT) by Reimers and Gurevych, 2019, RoBERTa by Liu et al., 2019b).

$$\mathbf{e}^{ct} = \text{Encoder}(\mathbf{x}) \quad (1)$$

where x denotes the input text, \mathbf{e}^{ct} is the semantic information representation vector.

3.2 Sentiment View

For sentiment feature extraction, we decompose the text into two parts: **Explicit Part**, which are sentiment words extracted following (Joshi et al., 2015), and **Implicit Part**, which is the rest of the text capturing the implicit sentiments.

Both segments are processed through the Sentiment Encoder, specifically the SiEBERT model (Hartmann et al., 2023), to separately obtain the vector representations $\mathbf{e}^{st,ep}$ and $\mathbf{e}^{st,ip}$ of the CLS token from the final hidden state. For brevity, we use \mathbf{e}^{st} to represent both of them in the following equation:

$$\mathbf{e}^{st} = \text{SiEBERT}(\mathbf{x}) \quad (2)$$

where \mathbf{x} denotes the input text, \mathbf{e}^{st} is the sentiment information representation vector. The explicit representation $\mathbf{e}^{st,ep}$ is labeled as $\mathbf{z}^{st,ep}$ with SiEBERT. The implicit representation $\mathbf{e}^{st,ip}$ is labeled as $\mathbf{z}^{st,ip}$ identically to $\mathbf{z}^{st,ep}$ for non-sarcastic inputs, whereas it is labeled oppositely for sarcastic inputs.

Fallback Strategy The implementation incorporates a fallback strategy for cases where no explicit sentiment cues are identified in the input text. Specifically, when the sentiment word detector returns an empty set, a predefined neutral-to-positive sentiment token sequence is used as a surrogate explicit sentiment representation. This surrogate input is encoded using SiEBERT following the same preprocessing and inference pipeline applied to detected sentiment words, ensuring architectural consistency.

3.3 Prototypical Layer

As shown in Figure 1, after the input is encoded into a latent semantic representation $\mathbf{e}^{ct} \in \mathbb{R}^{d_s}$ through Semantic View and two latent sentiment representation $\mathbf{e}^{st,ep} \in \mathbb{R}^{d_m}$, $\mathbf{e}^{st,ip} \in \mathbb{R}^{d_m}$ through

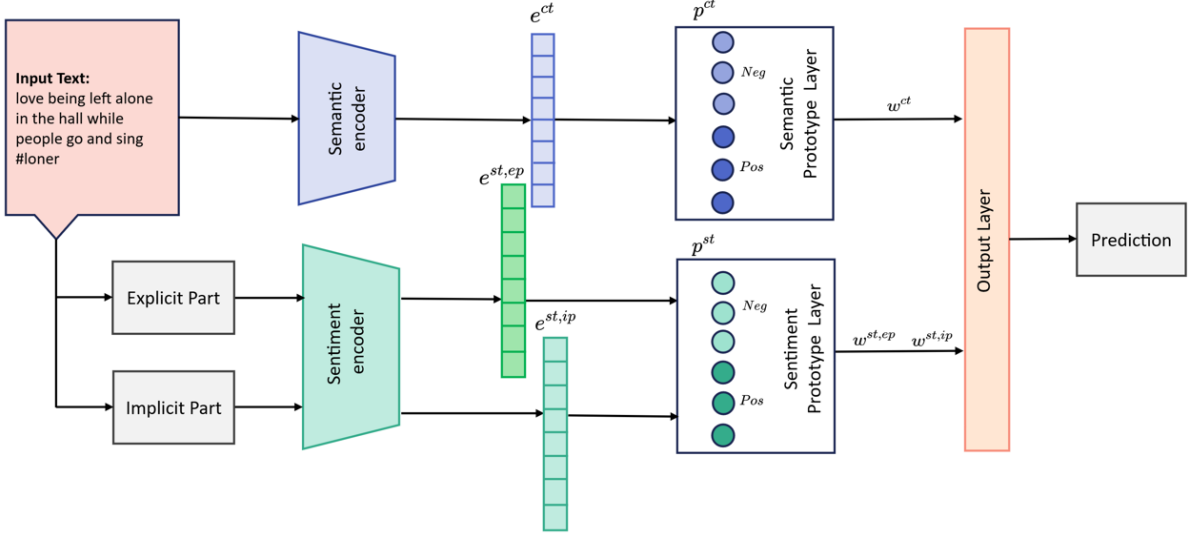


Figure 1: Diagram of the proposed model architecture and workflow.

Sentiment View separately, each representation is fed into a prototype layer respectively.

Semantic Prototype: The Semantic Prototype Layer consists of k_a prototypes $\mathbf{P}^{ct} = \{\mathbf{p}_j^{ct}\}_{j=1}^{k_a}$, where $\mathbf{p}_j^{ct} \in \mathbb{R}^{d_s}$, the same dimension of the encoded latent semantic feature \mathbf{e}^{ct} . Each prototype, represented as a tensor, encapsulates a cluster of training examples. To ensure that both sarcastic and non-sarcastic classes are effectively represented by the learned prototypes, we allocate a fixed number of prototypes to each class. Utilizing k-means clustering (Hartigan and Wong, 1979), we segment the training data of each class into multiple clusters and subsequently initialize the prototypes for each class with these cluster centers. These prototypes are trained through loss terms \mathcal{L}_{cls}^{ct} and \mathcal{L}_{sep}^{ct} defined in §3.5. This layer calculates the similarity between embedding \mathbf{e}^{ct} and each prototype \mathbf{p}_j^{ct} with a Radial basis function (RDF) kernel function as follows:

$$\text{sim}(\mathbf{e}^{ct}, \mathbf{p}_j^{ct}) = \exp\left(-\frac{\|\mathbf{e}^{ct} - \mathbf{p}_j^{ct}\|_2^2 + \varepsilon}{\sigma^2}\right) \quad (3)$$

This similarity score increases monotonically as the Euclidean distance increases. σ is a user-specified value that determines how quickly the similarity score increases as the distance between \mathbf{e}^{ct} and \mathbf{p}_j^{ct} decreases. A small σ makes the kernel function more sensitive to changes in distance, leading to a kernel matrix with more localized information about the data points. This can make the model more sensitive to noise in the data. A large

σ , on the other hand, produces a smoother kernel function that is less sensitive to the exact distance between data points, potentially making the model more robust. We also add a small value ε set as $1e-4$ for numerical stability. We get vector \mathbf{w}^{ct} by calculating the similarity score between \mathbf{e}^{ct} and each prototype vector within \mathbf{P}^{ct} .

Sentiment Prototype Layer: The Sentiment Prototype Layer comprises k_b prototypes $\mathbf{P}^{st} = \{\mathbf{p}_j^{st}\}_{j=1}^{k_b}$, where $\mathbf{p}_j^{st} \in \mathbb{R}^{d_m}$. These prototypes are categorized into positive prototypes where $\mathbf{p}^{st} \in \mathbf{P}_1^{st}$ and negative prototypes where $\mathbf{p}^{st} \in \mathbf{P}_0^{st}$. We initialize \mathbf{P}_1^{st} using the k-means cluster centers computed from the positive training data labeled by SiEBERT and \mathbf{P}_0^{st} using those derived from the negative training data. When clustering, we only use non-sarcastic training samples, without dividing the text into implicit and explicit parts.

We calculate the similarity between embedding \mathbf{e}^{st} and each prototype \mathbf{p}_j^{st} as follows:

$$\text{sim}(\mathbf{e}^{st}, \mathbf{p}_j^{st}) = \exp\left(-\frac{\|\mathbf{e}^{st} - \mathbf{p}_j^{st}\|_2^2 + \varepsilon}{\sigma^2}\right) \quad (4)$$

We calculate similarity scores for the explicit and implicit representations, $\mathbf{e}^{st,ep}$ and $\mathbf{e}^{st,ip}$, against each prototype vector within \mathbf{P}^{st} , yielding the similarity vectors $\mathbf{w}^{st,ep}$ and $\mathbf{w}^{st,ip}$, respectively.

3.4 Output Layer

The output layer is a fully connected layer followed by a sigmoid layer. It takes the concatenation of the

extracted similarity vectors \mathbf{w}^{ct} , $\mathbf{w}^{st,ep}$ and $\mathbf{w}^{st,ip}$ from prototype layers as input and predict the likelihood of a text being sarcastic.

We intentionally employ concatenation to fuse these representations in order to preserve the independence of prototype similarity vectors, which is essential for interpretability in prototype-based models, as it allows direct attribution of the final prediction to individual semantic or sentiment prototypes. More complex fusion mechanisms, such as cross-attention or gating, may introduce stronger interactions but would entangle prototype signals and obscure their individual contributions. Interactions between explicit and implicit sentiment are therefore not modeled at the fusion layer; instead, they are explicitly captured by the incongruity loss (Eq. 12), which encourages disagreement between explicit and implicit sentiment predictions for sarcastic inputs.

3.5 Loss and Training Algorithm

We construct the loss function with four different terms to ensure both accuracy and interpretability.

Accuracy Loss: The first term is accuracy loss, and it uses cross-entropy loss to optimize the predictive power of the network. In equation 5, f is the output classifier, \mathcal{P}_i is the predicted probability distribution, y_i is the label, and n is the total number of training data points. θ refers to the trainable weights in the classifier.

$$\mathcal{P}_i = f([\mathbf{w}_i^{ct}, \mathbf{w}_i^{st,ep}, \mathbf{w}_i^{st,ip}]) \quad (5)$$

$$\mathcal{L}_{acc} = -\frac{1}{n} \sum_{i=1}^n y_i \log P(\mathcal{P}_i = y_i | \mathbf{x}_i; \theta) \quad (6)$$

Division Loss: To distribute prototypes in the embedding space as much as possible, we design the \mathcal{L}_{div} indicated in equation 7 for both semantic and sentiment prototypes. This loss uses cosine similarity to measure the difference between any two prototypes \mathbf{p}_m and \mathbf{p}_n in P and penalizes it if their similarity is larger than λ . It is particularly beneficial when there are multiple prototypes for a single class, as it promotes the representation of diverse aspects of that class.

$$\mathcal{L}_{div} = \sum_{\mathbf{p}_j, \mathbf{p}_q \in \mathbf{P}, j \neq q} \max(0, \cos(\mathbf{p}_j, \mathbf{p}_q) - \lambda) \quad (7)$$

Clustering and Separation Loss: The clustering loss \mathcal{L}_{cls} and \mathcal{L}_{sep} are inspired by previous work, ProtoPNet (Chen et al., 2019). The clustering loss \mathcal{L}_{cls} ensures each embedding is close to at least one prototype in its own class, and separation loss \mathcal{L}_{sep} encourages each embedding to be distant from prototypes not of its class. Together, \mathcal{L}_{cls} and \mathcal{L}_{sep} push each prototype to focus more on training examples from the same class and less on training examples from other classes.

For **semantic prototypes**, they are defined as follows:

$$\mathcal{L}_{cls}^{ct} = \frac{1}{n} \sum_{i=1}^n \min_{\mathbf{p}_j^{ct} \in \mathbf{P}_{y_i}^{ct}} \|\mathbf{e}_i^{ct} - \mathbf{p}_j^{ct}\|_2^2 \quad (8)$$

$$\mathcal{L}_{sep}^{ct} = -\frac{1}{n} \sum_{i=1}^n \min_{\mathbf{p}_j^{ct} \notin \mathbf{P}_{y_i}^{ct}} \|\mathbf{e}_i^{ct} - \mathbf{p}_j^{ct}\|_2^2 \quad (9)$$

For **sentiment prototypes**, we promote proximity between positive prototypes and training data segments labeled positive ($z_i = 1$) by SiBERT, and likewise align negative prototypes with training data segments labeled negative ($z_i = 0$) through losses defined as follows:

$$\mathcal{L}_{cls}^{st} = \frac{1}{n} \sum_{i=1}^n \min_{\mathbf{p}_j^{st} \in \mathbf{P}_{z_i}^{st}} \|\mathbf{e}_i^{st} - \mathbf{p}_j^{st}\|_2^2 \quad (10)$$

$$\mathcal{L}_{sep}^{st} = -\frac{1}{n} \sum_{i=1}^n \min_{\mathbf{p}_j^{st} \notin \mathbf{P}_{z_i}^{st}} \|\mathbf{e}_i^{st} - \mathbf{p}_j^{st}\|_2^2 \quad (11)$$

The final Clustering and Separation Loss is $\mathcal{L}_{cls_sep} = \mathcal{L}_{cls}^{ct} + \mathcal{L}_{sep}^{ct} + \mathcal{L}_{cls}^{st} + \mathcal{L}_{sep}^{st}$

Incongruity Loss: We hypothesize the presence of incongruity between the explicit and implicit sentiments within a sarcastic comment. For instance, in the sarcastically labeled sentence, ‘‘Oh no, a rainy day again! This is great!’’ the explicit sentiment conveyed by the word ‘‘great’’ appears positive. However, upon closer examination of the context, it becomes evident that the speaker does not favor rainy days, revealing an underlying negative sentiment. Based on this observation, we introduce the incongruity loss defined with cross-entropy to effectively capture this disparity between explicit and implicit sentiment:

$$\mathcal{L}_{inco} = -\frac{1}{n} \sum_{i=1}^n \left(\mathbf{z}_i^{ep} \log P(h(\mathbf{w}_i^{st,ep}) = \mathbf{z}_i^{ep} | \theta) + \mathbf{z}_i^{ip} \log P(h(\mathbf{w}_i^{st,ip}) = \mathbf{z}_i^{ip} | \theta) \right) \quad (12)$$

where h is a MLP based classifier that predicts a probability distribution.

We construct the final loss function L by combining the previously defined loss components, each weighted by their respective coefficient λ . Additionally, we incorporate an L1 regularization term $\|\theta\|$ to promote sparsity in the weights of the output layer. The final loss \mathcal{L} is defined as follows:

$$\mathcal{L} = \mathcal{L}_{acc} + \lambda_1 \mathcal{L}_{div} + \lambda_2 \mathcal{L}_{cls_sep} + \lambda_3 \mathcal{L}_{inco} + \lambda_4 \|\theta\| \quad (13)$$

3.6 Prototype Projection

For improved interpretability, we visualize the semantic prototypes by projecting a prototype vector onto its closest datapoint in the training dataset, measured by Euclidean distance. For the large training dataset with over 120k comments, we did a random pre-sampling of one-tenth of the comments for computation efficiency. Each prototype’s embedding is replaced with the nearest comment’s embedding in the training data point. The alignment of prototypes with training set samples provides an intuitive and easily understandable interpretation for humans.

3.7 Prototype Initialization

We use k-means cluster centers on the training data to initialize prototypes, with a fixed small number of prototypes per class (20). We observed the model to be robust to prototype counts within a reasonable range; extremely small counts reduce coverage, and extremely large counts increase redundancy but yield similar performance due to the division loss.

4 Experimental Setup

In this section, we introduce the datasets used in the experiment and the baseline models.

4.1 Data

We evaluate our methods on the following three public benchmark datasets: (1) **SARC 2.0**¹ (Kho-

¹Link: <https://nlp.cs.princeton.edu/old/SARC/2.0/>

dak et al., 2018), a corpus comprising 1.3 million comments on Reddit. Each comment is self-annotated, and we focus on the primary main balanced variation of it, with 118,940 comments in training and 56,118 comments in the testing set. (2) **Twitter** (Riloff et al., 2013) dataset is sourced from the Twitter platform. Sarcastic tweets are identified using the hashtag #sarcasm, while non-sarcastic tweets lack this hashtag. The dataset consists of 1,368 training examples and 588 test examples. (3) **Sarcasm Corpus V2 Dialogues** (Oraby et al., 2017) is a diverse and richly annotated corpus of sarcasm in dialogue. This dataset is collected from a variety of dialogue sources to capture sarcasm in different conversational contexts, moving beyond traditional social media platforms like Twitter.

4.2 Models and Settings

We employed 5-fold cross-validation to evaluate our model’s performance and fine-tuned the hyperparameters on the validation data. The Optimizer for all neural networks is Adam, and the learning rate is $1e - 4$. We used one single GTX 3090 for each model’s training, and due to the limitation of GPU RAM, when training with LM encoders, we chose a batch size of 60 and an accumulated gradient step of 30. We use early stopping (Fomin et al., 2020) based on the loss of validation data.

Semantic Encoder: In prototype-based architectures, it is a common practice to use Euclidean distance to measure sentence similarity between training examples and prototypes. However, embeddings from transformer-based language models are not typically trained with contrastive loss that leverages Euclidean distance for measuring sentence similarity. To further explore whether this influences the model’s interpretability, we selected RoBERTa-large and SBERT as the encoders for comparison in our experiments:

(1) **RoBERTa-large:** RoBERTa-large is a transformer-based language model developed to improve upon the BERT architecture. We used the [CLS] token embedding from the last hidden states directly for the downstream task.

(2) **SBERT:** SBERT is post-trained on BERT. It uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity.

We used the pre-trained model all-mpnet-base-v2² developed by Microsoft from Hugging Face.

Sentiment Encoder: For sentiment feature extraction, we employed the SiEBERT model (Hartmann et al., 2023), which is a fine-tuned version of the RoBERTa-large model optimized for sentiment classification tasks.

Baseline : For comparative evaluation with our approach, we selected the following methodologies as baselines: Fracking Sarcasm (Ghosh and Veale, 2016), GRNN (Zhang et al., 2016), CNN-LSTM-DNN (Ghosh and Veale, 2016), SIRAN (Tay et al., 2018), MIRAN (Tay et al., 2018), ELMo-BiLSTM (Ilic et al., 2018), A2Text-Net (Liu et al., 2019a), SARC 2.0 (Khodak et al., 2018), and CASCADE (Hazarika et al., 2018). Additionally, we incorporated GRU-Attention model (Akula and Garibay, 2021), which offers intrinsic interpretability through attention scores. We further considered a graph-based method, BERT-GCN (Mohan et al., 2023), alongside ensemble strategies such as Fuzzy-Logic (Dai, 2024) and MULE (Vitman et al., 2023). BiGRU (Najafabadi et al., 2024) is included for its ability to measure sentiment incongruity.

Metrics: We used accuracy, recall, and F1-Score as metrics to evaluate models’ performance. Since we use 5-fold cross-validation, we calculated the average of 5 experiment results on the test dataset for each metric as the final result.

5 Results & Discussion

We mainly focus on answering the two questions through our experiments: **Q1:** What is the performance of our white box framework compared to other baselines, and **Q2:** How well does the given explanation represent the true reasoning of the model? We discuss them separately in the following sections.

5.1 Overall Performance

We represent our experiments on three public benchmark datasets in Table 2, Table 3, and Table 5 separately. Overall, our model achieved the highest accuracy, recall, and F1 scores on the Sarcasm Corpus V2 and SARC datasets, outperforming all other baselines. On the Twitter dataset, our model obtained the best accuracy and F1, while

²The model can be downloaded here: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

Models	Accuracy	Recall	F1
Fracking Sarcasm	89.2	87.9	88.1
GRNN	66.4	64.7	65.4
ELMo-BiLSTM	76.2	75.0	75.9
ELMo-BiLSTM FULL	77.4	73.5	75.3
ELMo-BiLSTM AUG	68.6	70.8	69.4
A2Text-Net	91.3	91.0	90.0
GRU-Attention	97.9	99.6	98.7
BERT-GCN	88.3	87.1	87.3
MULE	93.5	94.1	93.8
Our Model (SBERT)	98.0	97.3	98.7
Our Model (RoBERTa)	98.3	98.6	98.4

Table 2: Results on Twitter dataset.

Models	Accuracy	Recall	F1
GRNN	64.4	61.8	61.2
CNN-LSTM-DNN	67.3	66.7	65.7
SIARN	71.9	71.8	71.8
MIARN	74.2	72.9	72.7
ELMo-BiLSTM	74.6	74.7	74.7
ELMo-BiLSTM FULL	76.2	76.0	76.0
GRU-Attention	77.3	77.2	77.2
BiGRU	79.3	81.4	80.2
Fuzzy-Logic	81.8	80.3	81.0
Our Model (SBERT)	82.1	82.5	82.4
Our Model (RoBERTa)	83.6	83.7	83.6

Table 3: Results on Sarcasm Corpus V2 Dialogues dataset

GRU-Attention yielded the best recall. Additionally, to compare our approach with CASCADE on the SARC dataset, we incorporated personality features into our model. Of the two versions of our model, the one using a RoBERTa encoder generally outperformed its SBERT-based counterpart.

5.2 Case Study for Explanations

In Table 1, we present trained semantic prototypes after projection and the distance score between prototypes and the input comment. The prototypes exhibit a similar topic as the input text, and the keywords are highlighted in yellow by humans for easy illustration (The highlighted words in Table 1 were only meant for reader guidance to aid comparison). For instance, in Case 1, the SBERT prototypes consist of comments skeptical of the efficacy of specific government regulations on cigarettes, and drugs, showing significant overlap with the input comments—a pattern that is also observed in Case 2. Our case study demonstrates that prototype-based models provide more intuitive and human-readable explanations for sarcasm than analyzing a distribution of scores, especially in the absence of strong sarcasm cue words.

When comparing the explanations generated by SBERT and RoBERTa, we did not see a significant

	Input text or similar prototypes	Distance
Case 1	<i>Post:</i> Australian state to try and ban cigarettes to anyone born after 2000. <i>Comments:</i> yes , because drug prohibition always work so well.	
SBERT	<i>Post:</i> France proposes banning the use of underweight models as part of a campaign to stop anorexia. <i>Comment:</i> Finally, a problem government can solve.	1.1187
	<i>Post:</i> After kicking out the dea , the UN says Bolivia has successfully lowered its coca cultivation in each of the last four years. <i>Comment:</i> if the drug agency is literally a drug agency , can I use 1984 as a reference	1.1244
	<i>Post:</i> Gun control debate reignited by plan to import new shotgun into Australia. <i>Comment:</i> I don't understand why Australia has the gun bans that they do considering everything on the freaking continent is trying to kill them	1.1588
RoBERTa	<i>Post:</i> Next friday, the United Kingdom will join Australia and become the second country in the world to introduce plain packaging to cigarettes , removing all brands , logos , and colors and replacing them with standardized dark green packages with graphic health warnings. <i>Comment:</i> I'm sure that 'll stop people from smoking .	0.9662
	<i>Post:</i> Gunman shoots himself dead in sydney factory siege. <i>Comment:</i> Never happened because Australia has gun control	1.0821
	<i>Post:</i> I really want some people to die but I don't want to go to prison so i'm not going to do anything illegal . <i>Comment:</i> The system works .	1.1359
Case 2	<i>Post:</i> Campaigns must work very hard to pretend that presidential policy choices matter a great deal relative to the actions of the federal reserve. <i>Comment:</i> Yes , it 's a real stretch to imagine the fiscal stance of the federal government matters	
SBERT	<i>Post:</i> Looks like the Clinton campaign just pressed the big red button. <i>Comment:</i> Hey guys , wrap it up , campaign is over , we just got the memo saying that we can't run against a woman , it 's sexist	1.1428
	<i>Post:</i> Ron and Rand Paul now is the time to pass audit the fed. <i>Comment:</i> Yeah , thats really the big problem facing America right now .	1.1187
	<i>Post:</i> Bernie Sanders announces bill to abolish private prisons , hints at marijuana policy platform. <i>Comment:</i> So really important issues then.	1.1589
RoBERTa	<i>Post:</i> Senator Bernie Sanders and several other democratic lawmakers are urging the Obama administration to hold a public hearing to determine whether the national institutes of health should override the patent on a prostate cancer drug in an effort to drive down prices. <i>Comment:</i> Lol like the corrupt corporate structure is going to do anything for the greater good .	1.1005
	<i>Post:</i> Chris Christie being voted as potential VP pick for Trump. <i>Comment:</i> Because American needs more traffic jams.	1.1573
	<i>Post:</i> Senators are trying to rush through a massive online sales tax hike. <i>Comment:</i> New York state residents don't see what all the commotion is about.	1.5731

Table 1: Case examples of input texts, semantic prototypes of SBERT and RoBERTa after projection, and the Euclidean distance score between the prototype embeddings and input sentence vector. **Note:** The highlighted words are only meant for reader guidance and were manually annotated by us to aid comparison.

	w			w/o		
	Acc.	Rec.	F1	Acc.	Rec.	F1
Twitter	98.3 (+1.1%)	98.6 (-0.1%)	98.4 (+0.9%)	97.2	98.7	97.5
Dialogues	83.6 (+1.7%)	83.7(+1.3%)	83.6 (+1.3%)	82.2	82.6	82.5
SARC	82.4 (+2.2%)	85.8 (+2.4%)	83.0 (+2.9%)	80.2	83.4	81.1

Table 4: Ablation study on the effect of incongruity loss, where ‘w’ denotes models trained with incongruity loss, and ‘w/o’ refers to models trained without it. The percentage reflects the variation in performance when training with incongruity loss compared to training without it.

Models	Accuracy	Recall	F1
CASCADE	77.0	84.0	77.0
SARC 2.0	75.0	-	76.0
ELMo-BiLSTM	72.0	-	-
ELMo-BiLSTM FULL	76.0	-	76.0
GRU-Attention	81.0	82.1	81.0
BiGRU	69.4	68.6	69.0
MULE	75.2	83.8	80.1
Our Model (SBERT)	80.1	86.2	82.2
Our Model (RoBERTa)	82.4	85.8	83.0

Table 5: Results on Reddit dataset SARC 2.0

Method	Twitter	Dialogues	SARC
E_{ct}	97.3	82.7	81.02
$E_{ct} + E_{st}$ (w/o L_{inco})	97.5	82.5	81.1
$E_{ct} + E_{st}$ (with L_{inco})	98.4	83.6	83.0

Table 6: Ablation study on different modules

difference between them, indicating that although, unlike SBERT, RoBERTa is not additionally trained to represent semantic similarity, it still works well with the prototype structure.

5.3 Ablation Study

To fully evaluate our model, we conducted an ablation study evaluating the influence of our **Incongruity Loss**. We experimented with and without the incongruity loss with the RoBERTa encoder on all three datasets and the result is shown in Table 4.

Upon analyzing the results for the Dialogues and SARC datasets, we observed performance improvements ranging from -0.1% to 2.9% when training model with incongruity loss, with SARC dataset showing a particularly notable increase of over 2% on recall, accuracy, and F1. In contrast, the Twitter dataset demonstrated minimal improvement. We hypothesize that this is due to the already high baseline performance on the Twitter dataset, which constrains the extent of further performance gains.

We further perform an ablation study to examine the effect of incorporating the sentiment encoder module in addition to the semantic encoder. The corresponding F1 scores are presented in Table 6. Without the incongruity loss, adding the sentiment encoder yields a comparable F1 score to using only the semantic encoder. However, when the incongruity loss is introduced, the model’s performance improves substantially, achieving up to a 2.9% increase in F1.

6 Conclusion

We proposed a novel approach that leverages state-of-the-art LM encoders and prototype-based networks to build an intrinsically interpretable model for sarcasm detection. Our approach achieved state-of-the-art performance on three public benchmark datasets. By representing prototypes with the closest training sentences, our method can explain sarcasm detection with sentence-level, human-readable explanations.

7 Limitation

Our data primarily comes from English-speaking populations from specific platforms, which may not be generalizable. Also, our work does not provide examples to show reasoning with sentiment prototypes. Future research could investigate generating explanations by analyzing the attention scores associated with both negative and positive sentiment prototypes.

8 Ethics Statement

The development and deployment of sarcasm detection models present several ethical considera-

tions. First, our study recognizes the potential biases inherent in training data, particularly those that stem from subjective interpretations of sarcasm across different linguistic, cultural, and social contexts. We used three different datasets from diverse platforms to ensure generalizability. In addition, while our prototype-based explainability framework enhances interpretability, we stress that sarcasm detection remains an inherently complex and nuanced task. We encourage the responsible use of our model to enhance human understanding rather than replace human judgment in sensitive contexts. Therefore, we do not recommend over-reliance on automated sarcasm detection in complex tasks and advocate for human oversight in critical decision-making processes. Moreover, we are committed to transparency in our research by making our model and code publicly available to foster reproducibility. We also adhere to ethical guidelines and ensure that all datasets used in this study comply with proper licensing.

References

- Ramya Akula and Ivan Garibay. 2021. [Explainable detection of sarcasm in social media](#). In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–39, Online. Association for Computational Linguistics.
- Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.
- Kun Bu, Yuanchao Liu, and Xiaolong Ju. 2024. Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, 283:111148.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. [Multi-modal sarcasm detection in Twitter with hierarchical fusion model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.
- Paula Carvalho, Luís Sarmiento, Mário J. Silva, and Eugénio de Oliveira. 2009. [Clues for detecting irony in user-generated contents: oh...!! it's "so easy";-\)](#). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, TSA '09, page 53–56, New York, NY, USA. Association for Computing Machinery.
- Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. [This](#)

- looks like that: Deep learning for interpretable image recognition. *Preprint*, arXiv:1806.10574.
- James M. Cupello and David J. Mishelevich. 1988. **Managing prototype knowledge/expert system projects**. *Commun. ACM*, 31(5):534–550.
- Jiakun Dai. 2024. A bert-based with fuzzy logic sentimental classifier for sarcasm detection. In *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, pages 1275–1280. IEEE.
- Anubrata Das, Chitrang Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. **ProtoTEX: Explaining model decisions with prototype tensors**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2986–2997, Dublin, Ireland. Association for Computational Linguistics.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. **Semi-supervised recognition of sarcasm in Twitter and Amazon**. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.
- Richard Fikes and Tom Kehler. 1985. **The role of frame-based representation in reasoning**. *Commun. ACM*, 28(9):904–920.
- V. Fomin, J. Anmol, S. Desroziere, J. Kriss, and A. Tejani. 2020. High-level library to help with training neural networks in pytorch. <https://github.com/pytorch/ignite>.
- Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169.
- Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. **Identifying sarcasm in Twitter: A closer look**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.
- John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.
- Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.
- Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. **Cascade: Contextual sarcasm detection in online discussion forums**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848. Association for Computational Linguistics.
- Nivin A Helal, Ahmed Hassan, Nagwa L Badr, and Yasmine M Afify. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports*, 14(1):15415.
- Dat Hong, Tong Wang, and Stephen Baek. 2024. Protorynet - interpretable text classification via prototype trajectories. *J. Mach. Learn. Res.*, 24(1).
- E Ilavarasan et al. 2020. A survey on sarcasm detection and challenges. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1234–1240. IEEE.
- Suzana Ilic, Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7.
- Adam Ivankay, Ivan Girardi, Chiara Marchiori, and Pascal Frossard. 2022. **Fooling explanations in text classifiers**. *Preprint*, arXiv:2206.03178.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.
- Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. **A large self-annotated corpus for sarcasm**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Been Kim, Cynthia Rudin, and Julie Shah. 2015. **The bayesian case model: A generative approach for case-based reasoning and prototype classification**. *Preprint*, arXiv:1503.01161.
- Akshi Kumar, Shubham Dikshit, and Victor Albuquerque. 2021a. **Explainable artificial intelligence for sarcasm detection in dialogues**. *Wireless Communications and Mobile Computing*, 2021:1–13.
- Akshi Kumar, Shubham Dikshit, and Victor Hugo C Albuquerque. 2021b. Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing*, 2021(1):2939334.
- Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium*

- on *Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.
- Liyuan Liu, Jennifer Lewis Priestley, Yiyun Zhou, Herman E Ray, and Meng Han. 2019a. A2text-net: A novel deep neural network for sarcasm detection. In *2019 IEEE first international conference on cognitive machine intelligence (CogMI)*, pages 118–126. IEEE.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). *CoRR*, abs/1705.07874.
- Dina Mardaoui and Damien Garreau. 2021. [An analysis of lime for text data](#). In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3493–3501. PMLR.
- Anuraj Mohan, Abhilash M Nair, Bhadra Jayakumar, and Sanjay Muraleedharan. 2023. Sarcasm detection using bidirectional encoder representations from transformers and graph convolutional networks. *Procedia Computer Science*, 218:93–102.
- Maryam Khanian Najafabadi, Thoon Zar Chi Ko, Saman Shojae Chaeikar, and Nasrin Shabani. 2024. A multi-level embedding framework for decoding sarcasm using context, emotion, and sentiment feature. *Electronics*, 13(22):4429.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2017. Creating and characterizing a diverse corpus of sarcasm in dialogue. *arXiv preprint arXiv:1709.05404*.
- Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2020. Modeling the incongruity between sentence snippets for sarcasm detection. In *ECAI 2020*, pages 2132–2139. IOS Press.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). *Preprint*, arXiv:1908.10084.
- Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should I trust you?": Explaining the predictions of any classifier](#). *CoRR*, abs/1602.04938.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. [Sarcasm as contrast between a positive sentiment and negative situation](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.
- Rishabh Misra and Prahal Arora. 2018. [Sarcasm detection using hybrid neural network](#).
- Jaroslav Švelch. 2015. Excuse my poor english: language management in english-language online discussion forums. *International journal of the sociology of language*, 2015(232):143–175.
- Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856*.
- Joseph Tepperman, David R. Traum, and Shrikanth S. Narayanan. 2006. ["yeah right": Sarcasm recognition for spoken dialogue systems](#). In *Interspeech*.
- Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. [Icwsn - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews](#). *Proceedings of the International AAAI Conference on Web and Social Media*.
- Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19089–19097.
- Oxana Vitman, Yevhen Kostiuk, Grigori Sidorov, and Alexander Gelbukh. 2023. Sarcasm detection framework using context, emotion and sentiment features. *Expert Systems with Applications*, 234:121068.
- Ximing Wen. 2024. Language model meets prototypes: Towards interpretable text classification models through prototypical networks. *arXiv preprint arXiv:2412.03761*.
- 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.
- Zhe Yu, Di Jin, Xiaobao Wang, Yawen Li, Longbiao Wang, and Jianwu Dang. 2023. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2423–2431.
- Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, pages 2449–2460.

Multimodal Claim Extraction for Fact-Checking

Joycelyn Teo^{1,*}, Rui Cao², Zhenyun Deng², Zifeng Ding²
Michael Sejr Schlichtkrull³, Andreas Vlachos²

¹Defence Science and Technology Agency, Singapore

²University of Cambridge, UK, ³Queen Mary University of London, UK

tlimeijo@dsta.gov.sg, {rc990,zd302,zd320,av308}@cam.ac.uk
m.schlichtkrull@qmul.ac.uk

Abstract

Automated Fact-Checking (AFC) relies on claim extraction as a first step, yet existing methods largely overlook the multimodal nature of today’s misinformation. Social media posts often combine short, informal text with images such as memes, screenshots, and photos, creating challenges that differ from both text-only claim extraction and well-studied multimodal tasks like image captioning or visual question answering. In this work, we present the first benchmark for multimodal claim extraction from social media, consisting of posts containing text and one or more images, annotated with gold-standard claims derived from real-world fact-checkers. We evaluate state-of-the-art multimodal LLMs (MLLMs) under a three-part evaluation framework (semantic alignment, faithfulness, and decontextualization) and find that baseline MLLMs struggle to model rhetorical intent and contextual cues. To address this, we introduce MICE, an intent-aware framework which shows improvements in intent-critical cases.¹

1 Introduction

The spread of misinformation on social media is increasingly multimodal, amplifying both its credibility and viral reach (Akhtar et al., 2023). Recent studies estimate that over one-third of debunked claims involve both text and images, highlighting the need for automated systems capable of reasoning across modalities (Zeng et al., 2024; Van Der Meer et al., 2025). This shift has motivated interest in multimodal fact-checking (Alam et al., 2025), where prior work has primarily focused on claim detection (Cheema et al., 2022) and verification (Yao et al., 2023; Braun et al., 2025). How-

*This work was partially done during Joycelyn’s research visit at Cambridge.

¹The MMCE dataset is available at <https://huggingface.co/datasets/joycelynt/MMCE>, and the code can be accessed at <https://github.com/jt9080/MMCE-Paper>.

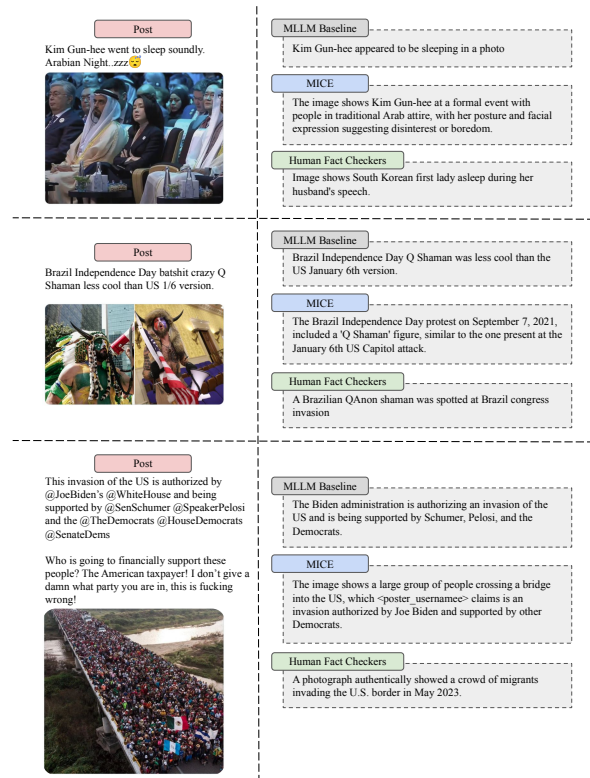


Figure 1: Examples of claim extraction from the MMCE dataset illustrating how baseline MLLMs (with in-context learning) underperform on intent-heavy multimodal posts. These cases highlight situations where structured reasoning (MICE) can help, but they are not representative of our dataset as a whole.

ever, multimodal claim extraction remains largely unexplored, despite being a crucial step in the fact-checking process (Hassan et al., 2015).

Extracting claims from multimodal content is challenging as systems must interpret informal language which often contains errors, integrate complementary or contradictory signals from images, and remove irrelevant context while preserving the factual core of a post (Wang et al., 2025). Figure 1 shows examples in which the Gemini 2.0 Flash baseline fails to capture these nuances, as it tends to extract the claim literally as the text states or the

image depicts, rather than inferring any underlying implication or context about the scenario.

To address this gap, in this paper, we make the following contributions:

1. We introduce MMCE (MultiModal Claim Extraction for fact checking), a new real-world benchmark dataset of 732 social media image-text posts paired with gold-standard claims, derived from fact-checking sources.
2. We introduce a scalable three-part evaluation framework for multimodal claim extraction, measuring semantic alignment with a manually annotated claim, faithfulness to the source, and decontextualization (i.e. whether a claim contains all necessary contextual information).
3. We provide a systematic study of MLLMs for claim extraction, showing that while they capture surface information, they often miss rhetorical intent and contextual cues.
4. We introduce the Multimodal Intent-aware Claim Extraction framework (MICE), which is a practical framework that combines vision-based semantic analysis with MLLMs to perform intent- and context-aware claim extraction. MICE breaks down the claim extraction process, extracting visual information and modeling underlying intent and context before performing claim extraction, so that implicit claims can be surfaced rather than literal descriptions. Our analysis shows that MICE can act as a mitigation strategy for MLLMs when handling high-nuance, intent-critical claims.

2 MMCE (Multimodal Claim Extraction)

Dataset We construct MMCE using post-claim pairs extracted from AVerImaTeC, a recently proposed dataset for real-world image-text claim verification based on data from fact-checking organizations (Cao et al., 2025). The dataset focuses on out-of-context (OOC) image-text claims because recent studies had observed that context-manipulated claims are the most common type of media-based misinformation (Dufour et al., 2024). Additionally, extraction for OOC claims is the most different from traditional vision-language tasks, such as image captioning, where the image and the text align. The task deviates from the pre-training objectives of MLLMs, presenting a significant challenge.

Each post-claim pair in the dataset consists of the original text and image(s) from the social media post, as well as the extracted claim from the corresponding fact-checking article. To do this, we first filter out data without links to the original social media posts, as well as data with claims that do not originate from social media sites. Next, we retrieve the source text directly from the linked social media posts. The final dataset contains 732 real-world post-claim pairs.²

For our experiment, we also curated a subset of 50 intent-critical claim examples that represent the failure modes that MICE is engineered to mitigate, to further analyze the utility of the intent-aware approach.

Evaluation In evaluating multimodal claim extraction, our goal is to go beyond simple similarity with a gold reference and instead capture multiple aspects of claim quality that matter for downstream verification in practice. Thus, we draw on the insights from previous work, which demonstrate that LLMs can serve as effective evaluators, offering scalable, consistent, and context-sensitive judgments (Liu et al., 2023; Fu et al., 2024; Muhamed, 2025; Es et al., 2024). Moreover, existing work on the evaluation of text-only claim extraction goes beyond evaluating similarity, but also considers faithfulness to and coverage of the original content, as well as whether the claim contains all necessary contextual information (Metropolitansky and Larson, 2025; Ullrich et al., 2025; Deng et al., 2024). These more comprehensive evaluation protocols provide a fuller picture of how well an extraction system supports downstream verification.

Motivated by these insights, we adopt an evaluation scheme consisting of both reference-based and reference-free metrics: (1) Reference-based evaluation uses an LLM to judge the degree of semantic alignment between the generated claim and the gold reference claim; (2) Entailment evaluation measures the extent to which the generated claim is faithful to the original social media post from which it was derived, while assuming that the post is true; (3) Decontextualization evaluation determines whether the extracted claim is interpretable as a stand-alone factual statement, without requiring additional context.

Our choice of model for evaluation is informed by previous studies, which have shown that Gemini

²The final dataset includes 618 train / 114 dev pairs from AVerImaTeC; the original test split was not publicly available.

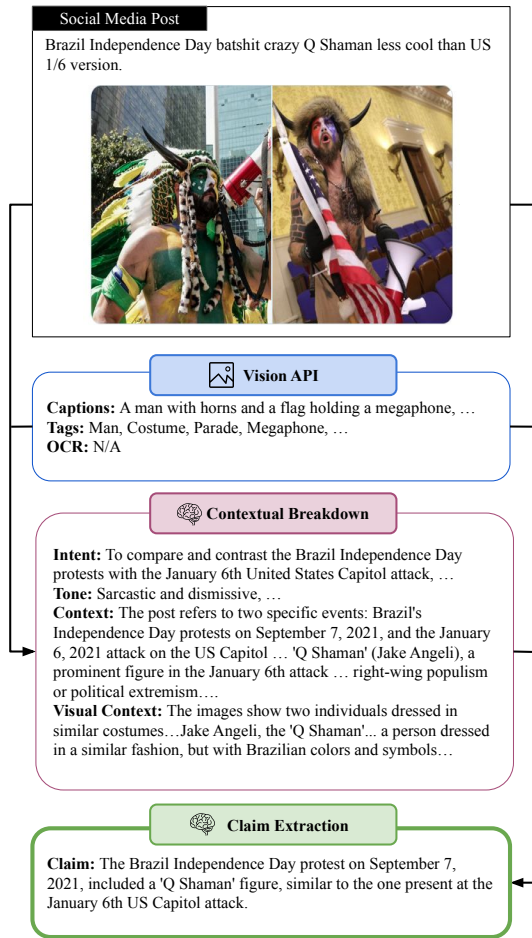


Figure 2: Overview of the MICE framework, which leverages visual understanding tools and MLLMs to reason across modalities.

models correlate relatively more with human assessments compared to other models (Akhtar et al., 2025; Gu et al., 2025). As such, we use Gemini 2.5 Flash Lite as a judge.

3 MICE (Multimodal Intent-aware Claim Extraction)

We introduce MICE, a novel approach for extracting factual claims from social media posts. While baseline MLLMs can produce fluent claim outputs, our analysis shows that they often interpret text and image content too literally without inferring underlying intent and miss contextual or rhetorical cues. To address these shortcomings, MICE first extracts visual semantics from the image of the social media post to improve visual understanding, which is an approach supported by previous work showing that combining vision APIs with LLMs is effective for multimodal reasoning (Khademi et al.,

2023). Next, MICE employs a contextual breakdown module to model intent, tone, and situational context, before integrating all signals to generate concise and accurate factual claims. The overall architecture is illustrated in Figure 2.

Vision API Given an image-text social media post, we first apply the off-the-shelf Azure Vision API to obtain structured descriptions of visual content. Specifically, we extract (1) Dense captions: natural language descriptions of salient objects, attributes, and relations, (2) OCR text: textual elements embedded in the image, and (3) Tags: high-level labels (e.g. protest, flooding).

Contextual Breakdown To capture intent and rhetorical context often present in social media posts, we prompt an MLLM for attributes which align with previous NLP-based frameworks for deception analysis, which models intention, tone, and contextual factors as core dimensions of deceptive content (Rani et al., 2025). To strike a balance between granular taxonomy and to avoid over-constraining the model outputs, we categorize the contextual breakdown into four crucial attributes: (1) Intention captures what the author tries to achieve with the post (e.g. joking, criticizing), which helps distinguish factual claims from rhetorical or performative speech, (2) Tone reflects the emotional or stylistic delivery (e.g. sarcastic, ironic, outraged), which signals whether a claim should be interpreted literally or as exaggeration, (3) Context accounts for the surrounding situational and cultural cues (e.g. ongoing events/conflicts, pop culture), which provide the background necessary to construct a sound claim, and (4) Image Context identifies specific visual elements (people, objects) shown in accompanying images, enabling the model to ground its understanding in both textual and visual information.

Claim Extraction Finally, the claim extractor MLLM receives the post text and image(s), the visual representation, and the contextual information. It is prompted to produce factual claims with In-Context Learning (ICL), in which similar demonstrations are retrieved from the training data using BM25 (Robertson and Zaragoza, 2009) under a leave-one-out strategy to prevent data leakage.

4 Experiments

Methods We evaluate a range of methods to establish performance across a diverse set of state-of-

Method	Model (used in the method)	Reference-Based (1–4) (↑)	Entailment (%)		Decontextualization (%)	
			Strict (↑)	Lenient (↑)	Strict (↑)	Lenient (↑)
MMCE dataset						
<i>LLM-based Evaluation</i>						
MLLM (text input only)	Gemini 2.0 Flash	2.80	80.0	85.9	96.9	99.7
	Qwen2.5 VL 32B Instruct	2.85	77.7	86.3	96.4	99.6
	GPT 4o Mini	2.83	80.6	88.0	97.4	99.7
MLLM	Gemini 2.0 Flash	3.11	75.5	85.4	97.9	99.9
	Qwen2.5 VL 32B Instruct	3.14	77.3	86.6	98.5	100.0
	GPT 4o Mini	3.15	74.4	85.2	97.8	99.9
MLLM with ICL	Gemini 2.0 Flash	3.21	70.5	82.6	98.4	100.0
	Qwen2.5 VL 32B Instruct	3.24	71.8	83.6	98.4	100.0
	GPT 4o Mini	3.22	69.9	82.4	98.1	100.0
MICE	Gemini 2.0 Flash	3.25	54.9	74.6	98.8	100.0
	Qwen2.5 VL 32B Instruct	3.15	52.7	78.9	98.1	100.0
	GPT 4o Mini	3.13	65.8	83.9	98.5	100.0
Intent-Critical Subset of MMCE						
<i>Human Evaluation</i>						
MLLM with ICL	Gemini 2.0 Flash	2.60	76.0	96.0	83.0	99.0
MICE	Gemini 2.0 Flash	3.31	81.0	97.0	92.0	99.0
<i>LLM-based Evaluation</i>						
MLLM with ICL	Gemini 2.0 Flash	2.08	68.0	82.0	94.0	100.0
MICE	Gemini 2.0 Flash	3.56	48.0	68.0	100.0	100.0

Table 1: Experimental results in the full MMCE dataset, as well as the intent-critical subset of it. The bold values represent the best performing method for each dataset, and for each evaluation method (human and LLM-based). Scores in the *Reference-Based* column are on a 1–4 scale (1 = lowest, 4 = highest). *Entailment* and *Decontextualization* are a 3-class categorical value, and is shown as strict (% fully entailed / fully decontextualized) and lenient (% fully or partially entailed / partially decontextualized).

the-art MLLMs: Google Gemini 2.0 Flash (DeepMind, 2024), Qwen2.5 VL 32B Instruct (Bai et al., 2025), and OpenAI GPT 4o Mini (OpenAI, 2024). We compare four methods: (1) MLLMs with text input only to evaluate the significance of visual information, (2) MLLMs with image-text inputs, (3) MLLMs with image-text inputs and ICL examples selected using BM25, and (4) the intent-aware framework, MICE.

Main results Table 1 summarizes the results across the three evaluation metrics. In reference-based scoring, we find that MLLMs with image-text inputs consistently outperform instances with text-only inputs, underscoring the importance of incorporating visual signals for claim extraction in noisy social media posts. The further gains from MLLM with in-context learning (ICL) suggest that multimodal intent is best captured when models are guided by examples. In the subset of intent-critical cases, both human and LLM-based evaluations agree that the MICE framework improves reference-based scores, though its utility is not reflected across the entire dataset because of other challenges. E.g, when interpreting an image about the packaging of COVID-related medication, MLLMs and MICE fail to spot the insinuation that COVID is the same as influenza, which is indica-

tive of the challenges MMCE poses to MLLMs more broadly (see error analysis in Appendix H).

MICE also improves decontextualization scores for Gemini 2.0 flash and GPT 4o Mini, producing claims that are more stand-alone and interpretable without the original post. This is crucial for downstream fact-checking pipelines, where claims are often checked in isolation. Our experiments also reveal that improvements in reference alignment and decontextualization introduce a trade-off with entailment. This is likely because explicitly modeling intent and context encourages the model to abstract away from the literal text and image.

Human Alignment with Automated Evaluations

To validate the LLM-based evaluation metrics and verify the utility of the MICE approach in handling intent-critical claims, we had 4 expert annotators independently score the claims extracted from the intent-critical subset (details in Appendix G), and calculated the agreement between the human evaluators and the LLM scorer. The agreement is measured using Krippendorff’s α (Krippendorff, 2013) and Spearman’s ρ (Spearman, 2010). Results (Appendix B1) show that reference-based and decontextualization alignment achieved moderate correlation with humans ($\alpha = 0.59$ and $\alpha = 0.54$ respectively), but entailment score achieved a lower

correlation ($\alpha = 0.07$). This suggests that judging entailment involves a deeper inferential reasoning of the post compared to the other metrics, leading to a larger divergence from human judgment.

Experiment on Temporal Leakage To further examine the extent to which MLLMs rely on pre-training knowledge when generating claims, we compared Gemini 2.0 flash (training cutoff: June 2024) (DeepMind, 2024) with Gemini 2.5 flash (training cutoff: January 2025) (DeepMind, 2025). Experiments on 50 randomly sampled image–text claims posted between July and December 2024 (Appendix C) showed that Gemini 2.5 flash did not have a significant improvement in scoring metrics (Table C2). On closer analysis of the improved claims, we also found no clear evidence that Gemini 2.5 flash relied on additional pre-training knowledge unavailable to Gemini 2.0 flash.

5 Conclusion

In this work, we introduce MMCE, the first dataset for multimodal claim extraction³ from social media, and established a three-part evaluation framework that measures semantic alignment, faithfulness, and decontextualization. Our analysis demonstrates that while baseline MLLMs can extract literal claims, they often miss the rhetorical intent and contextual framing crucial for understanding social media content. To address this gap, we propose MICE, an intent-aware framework that improves performance in these nuanced, intent-critical cases. Overall, this research contributes a new benchmark and an intent-aware framework, supporting the ongoing development of more effective automated tools for fact-checking.

Acknowledgments

This research was supported by the Alan Turing Institute and DSO National Laboratories in Singapore Partnership (ref DCfP2\100063). Zhenyun Deng, Zifeng Ding and Andreas Vlachos were further supported by the ERC grant AVeriTeC (GA 865958). Andreas Vlachos is also supported by the DARPA program SciFy. Michael Schlichtkrull is supported by the Engineering and Physical Sciences Research Council (grant number EP/Y009800/1), through funding from Responsible AI UK (KP0016).

³During the reviewing process we became aware of concurrent work by (Geng et al., 2026) that also proposes a dataset for this task.

Limitations

We acknowledge that claim extraction can inherently be a subjective task, whereby extracted claims with different semantic meaning can still be considered valid and check-worthy. Our rationale is to model the claim extraction process as closely as possible to that of professional fact checkers, which motivated the decision to curate a dataset from real-world fact-checking articles.

Moreover, the dataset we curated explicitly focuses on out-of-context images, where we deemed it the most urgent to focus our efforts on these types of claim. However, this excludes image-text claim types that could also benefit from the MICE framework, such as memes. In future work, we propose to extend the experiments to other types of image-text claims for diversity.

Lastly, due to resource constraints, we performed human evaluation on 100 generated claims. Although this sample size provided a representative assessment of the claims’ quality, it limits a more comprehensive statistical analysis of the framework’s performance variability across different claim domains and styles. A larger scale human evaluation would be beneficial for future iterations of this work to improve statistical reliability and reproducibility.

Ethical Considerations

We rely on fact-checks from real-world fact-checkers to develop and evaluate our models. Nevertheless, as any dataset, it is possible that it contains biases which influenced the development of our approach. Given the societal importance of fact-checking, we advise that any automated system is employed with human oversight to ensure that the fact-checkers fact-check appropriate claims.

References

- Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. [Multimodal automated fact-checking: A survey](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.
- Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2025. [Ev2r: Evaluating evidence retrieval in automated fact-checking](#).
- Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna

- Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venkatesh V. 2025. [The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and & retrieval](#). In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, page 467–478, Berlin, Heidelberg. Springer-Verlag.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. [Qwen2.5-vl technical report](#).
- Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. [DEFAME: Dynamic evidence-based FAct-checking with multimodal experts](#). In *Forty-second International Conference on Machine Learning*.
- Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. [Averimatec: A dataset for automatic verification of image-text claims with evidence from the web](#).
- Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. [MM-claims: A dataset for multimodal claim detection in social media](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.
- DeepMind. 2024. [Gemini 2.0 flash model card](https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf). <https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf>.
- DeepMind. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#).
- Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. [Document-level claim extraction and de-contextualisation for fact-checking](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.
- Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Duffield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. [Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild](#).
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. [GPTScore: Evaluate as you desire](#). 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 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2026. [M4fc: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset](#).
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#).
- Naemul Hassan, Chengkai Li, and Mark Tremayne. 2015. [Detecting check-worthy factual claims in presidential debates](#). In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, page 1835–1838, New York, NY, USA. Association for Computing Machinery.
- Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. [MM-reasoner: A multimodal knowledge-aware framework for knowledge-based visual question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, Singapore. Association for Computational Linguistics.
- K. Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Dasha Metropolitansky and Jonathan Larson. 2025. [Towards effective extraction and evaluation of factual claims](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6996–7045, Vienna, Austria. Association for Computational Linguistics.
- Aashiq Muhamed. 2025. [Ccrs: A zero-shot llm-as-a-judge framework for comprehensive rag evaluation](#).
- OpenAI. 2024. [Gpt-4o system card](#).
- Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2025. [SEPSIS: I can catch your lies – a new paradigm for deception detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for*

Computational Linguistics (Volume 4: Student Research Workshop), pages 97–128, Vienna, Austria. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.

C Spearman. 2010. [The proof and measurement of association between two things](#). *International Journal of Epidemiology*, 39(5):1137–1150.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2025. [Claim extraction for fact-checking: Data, models, and automated metrics](#).

Michiel Van Der Meer, Pavel Korshunov, Sébastien Marcel, and Lonneke Van Der Plas. 2025. [HintsOfTruth: A multimodal checkworthiness detection dataset with real and synthetic claims](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31274–31291, Vienna, Austria. Association for Computational Linguistics.

Haoran Wang, Aman Rangapur, Xiong Xiao Xu, Yueqing Liang, Haroon Gharwi, Carl Yang, and Kai Shu. 2025. [Piecing it all together: Verifying multi-hop multimodal claims](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7453–7469, Abu Dhabi, UAE. Association for Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. [End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. 2024. [Multimodal misinformation detection by learning from synthetic data with multimodal LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10467–10484, Miami, Florida, USA. Association for Computational Linguistics.

A MMCE Dataset

The breakdown of the social media sites of the data from MMCE is shown in table A1.

Social Media Site	Number of Claims
X (Formally Twitter)	360
Facebook	319
Instagram	32
Reddit	13
Telegram	2
Weibo	2
Band (Naver)	1
Flickr	1
Truth Social	1

Table A1: Breakdown of social media sites

Although we converted the original post text to English for our experiments, we also retained the original language of the social media posts for future experiments. Table A2 shows the breakdown of the language of the original posts in MMCE.

Original Post Language	Number of Claims
English	551
Hindi	100
Korean	14
Urdu	10
Chinese	8
Bengali	7
Thai	6
Sinhala	6
Khmer	7
Burmese	6
Filipino	4
Amharic	4
Tamil	3
Indonesian	2
Punjabi	2
Telugu	2
Kannada	1
French	1
Pashto	1
Swahili	1

Table A2: Breakdown of the language of the original posts in MMCE. Some posts use more than one language. In those cases, we count it under all the languages used.

B Human Alignment Analysis

Table B1 reports detailed agreement statistics between LLM-based evaluations and human annotators, as well as human–human agreement.

Due to resource constraints, one annotator labeled all 100 examples; the other three split the set (approximately 33–34 claims each).

C Temporal Leakage Analysis

Data for the temporal leakage analysis was drawn from four fact-checking sources via the Google Fact Check API. The breakdown of the source URLs of the data used in the temporal leakage analysis is shown in table C1.

Fact Checking Site	Number of Claims
factcheck.afp.com	30
leadstories.com	13
politifact.com	5
fullfact.org	2

Table C1: Breakdown of fact checking sites used in the temporal leakage experiments

The results of the temporal leakage experiment are shown in table C2.

D Model Implementation and Resources

For few-shot settings, we set the number of shots to be 5 to strike a balance between providing sufficient demonstrations, and to prevent exceeding the context length limit.

For experiments with Qwen2.5 VL 32B Instruct, we faced a 30-image limit per input. In the handful of cases where the limit is exceeded (due to the images from the few shot demonstrations), we limit the number of few-shot inputs to fit the image constraints.

All experiments were conducted using API access to commercial and open-source multimodal LLMs via the OpenRouter⁴ platform. We report results for Gemini 2.0 Flash, Gemini 2.5 Flash, Qwen2.5 VL 32B Instruct, and GPT-4o Mini. Since these models are hosted services, we do not control the underlying hardware; however, we record the model versions and set temperature to zero to facilitate reproducibility.

⁴<https://openrouter.ai/>

Method	Model	Reference-Based	Entailment	Decontextualization
LLM-Human	Krippendorff’s α	0.59	0.07	0.54
	Spearman ρ	0.56	0.11	0.41
	Agreement (%)	46.5	50.0	89.5
Human-Human	Krippendorff’s α	0.67	0.14	0.80
	Spearman ρ	0.70	0.30	0.81
	Agreement (%)	55.0	69.0	94.0

Table B1: Human–LLM and human–human alignment statistics across the three evaluation metrics used in the experiment.

Model	Reference-Based (1–4) (↑)	Entailment (%)		Decontextualization (%)	
		Strict (↑)	Lenient (↑)	Strict (↑)	Lenient (↑)
Gemini 2.0 Flash	3.08	88.0	90.0	94.0	98.0
Gemini 2.5 Flash	3.10	86.0	86.0	84.0	90.0

Table C2: Results of the temporal leakage experiment, on 50 post-claim pairs that surfaced after the training cut-off for Gemini 2.0 Flash, but before the training cut-off for Gemini 2.5 Flash. Scores in the *Reference-Based* column are on a 1–4 scale (1 = lowest, 4 = highest). *Entailment* and *Decontextualization* are shown as strict (% fully entailed / fully decontextualized) and lenient (% fully or partially entailed / partially decontextualized).

E Licensing Information

If accepted, we will publicly release and maintain the dataset and baseline code, which will be licensed under the CC BY-NC 4.0 license.

F Prompts

Here we provide the prompts used in our experiments. This includes the prompts used in the LLM-based evaluation, as well as the prompts used in the baseline approaches and the MICE framework.

For the MICE framework, the prompt used in the claim extraction stage is a concatenation of the image-text input prompt (F.5), the output of the Vision API, and the output of the contextual extraction (F.6).

F.1 Prompt for reference-based evaluation

```
# MISSION
You will get two claims below, a
generated claim and a reference claim.
Your task is to perform a comprehensive
similarity assessment between the
generated claim and the reference
claim, in the context of fact checking.
Provide a similarity score from 1 to 4.

# CRITICAL INSTRUCTIONS:
- Be CONSISTENT in your scoring.
Similar claims should always receive
similar scores.
- Focus on the core implied
meaning/content of the sentence,
wording differences should be
acceptable.
```

- Extra details that don’t contradict the core claim is beneficial. Additional specific details (e.g. names, dates, locations, numbers) on the core factual assertion is a good thing. Do not penalize for verbosity and specificity.

Scoring Guidelines:

- 1: Completely different, no overlap in themes, topics, or entities mentioned.
- 2: Minimal similarity in themes, topics, or entities mentioned, sentences have different meanings. The core factual statement to be fact checked is different.
- 3: Partial alignment in message conveyed, with significant differences that could potentially affect downstream fact checking.
- 4: Strong conceptual similarity with minor variations or near-identical meaning.

```
# RESPONSE FORMAT
Return the response in the following
JSON format:
““json
{
  "score": 3,
  "reasoning": "Brief explanation of
why this score was given"
}
““

# INPUT
Generated Claim: <generated claim>
Reference Claim: <reference claim>
```

F.2 Prompt to evaluate entailment

```
# MISSION
You will be given a social media post via the text and image of the post, we well as a candidate claim extracted from the post. Your task is to assess whether the claim is fully faithful to and entailed by the combined content of the image and text. This means that assuming the social media post is true, the extracted claim must also be true.
```

Do NOT check if the claim is true in reality, only whether it is faithful to the content of the image and text of the post.

Ignore whether the correct factual content had been extracted, focus on whether the extracted sentence is faithful (i.e. no hallucinations).

```
# EVALUATION CRITERIA
Classify the claim into one of three categories:
- **entailed**:: The claim is fully aligned with the post content without any contradictions, hallucinations or unsupported additions.
- **partially_entailed**:: The claim is partially aligned with the post content but contains minor variations or additional context not stated or implied in the post.
- **not_entailed**:: The claim contains significant misaligned inferences, exaggerations beyond what's stated, major contradictions, hallucinations, or completely misaligns with the post content.
```

```
# INPUT
Generated Claim: <generated claim>
Text: <social media post text>
Image(s): <social media post image(s)>
```

F.3 Prompt to evaluate decontextualization

```
# MISSION
You will be given a candidate claim that was extracted from a social media post. Your task is to assess whether this claim is understandable in isolation, without access to the original post or any external context. A decontextualized claim should be fully interpretable and self-contained to an average reader, who has no knowledge of the post.
```

```
# CRITICAL INSTRUCTIONS:
- Focus only on clarity and completeness of meaning. Do not check whether the claim is factually true or faithful to the post, only whether the claim can stand alone and be understood independently.
```

```
# Scoring Guidelines:
- **fully_decontextualized**:: Understandable in isolation. The claim is completely self-contained, unambiguous, and requires no edits to be understood on its own. (Example: The mayor of NYC announced a new recycling program on June 1, 2024.)
- **partially_decontextualized**:: The claim is mostly clear and contains some context, but has gaps, vague references or unresolved pronouns. The claim could benefit from some edits. (Example: Vaccination rates rose after that. -> could be rewritten to -> Vaccination in the UK rates rose after the 2023 campaign.)
- **not_decontextualized**:: Not understandable in isolation. The claim cannot be interpreted on its own; key entities, referents, or context are missing. Major rewriting is needed. (Example: He did something yesterday.)
```

```
# INPUT
Generated Claim: <generated claim>
```

F.4 Prompt for text-only input for claim extraction

```
# MISSION
You are an expert fact-checking analyst specializing in social media content verification. Your primary objective is to precisely extract and articulate the core factual claim(s) from the given text.
```

```
# CONTEXT ANALYSIS
Before extracting the claim, perform a comprehensive context analysis:
- Examine the full text carefully
- Consider the platform type and its typical communication style
- Identify potential implicit or explicit claims
```

```
# CLAIM EXTRACTION METHODOLOGY
1. Identify Potential Claims:
- Look for definitive statements
- Detect implied assertions
- Recognize potentially misleading or exaggerated claims
2. Claim Criteria:
- Clarity: Can the claim stand alone and be understood without the original context?
- Specificity: Does the claim capture the most significant factual assertion?
- Verifiability: Does the claim provide enough detail to enable fact-checking?
3. Claim Refinement Process:
- Remove subjective language
- Distill the core factual assertion
```

- Ensure the claim is neutral and objective

CLAIM SELECTION STRATEGY

- Always try to extract just one main claim first
- If the text contains one main factual assertion, extract only that claim
- If multiple statements can be combined into one coherent claim, do so

Multiple claims should only be used when:

- The text contains completely separate factual statements about different topics that cannot be combined
- Each claim is independently verifiable and fact-checkable
- Combining them would create a confusing or overly complex claim

ADDITIONAL CONSIDERATIONS

- If multiple potential claims exist, first try to identify the most significant or impactful one
- If the claim is ambiguous, provide the most reasonable interpretation based on context
- Avoid introducing personal bias or speculation
- Always prioritize extracting a single, comprehensive claim over multiple separate claims

RESPONSE FORMAT

Return the response in the following JSON format:

```

{
  "claims": ["main claim"]
}

```

INPUT

Extract the claim(s) from the following text: <social media post text>

F.5 Prompt for image-text input for claim extraction

MISSION

You are an expert fact-checking analyst specializing in social media content verification. Your primary objective is to precisely extract and articulate the core factual claim(s) from the given text and accompanying image.

CONTEXT ANALYSIS

Before extracting the claim, perform a comprehensive context analysis:

- Examine the full input text and image carefully. Consider how the image contributes to the messaging
- Consider the platform type and its typical communication style
- Identify potential implicit or explicit claims

CLAIM EXTRACTION METHODOLOGY

1. Identify Potential Claims:
 - Look for definitive statements
 - Detect implied assertions
 - Recognize potentially misleading or exaggerated claims

2. Claim Criteria:

- Clarity: Can the claim stand alone and be understood without the original context?
- Specificity: Does the claim capture the most significant factual assertion?
- Verifiability: Does the claim provide enough detail to enable fact-checking?

3. Claim Refinement Process:

- Remove subjective language
- Distill the core factual assertion
- Ensure the claim is neutral and objective
- Consider whether the image alters, reinforces, or creates the perceived claim

CLAIM SELECTION STRATEGY

- Always try to extract just one main claim first
- If the text contains one main factual assertion, extract only that claim
- If multiple statements can be combined into one coherent claim, do so

Multiple claims should only be used when:

- The text contains completely separate factual statements about different topics that cannot be combined
- Each claim is independently verifiable and fact-checkable
- Combining them would create a confusing or overly complex claim
- The image introduces additional factual assertions that cannot be combined with the text claims

ADDITIONAL CONSIDERATIONS

- If multiple potential claims exist, first try to identify the most significant or impactful one
- If the claim is ambiguous, provide the most reasonable interpretation based on context
- Avoid introducing personal bias or speculation
- Always prioritize extracting a single, comprehensive claim over multiple separate claims
- Consider whether the image alters, reinforces, or creates the perceived claim

RESPONSE FORMAT

Return the response in the following JSON format:

```

{
  "claims": ["main claim"]
}

```

```

}
```

INPUT
Extract the claim(s) from the following
text: <social media post text>

```

## F.6 Prompt to extract contextual insights

```

MISSION
Analyze this social media post and
provide contextual insights to help
identify the main factual claim.

CONTEXTUAL ANALYSIS
Focus on these key insights:
1. INTENT: What's the main purpose of
the post?
(inform/persuade/entertain/satire/etc.)
2. TONE: What's the emotional tone of
the post?
(serious/humorous/sarcastic/anger/etc.)
3. CONTEXT: What real-world
events/issues does this relate to?
Include specific details like dates,
locations, people, organizations.
4. VISUAL_CONTEXT: What specific
people, objects, locations, or events
are shown in the image that provide
context for the claim?

RESPONSE FORMAT
Return your analysis as a JSON object
with the following structure:
```json
{
  "intent": "description of the
poster's main purpose",
  "tone": "description of the
emotional tone",
  "context": "brief context with
specific details about real-world
events/issues",
  "visual_context": "description of
what's shown in the image that
provides context"
}
```

```

## G Annotation Guidelines

The annotators involved in the human evaluation are also co-authors but were unaware of which system generated which claim. Here, we provide the annotation guidelines provided to those annotators, alongside an excel sheet containing the intent-critical subset of the 100 post-claim pairs, as well as the images relevant to the posts.

### G.1 Instructions given to the annotators

There are 100 generated claims in the excel file, and for each claim, the original social media post (text+image(s)) and the reference claim, are provided. Fill in the green

columns (E and F): reference-based score and entailment score.

Here are the guidelines for the scoring:

Reference-based  
[Look at columns C and D only]  
Given 2 claims, a generated claim and a reference claim, perform a comprehensive similarity assessment between the generated claim and the reference claim. Provide a similarity score from 1 to 4.

Focus on the core meaning and factual content, not minor wording differences. Ignore extra details that don't change the core factual claim, no need to penalize for that.

Scoring Guideline (1-4 scale)  
1: Completely different, no overlap in themes, topics, or entities mentioned.  
2: Minimal similarity in themes, topics, or entities mentioned, sentences have different meanings. The core factual statement to be fact checked is different.  
3: Partial alignment in message conveyed, with significant differences that could potentially affect downstream fact checking.  
4: Strong conceptual similarity with minor variations or near-identical meaning.

Entailment  
[Look at columns A, B and D only]  
Given a social media post (text and image(s)), as well as a claim extracted from the post, assess whether the claim is fully faithful to and entailed by the combined content of the image and text. This means that assuming the social media post is true, the extracted claim must also be true.

Ignore whether the correct factual content had been extracted, focus on whether the extracted sentence is faithful to the post (i.e. no hallucinations).

Scoring Guideline (3-class classification)  
1. entailed: The claim is fully aligned with the post content without any contradictions, hallucinations or unsupported additions.  
2. partially\_entailed: The claim is partially aligned with the post content but contains minor variations or additional context not stated or implied in the post.  
3. not\_entailed: The claim contains significant misaligned inferences, exaggerations beyond what's stated, major contradictions, hallucinations, or completely misaligns with the post content.

## Decontextualization

[Look at column D only]

You will be given a candidate claim that was extracted from a social media post. Your task is to assess whether this claim is understandable in isolation, without access to the original post or any external context. A decontextualized claim should be fully interpretable and self-contained to an average reader, who has no knowledge of the post.

Scoring Guideline (3-class classification)

1. `fully_decontextualized`:

Understandable in isolation. The claim is completely self-contained, unambiguous, and requires no edits to be understood on its own. (Example: The mayor of NYC announced a new recycling program on June 1, 2024.)

2. `partially_decontextualized`: The claim is mostly clear and contains some context, but has gaps, vague references or unresolved pronouns. The claim could benefit from some edits. (Example: Vaccination rates rose after that. -> could be rewritten to -> Vaccination in the UK rates rose after the 2023 campaign.)

3. `not_decontextualized`: Not understandable in isolation. The claim cannot be interpreted on its own; key entities, referents, or context are missing. Major rewriting is needed. (Example: He did something yesterday.)

## H Error Analysis

To surface the key challenges faced in image-text social media claim extraction, we select erroneous samples (scored 1 or 2 for reference-based scoring) for analysis, shown in Table H1. We also select instances where the baseline MLLM with ICL outperforms MICE in reference-based scoring, shown in Table H3, in order to highlight some potential drawbacks of MICE. Note that in cases where the system outputs more than one claim, we select the highest scoring claim based on the reference-based evaluation.

**Failure Cases for both baseline MLLMs and MICE**

| Social Media Post                                                                                                                                                                                                                                                                                                                                                  | Claims                                                                                                                                                                                                                                                                                                                                                                                                                | Analysis                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>India's clothing minister - Prime Minister of Italy. Very cultured.</p>                                                                                                                                                                                                        | <p><b>Reference:</b> Italy's Prime Minister Giorgia Meloni refuses to shake hands with PM Narendra Modi.</p> <p><b>MLLM+ICL:</b> Image shows Narendra Modi and the Prime Minister of Italy. [2]</p> <p><b>MICE:</b> The image shows Narendra Modi opening a car door for Italian Prime Minister Giorgia Meloni. [1]</p>                                                                                               | <p>The LLM-based approaches fail to capture the nuance of the gestures and context in the image. The phrase "Very cultured" adds a subjective commentary on behavior rather than merely describing the scene. Though MICE attempts to interpret the actions of the people in the image, it overlooks intended meaning of the post. This reflects a broader limitation in MLLMs' ability to integrate non-literal cues from language with the contextual visual semantics of human gestures.</p>                                  |
| <p>Krabi people are enlightened!!! Move Forward Party is no longer a party that will overthrow the monarchy. #MoveForward #themalaengtad</p>  <p>The text overlay in the image is: "Krabi people's eyes are open! If you don't take the party forward, the party will fall."</p> | <p><b>Reference:</b> The Move Forward Party's political rally in Southern Krabi province, Thailand had a very small crowd attendance.</p> <p><b>MLLM+ICL:</b> The Move Forward Party is no longer a party that will overthrow the monarchy. [1]</p> <p><b>MICE:</b> On April 23, 2023, the Move Forward Party held an event in Krabi, and the party is no longer viewed as wanting to overthrow the monarchy. [2]</p> | <p>The LLM-based models misinterpret the underlying sarcasm and political sentiment embedded in the post. The Thai text, combined with the image of a sparse crowd, is meant to mock the Move Forward Party's rally attendance. However, both MICE and MLLM+ICL treat the post as a straightforward political statement, failing to synthesize the ironic signals between text and visuals to understand the true rhetorical intent.</p>                                                                                         |
| <p>Packaging update 🤔</p>                                                                                                                                                                                                                                                       | <p><b>Reference:</b> COVID19 is the same as influenza.</p> <p><b>MLLM+ICL:</b> Fanttest sells a COVID-19/Influenza A&amp;B Antigen Test Kit. [1]</p> <p><b>MICE:</b> The packaging for the Fanttest COVID-19/Influenza A&amp;B Antigen Test Kit has been updated. [1]</p>                                                                                                                                             | <p>The LLM-based approaches fail to identify the implicit misinformation within the post. Both MICE and MLLM+ICL interpret the post literally without recognizing the underlying insinuation that conflates two distinct diseases. The emoji also serves as a rhetorical device signaling skepticism and insinuation, which was not picked up by the systems. Consequently, their outputs remain factually descriptive but semantically shallow, demonstrating a persistent limitation in detecting implicit misinformation.</p> |

Table H1: Analysis on instances where both baseline MLLMs and the MICE framework yields poor results, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score (Part 1).

**Failure Cases for both baseline MLLMs and MICE**

**Social Media Post**

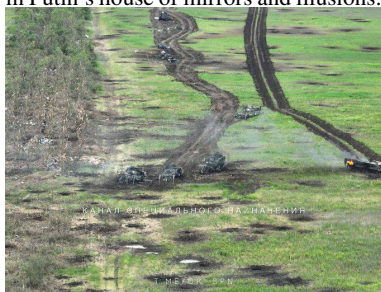
The name of this Chinese firm is PMC Projects and its owner is 'Chang Chien-ting', better known as Morris Chang. Surprisingly, Morris Chang is the son of Chinese citizen Chang Chung-ling, who has been a director of Adani's companies and is a business partner of Adani's brother Vinod Adani.

2. Particulars of the holder of the significant beneficial interest:

|                                                                     |                                                                                              |
|---------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| Name of the Significant Beneficial Owner (Given name and last Name) | Chang Chien-Ting                                                                             |
| Address and Email Id                                                | 11F-1, No.290, Xianfu Road, Taoyuan Dist., Taoyuan city, Taiwan.<br>pmcinfra.maust@gmail.com |
| Date of Birth/Age                                                   | 14 November 1976 / 41                                                                        |
| Father's/<br>Spouse's name                                          | Mr. Chang Chung Ling                                                                         |
| Occupation                                                          | Business                                                                                     |
| Nationality                                                         | Republic of China                                                                            |
| Passport No. (in case of foreign national)                          | 316345230                                                                                    |



This image is a fraud, created via AI. Supporters of Russia, you go right ahead and believe this fraud. I have no objections at all. Supporters of Ukraine, Ukraine has taken at least five villages in the last day. The Russian army in the last two days has lost immense amounts of men, far more than the daily average for this war. In the last village taken alone, fifty occupiers were slain, and four captured. Russian supporters, you won't believe me, and I absolutely do not need you to. What you believe or don't believe is immaterial. As we freed my family in Kherson, as we freed Kharkiv, so also now we free more villages from the horrible dystopia of the 'multi-polar paradise'. But you please do keep making your AI images and living in Putin's house of mirrors and illusions.



**Claims**

**Reference:** Chang Chien-Ting, an associate of the Adani Group, is a Chinese national.

**MLLM+ICL:** Morris Chang, owner of PMC Projects, is the son of Chang Chung-ling, who is a director of Adani's companies and a business partner of Adani's brother, Vinod Adani. [1]

**MICE:** Morris Chang, owner of PMC Projects, is the son of Chang Chung-ling, who is a director of Adani's companies and a business partner of Vinod Adani. [1]

**Analysis**

The LLM-based approaches fail to detect the deceptive framing and the nationality misattribution at the core of this post. This is likely because the post relies on background knowledge of India–China relations and domestic narratives around economic nationalism. The post falsely claims that the Adani Group's (An Indian Multinational Conglomerate) associate is linked to a "Chinese" businessman, exploiting ongoing public suspicion toward Chinese influence in Indian infrastructure projects. Both MICE and MLLM+ICL extract the literal factual components (names, companies, and family relations) but fail to recognize the misleading intent: the deliberate conflation of Taiwan ("Republic of China") with mainland China ("People's Republic of China") to frame the relationship as evidence of Chinese involvement. This illustrates a broader limitation of automated claim extraction, which in some cases would need context from real-world and real-time socio-political context.

**Reference:** Image Showing Ukrainian Losses Is Generated By Artificial Intelligence.

**MLLM+ICL:** Ukraine has taken at least five villages in the last day, and in the last village taken alone, fifty Russian soldiers were slain, and four captured. [1]

**MICE:** Ukraine has taken at least five villages in the last day, and the Russian army has suffered immense losses in the last two days. [1]

Both models extract a secondary claim from the lengthy post text but miss the primary, overarching claim that "This image is a fraud, created via AI" (the image is verified as legitimate). The models appear to latch onto the more concrete, detailed statements about recent military events later in the paragraph. This limitation likely arises from the claim extraction prompt, which directs the system to identify minimal or singular claims, a strategy that proves inadequate for posts containing longer, more complex paragraphs.

Table H2: Analysis on instances where both baseline MLLMs and the MICE framework yields poor results, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score (Part 2, continued from H1)).

Cases where baseline MLLMs outperforms MICE


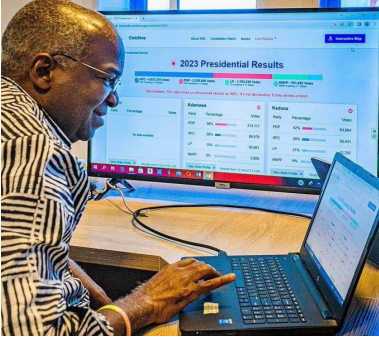

| Social Media Post                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Claims                                                                                                                                                                                                                                        | Analysis                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Tunisian Club African CEO being interviewed by a journalist about their game with Yanga SC</p>                                                                                                                                                                                                                                                                                                                                                                                 | <p><b>Reference:</b> Image shows the CEO of Club Africain of Tunisia being interviewed after their match with Tanzania's Yanga SC.</p>                                                                                                        | <p>This instance illustrates the potential drawback of applying MICE's complex reasoning to straightforward, literal posts. The MICE framework fails because of its over-emphasis on visual analysis where none is needed. The post's claim is fully contained in the text, but MICE's methodology compels it to integrate visual features. This shows that for simple, text-dominant claims, MICE's intent- and context-aware modules can introduce noise and errors rather than adding value.</p> |
|                                                                                                                                                                                                                                                                                                                                                                                                  | <p><b>MLLM+ICL:</b> The man in the photo is the CEO of Tunisian Club African being interviewed about their game with Yanga SC. [4]</p>                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <p>Someone please ask Fashola how he got access to this server that has the election result data. INEC and APC rigged the election even before it started. They should come out and defend this.</p>                                                                                                                                                                                                                                                                              | <p><b>MICE:</b> The CEO of Tunisian Club African is depicted as a man wearing a turban and traditional clothing, sitting at a desk with a laptop and an AK-47 assault rifle. [1]</p>                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|                                                                                                                                                                                                                                                                                                                                                                                                 | <p><b>Reference:</b> Nigerian housing minister Babatunde Fashola illegally accessed election data on a server prior to the official release of results.</p>                                                                                   | <p>MICE mentioned "result data from Civichive," a detail not mentioned in the text and likely derived from misinterpreting the website screenshot in the image. This shows that the Vision API or contextual breakdown module likely misinterpreted the UI elements in the image, whereas the baseline's more direct approach remained more faithful to the source text.</p>                                                                                                                        |
| <p>U.S. District Judge Aileen Cannon (on the left side in photo) has been initially assigned to oversee the case...could preside over the trial as well. Cannon, appointed by Trump in 2019, made decisions that favored him in legal skirmishes during the documents investigation last year. Her rulings were overturned on appeal. Cannon would determine, among other things, when a trial would take place and what Trump's sentence would be if he were found guilty. 😊</p> | <p><b>MLLM+ICL:</b> A photo shows Babatunde Fashola accessing a server containing election result data. [4]</p>                                                                                                                               | <p>However, the MLLM+ICL baseline claim also does not capture the key detail that Fashola allegedly illegally accessed the server. The high score is arguably inflated, as the claim omits the critical detail from the reference that the access was "illegally" obtained. The automated judge rewards the partial semantic overlap, highlighting a weakness in the reference-based scoring metric itself.</p>                                                                                     |
| <p><b>MICE:</b> The image shows election result data from Civichive for the 2023 Nigerian presidential election. [1]</p>                                                                                                                                                                                                                                                                                                                                                          | <p><b>Reference:</b> A photo shows Aileen Cannon, the US district court judge assigned to the criminal case against Donald Trump for retaining classified documents, wearing Trump apparel outside one of the former president's rallies.</p> | <p>Here, the MLLM+ICL baseline outperforms MICE by successfully integrating key details from both the text and the image. The reference claim's core is that Judge Cannon is wearing "Trump apparel," a visual fact explicitly mentioned in the baseline's output.</p>                                                                                                                                                                                                                              |
| <p><b>MLLM+ICL:</b> Judge Aileen Cannon, who was appointed by Trump in 2019 and initially assigned to oversee his case, is shown in the photo wearing pro-Trump attire. [4]</p>                                                                                                                                                                                                                                                                                                   | <p><b>MICE:</b> Aileen Cannon, a Trump appointee, has been assigned to oversee the classified documents case against Donald Trump, and she previously made decisions favorable to him that were later overturned. [2]</p>                     | <p>In contrast, the MICE framework's treated the image as merely illustrative. Its reasoning process failed to grasp that the visual information was the most salient part of the claim.</p>                                                                                                                                                                                                                                                                                                        |
|                                                                                                                                                                                                                                                                                                                                                                                                | <p><b>MICE:</b> Aileen Cannon, a Trump appointee, has been assigned to oversee the classified documents case against Donald Trump, and she previously made decisions favorable to him that were later overturned. [2]</p>                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

Table H3: Analysis on instances where the baseline MLLMs outperforms the MICE framework, using Gemini 2.0 Flash as the underlying model. The numbers within the square brackets represent the LLM-based reference based score.

# A Multi-Aspect Evaluation Framework for Synthetic Data: Case Study on Irony and Sarcasm

Laura Majer<sup>♡</sup> Ana Barić<sup>♡</sup>

Florijan Sandalj Ivan Unković Bojan Puvača Jan Šnajder

TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb  
{name.surname}@fer.hr

## Abstract

Data augmentation (DA) using large language models (LLMs) is a cost-effective method for generating synthetic data, particularly for tasks with scarce datasets. However, its potential remains largely underexplored, both in terms of augmentation configuration and evaluation of synthetic data. This paper investigates LLM-based synthetic data generation for irony and sarcasm, two subjective and context-dependent forms of figurative language. We propose a multi-aspect evaluation framework assessing synthetic data’s *utility-plausibility* and *extrinsic-intrinsic* dimensions through four aspects: predictive performance, sample diversity, linguistic properties, and human judgment. Our findings indicate that other aspects of evaluation, like diversity and linguistic features, do not necessarily correlate with an increase in predictive performance, underscoring the importance of multi-faceted evaluation. This work highlights the potential of LLM-based DA for irony and sarcasm detection, offering insights into the linguistic competence of LLMs. As synthetic data becomes increasingly prevalent, our framework offers a broadly applicable and crucial evaluation method, particularly for linguistically complex tasks.

 [takelab/llm-irony-sarcasm](https://github.com/takelab/llm-irony-sarcasm)

## 1 Introduction

Irony and sarcasm are complex pragmatic phenomena that present a significant challenge in both linguistics and NLP. This is primarily due to their ambiguity, contextual dependency, subjectivity, and the inherent incongruence between expressed and intended meaning. Some datasets have only a few dozen examples per category, which is insufficient for training supervised models or fine-tuning pre-trained classifiers for irony and sarcasm detection. The challenge is further aggravated by the fact that

<sup>♡</sup>Equal contribution

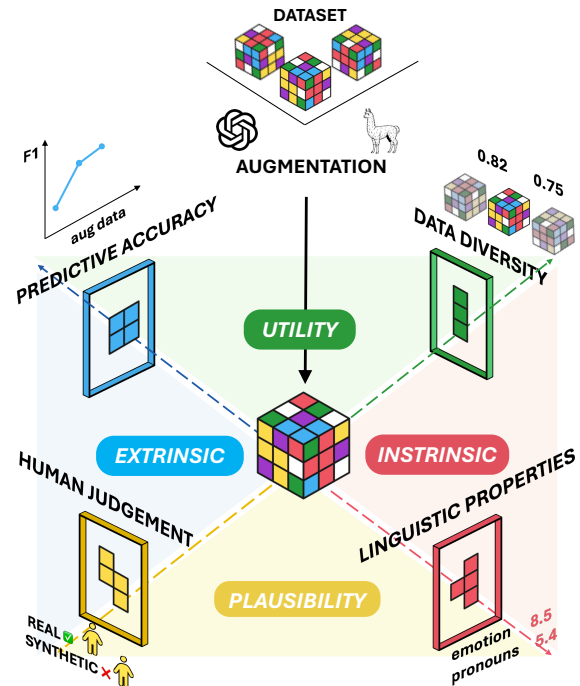


Figure 1: Multi-aspect framework for evaluating synthetic data quality through four aspects spanning the *utility-plausibility* and *extrinsic-intrinsic* dimensions.

linguistics lacks clear definitions of sarcasm and irony, as well as a consistent distinction between the two (see Appendix 5 for examples). This inconsistency is reflected in dataset annotations (Jang and Frassinelli, 2024), which makes it difficult to align or merge datasets. Thus, despite increasingly sophisticated language models, detecting irony and sarcasm remains a challenging task and a benchmark of true language understanding.

In NLP, data augmentation (DA) is a widely adopted technique used to expand existing datasets by generating additional training examples through various transformations, improving model robustness and performance, especially in low-resource settings (Feng et al., 2021; Chen et al., 2023). While previous DA approaches relied on simple linguistic transformations, such as paraphrasing,

synonym replacement, or back-translation (Wei and Zou, 2019), Large Language Models (LLMs) have proven more effective due to their strong language fluency, enhancing the plausibility and diversity of generated data (Ding et al., 2024). Another advantage is LLMs’ capacity for in-context learning (ICL) (Brown et al., 2020), which enables the use of detailed instructions and few-shot demonstrations to produce higher-quality synthetic data.

However, despite its proven utility, LLM-based DA remains underexplored, both regarding augmentation configurations and the evaluation of synthetic data. The latter is especially important for linguistically complex and subjective tasks, such as sarcasm and irony detection. While prior work often used predictive accuracy as a proxy for synthetic data quality (Yoo et al., 2021), this metric alone can be misleading. Performance gains may stem from spurious correlations (Zhou et al., 2024), and may be limited when synthetic instances lack complexity or closely resemble the original data. Finally, improvements from synthetic data, even when the data is diverse and boosts in-domain performance, do not demonstrate that it genuinely captures the targeted linguistic phenomena or that it is indistinguishable from real data to proficient language users. This makes evaluation beyond downstream performance essential, not only for validating the synthetic data itself but also for analyzing the generative model that produced it. Systematic analysis of the generated instances can reveal which linguistic properties are consistently encoded and which are lacking, indicating that the model relies on shortcuts or artifacts. In this sense, synthetic data serves as a diagnostic probe: examining its structure and limitations provides direct evidence of the model’s inductive biases and the extent to which it reproduces the target phenomenon, insights that zero- and few-shot classification performance alone cannot provide.

In this paper, we address the lack of systematic assessment of synthetic data for complex tasks by proposing a multi-aspect framework for evaluating its quality (Figure 1). This framework considers four key aspects of synthetic data: predictive accuracy, data diversity, linguistic properties, and human judgment. Together, these aspects span two key dimensions of data evaluation: the *extrinsic–intrinsic* dimension (examining how synthetic data appears externally versus its inherent characteristics) and the *utility–plausibility* dimension (assessing its usefulness for a given task versus its

resemblance to real data).

We apply the proposed framework to LLM-based DA for irony and sarcasm detection. Considering various models, prompt configurations, and numbers of demonstrations, we aim for novel insights into the impact of DA configuration choices on the different aspects of the generated synthetic data. In particular, we focus on the following research questions: **RQ1**. How does synthetic data influence fine-tuning, and is it sufficiently diverse both internally and relative to the few-shot demonstrations? **RQ2**. Does synthetic data accurately capture the linguistic phenomena it is intended to represent, and how closely it resembles – or differs from – the original data it was modeled on? **RQ3**. Which augmentation setups produce synthetic data that scores best across multiple evaluation aspects, and do the different aspects correlate?

Our analyses show that while adding synthetic data improves predictive performance, the synthetic data points are mostly easy-to-learn and exhibit lower diversity. We also demonstrate that human annotators do not easily distinguish synthetic from original data, verifying the potential of using DA for irony and sarcasm detection. Overall, our multi-aspect evaluation shows that different DA configurations optimize for various quality aspects of synthetic data. Thus, rather than considering only the predictive performance on augmented data, our approach gives rise to Pareto-optimal sets of DA configurations, from which DA practitioners should choose based on the specific aspects they aim to maximize. Our work emphasizes the importance of thorough and multifaceted evaluation of synthetic data, particularly for subjective and linguistically complex tasks, and provides a framework applicable across domains. We make all synthetic datasets generated and used in this study publicly available.

## 2 Related Work

Our work connects two research areas: (1) irony and sarcasm classification, and (2) data augmentation using LLMs. The following is an overview of the main research done in these areas.

### 2.1 Irony and sarcasm detection

In linguistics, there is no agreement on definitions for sarcasm and irony, or their clear distinction (see Appendix 5 for examples). The primary distinction between the two categories lies in the target. According to this view, irony is general, whereas

sarcasm is more directed and malevolent (Averbeck, 2013), often linked to ridicule (Lee and Katz, 1998), and characterized by a mocking, sharper tone. Due to the lack of standardized definitions for these phenomena, distinguishing between intended and perceived irony or sarcasm can be difficult – especially in contexts with limited cues. This ambiguity makes manual data annotation challenging, as guidelines are often imprecise. As a result, many datasets are constructed by scraping the web for ironic hashtags without involving human annotators (Khodak et al., 2018), fail to differentiate between sarcasm and irony (Van Hee et al., 2018), or include only a small number of examples for each category (Abu Farha et al., 2022).

In response to these annotation and dataset limitations, recent research has sought to improve the robustness and nuance of sarcasm and irony detection. For instance, Jang and Frassinelli (2024) conduct both intra- and cross-dataset evaluations to assess the generalizability of sarcasm detection models, while Casola et al. (2024) emphasize the importance of a perspectivist approach by building an irony corpus from social media that includes disaggregated annotations and annotator metadata.

## 2.2 Data augmentation with LLMs

Since data annotation is both expensive and time-consuming (Wang et al., 2021), data augmentation (DA) often provides a more cost-effective and scalable alternative. In NLP, DA takes many forms, ranging from simple techniques such as word insertion, swapping, and synonym replacement (Wei and Zou, 2019) to more sophisticated methods utilizing pre-trained transformers for low-resource scenarios (Kumar et al., 2020).

Recently, LLMs have garnered significant attention for DA. Piedboeuf and Langlais (2025) demonstrate that LLM-based methods – particularly zero- and few-shot prompting – can outperform classical DA techniques, confirming their effectiveness. One early method, GPT3Mix (Yoo et al., 2021), introduced a prompt-based approach that selects two random samples from the dataset and includes them in the prompt alongside dataset-specific meta-information. Building on this idea, Møller et al. (2024) enhanced it with a task-specific system prompt to better guide generation. In a different line of work, Yang et al. (2024) proposed a multi-step framework that iteratively trains the model and augments the dataset using challenging, model-generated examples.

When evaluating DA methods, most approaches focus solely on predictive model performance when synthetic data is added to the training set. However, for subjective and linguistically rich tasks like irony and sarcasm detection, this can overlook important aspects such as semantic diversity and linguistic adherence. In this work, we broaden the evaluation criteria for data augmentation methods to include not only model performance but also data quality, focusing on semantic diversity and linguistic properties through the multi-aspect framework.

## 3 Experimental Setup

**Datasets.** For modeling irony and sarcasm, we utilize a dataset from SemEval-2022 Task 6, introduced by Abu Farha et al. (2022). We use subtasks A and B that determine whether a tweet is either sarcastic or non-sarcastic, and classify the specific type of ironic speech, respectively. More concretely, subtask B classifies each tweet into one of six ironic speech categories defined by Leggitt and Gibbs (2000). For our experiments, we used two main categories: sarcasm and irony, consisting of 893 and 175 tweets, respectively. In the original train-test split, the training set consisted of 713 sarcasm tweets and 155 irony tweets, whereas the test set consisted of 180 sarcasm tweets and 20 irony tweets. Furthermore, we randomly sampled non-sarcastic tweets from subtask A dataset. Using these, we constructed two balanced datasets: one for sarcasm, consisting of 1,786 tweets, and one for irony, consisting of 350 tweets.

**Models.** We use instruction-tuned LLMs since we formulate our prompts as n-shot instructions. We opt to include both an open-source and a closed-source instruction-following model: Llama-3.1-8B-Instruct (llama3; AI@Meta (2024)) and GPT-4o (gpt4; OpenAI et al. (2024)), respectively.

**Prompt construction.** To construct LLM prompts for ironic and sarcastic tweet augmentation, we follow Møller et al. (2024), who focused on tasks involving evaluative language, including sentiment analysis, offensive speech detection, and emotion recognition. We divide the prompt format into two parts, following the taxonomy proposed by Jeoung et al. (2025): (1) fixed and (2) variable parts. The fixed part of the prompt is present in all configurations and includes a short task description (e.g., *Your task is to generate X ironic tweets.*) and output format instruction (e.g., *Separate each new instance with ordinal numbers.*), along with a

list of tweet examples used as demonstrations for the few-shot prompts (either 1- or 3-shot). In the 1-shot setting, the model is prompted to generate three new instances, while in the 3-shot setting, it is prompted to generate nine instances. Unlike the fixed part, the variable parts, which include the task definition and stylistic guidance, are either added to or omitted from the prompt, depending on the prompt configuration. In total, we evaluate 16 distinct prompt configurations for each task and model, generated by combining two variable components with both 1- and 3-shot scenarios.

To ensure we cover all relevant aspects of the datasets we aim to augment, we obtain definitions for irony and sarcasm from the iSarcasm dataset (Oprea and Magdy, 2020). Furthermore, we create stylistic guidance instructions by describing the informal and unstructured style typical of tweets. Before selecting a specific wording for both the fixed and variable parts of the prompt, we conduct an initial evaluation of different prompt variants on a subset of data for each task to identify the most effective one in terms of instruction-following quality. In subsequent analyses, we label each prompt configuration with ‘1’ or ‘3’ (for 1-shot and 3-shot, respectively), followed by letters indicating the included variable parts (‘D’ for Definition, ‘S’ for Stylistic guidance, and ‘X’ for neither). For example, ‘3DS’ denotes a 3-shot prompt with both Definition and Stylistic guidance included. The complete prompts, including all variable parts, are shown in Table 7 in Appendix E.1. In what follows, we refer to a specific combination of an LLM and prompt configuration as *augmentation setup*.

**Demonstration selection.** We utilize all training examples as demonstrations in 1-shot configurations and employ a cluster-based selection method to choose semantically similar examples for each prompt in 3-shot configurations for each task. Concretely, we obtain training data embeddings from BERT (Devlin et al., 2019), reduce dimensionality with PCA, and apply K-Means to cluster semantically similar examples.

## 4 Utility

We now apply the proposed evaluation framework to analyze the synthetic data generated by the DA procedure described above. Our results are presented along the *utility-plausibility* dimension, beginning with utility. Utility captures the extent to which synthetic data supports model training and performance. Specifically, we assess utility through

two aspects: predictive accuracy (an *extrinsic* aspect) and data diversity (an *intrinsic* aspect).

### 4.1 Predictive Accuracy

Since DA expands datasets often too small for model fine-tuning, improved predictive performance in classification is typically used as a proxy for synthetic data quality (Wei and Zou, 2019; Møller et al., 2024; Piedboeuf and Langlais, 2025).

**Original data.** To establish baseline performance on the original data, we fine-tune three transformer-based models commonly used for irony and sarcasm detection: BERT (Devlin et al., 2019), BERTweet (Nguyen et al., 2020), and RoBERTa (Zhuang et al., 2021). Each model is fine-tuned five times with different random seeds on the task-specific training sets, using default hyperparameters (see Appendix B). In addition, we evaluate two LLMs in 0-shot and 2-shot settings: Llama-3.1-8B-Instruct (llama3; AI@Meta (2024)) and Gemma-3-4B (gemma3; Team et al. (2025)). For prompting, we use the model-specific chat templates and provide test examples as input. The full set of baseline prompts is provided in Appendix C.1. Table 1 reports the average F1 scores on the test sets. For the fine-tuned models, results are averaged over five runs evaluated on a shared test set. To assess overall performance differences among models, we apply the Friedman test. For the irony task, the test does not reveal statistically significant differences among models ( $\chi^2(df = 4) = 10.03$ ,  $p = 0.124$ ); therefore, no post-hoc pairwise comparisons are conducted. For the sarcasm task, the Friedman test indicates statistically significant differences among models ( $\chi^2(df = 4) = 26.57$ ,  $p < 0.001$ ). We subsequently perform post-hoc pairwise comparisons using Wilcoxon signed-rank tests with Holm correction to account for multiple comparisons. However, none of the pairwise comparisons remain statistically significant after correction (all Holm-adjusted p-values  $> 0.05$ ). Since none of the models significantly outperforms the others across both tasks, we select BERTweet as the baseline model for the remainder of the study, as it achieves the highest average F1 score.

**Synthetic data.** We measure the influence of synthetic data on model performance by fine-tuning BERTweet, our best-performing model on original data, with real and synthetic data for all prompt configurations. To examine how performance changes with different synthetic-to-real data ratios, we grad-

|        | Model    | Irony           | Sarcasm         |
|--------|----------|-----------------|-----------------|
| FT     | BERT     | $0.60 \pm 0.17$ | $0.53 \pm 0.02$ |
|        | BERTweet | $0.68 \pm 0.08$ | $0.70 \pm 0.07$ |
|        | RoBERTa  | $0.39 \pm 0.22$ | $0.64 \pm 0.07$ |
| 0-shot | llama3   | $0.68 \pm 0.06$ | $0.56 \pm 0.01$ |
|        | gemma3   | $0.44 \pm 0.02$ | $0.37 \pm 0.01$ |
| 2-shot | llama3   | $0.64 \pm 0.03$ | $0.73 \pm 0.01$ |
|        | gemma3   | $0.65 \pm 0.02$ | $0.73 \pm 0.01$ |

Table 1: F1 scores (mean  $\pm$  std) of models trained on the original dataset for irony and sarcasm detection tasks.

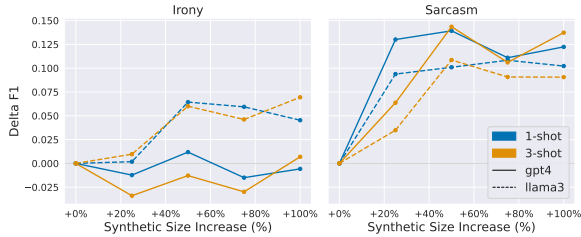


Figure 2: F1 score change relative to original training data across increasing synthetic data sizes, grouped by few-shot setting (color) and model type (line style), for irony and sarcasm detection. Averaged over five runs and across configurations; positive values indicate performance gains from synthetic augmentation.

ually increase the amount of synthetic data in the training set following the approach of Møller et al. (2024). This setup also helps identify whether there is an upper limit to the amount of synthetic data beyond which model performance plateaus. Figure 2 shows the change in F1 score aggregated across different few-shot settings (either 1-shot or 3-shot). The addition of synthetic data yielded a maximum performance improvement of 10% for irony and a higher improvement of 15% for sarcasm; similar performance gains were also reported in (Møller et al., 2024). Additionally, we observe that, across both tasks and all few-shot settings, performance stabilizes once synthetic data comprises approximately 50% of the total training set. Additionally, we notice a larger performance gain for sarcasm. Similar trends are observed for prompts extended with definition and style parts (see Appendix C.2).

Motivated by the observed plateau in model performance as the amount of synthetic data increases, we employ dataset cartography (Swayamdipta et al., 2020) to assess if this stagnation stems from differences in example difficulty distribution between the original and synthetic datasets. We fine-tune BERTweet for 3 epochs using a balanced set of

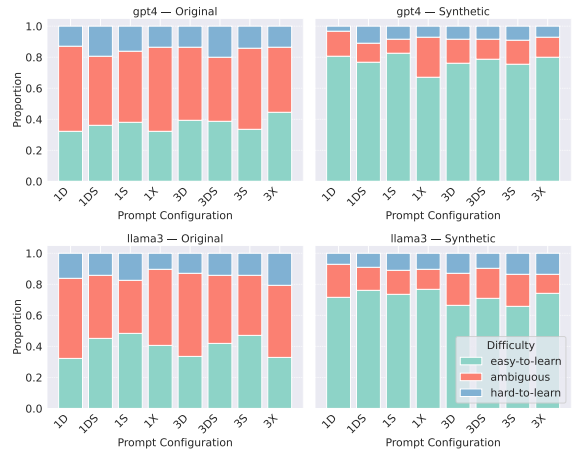


Figure 3: Distribution of difficulty levels (easy-to-learn, ambiguous, and hard-to-learn) across different prompt configurations for the irony detection task, with the original data (left) and the synthetic data (right). Each bar shows the normalized proportion of each difficulty category for each prompt configuration.

original and synthetic data and default hyperparameters, and record the confidence, correctness, and variability scores, later used to identify easy-to-learn, ambiguous, and hard-to-learn examples. Figure 3 shows the differences in difficulty distribution between original and synthetic data for irony. We observe higher proportions of easy-to-learn examples in synthetic data across both tasks and prompt configurations. A similar trend is observed for sarcasm (see Figure 7 in the appendix).

## 4.2 Data Diversity

The stagnation in F1 scores despite adding more synthetic data, coupled with generally volatile training curves and the observation that many synthetic samples fall into the easy-to-learn region of the data map, directly motivates data diversity as another evaluation aspect.

**Method.** We quantify the diversity of the synthetic dataset through the semantic similarity between pairs of individual instances. We use the SBERT model (Reimers and Gurevych, 2019) to calculate instance-level similarity, with the zero-one output range, reflecting the degree of semantic overlap between a pair of instances. We then investigate the semantic diversity of synthetic datasets across different configurations, approaching it from two angles. The first type of diversity we consider is *intra-diversity*, motivated by the observation that synthetic examples often contain repetitive phrases or recurring themes that may superficially boost

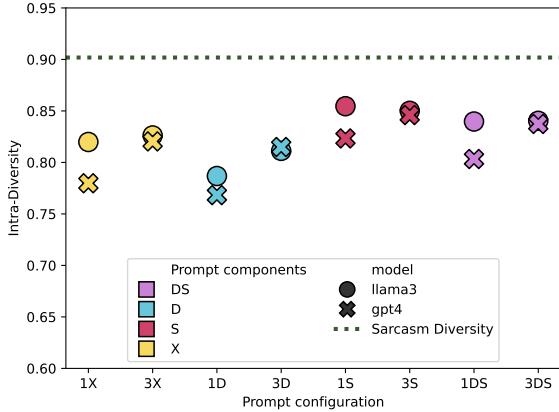


Figure 4: Intra-diversity plots for sarcasm synthetic datasets across prompt setups (shown in color) and models (shown in symbols).

performance but ultimately offer limited variety – potentially explaining the observed performance plateau. Given a synthetic dataset of size  $M$  and semantic similarity  $\text{sim}(\cdot, \cdot)$  between two synthetic instances,  $\tilde{x}_i$  and  $\tilde{x}_j$ , we compute intra-diversity as:

$$d_{\text{intra}} = 1 - \frac{1}{M} \sum_{i \neq j} \text{sim}(\tilde{x}_i, \tilde{x}_j)$$

Second, since the few-shot demonstrations are the only component that varies between runs, the generation process may produce instances that are too similar to the demonstrations. To assess this, we calculate *inter-diversity*, which quantifies the extent to which the synthetic instances are different from the few-shot demonstrations included in the prompt used to generate them (see Appendix C.3 for the formula).

**Original data.** To compare augmented datasets with the original dataset, we calculate intra-diversity for the original irony and sarcasm datasets. The resulting scores are 0.888 for irony and 0.902 for sarcasm, respectively.

**Synthetic data.** Figure 4 shows the intra-diversity values across augmentation setups for sarcasm. Prompt configurations are color-coded, and marker shapes indicate the model. Although synthetic datasets show lower diversity than the original, the difference of  $\sim 0.05$  of top-performing configurations is small – especially considering sets are three times larger. Prompt configurations strongly influence diversity: style-only prompts produce the most diverse data, while definition-only prompts result in the least. For gpt4, three-shot prompts clearly outperform one-shot prompts.

We observe similar patterns for sarcasm, shown in Figure 8. Inter-diversity behaves similarly, but varies less across prompt types and more between one-shot and three-shot setups (see Figure 9 in the appendix). Three-shot setups yield higher diversity across tasks, suggesting that not all examples influence every output, increasing batch diversity.

## 5 Plausibility

Predictive performance and data diversity assess how synthetic data impacts a model’s training and performance when included in its training set, focusing on its *utility*. In contrast, *plausibility* concerns whether the synthetic data exhibits linguistic properties associated with irony and sarcasm and appears realistic compared to the original data. We evaluate plausibility through two aspects: linguistic properties (an *intrinsic* aspect) and human judgment (an *extrinsic* aspect).

### 5.1 Linguistic Properties

Since irony and sarcasm are inherently linguistic phenomena, synthetic data should ideally conform to specific definitions and linguistic properties for accurate representation. Generating misaligned samples can introduce flawed data points and encourage shortcut learning during fine-tuning, resulting in a model that incorrectly represents irony and sarcasm. Conversely, a model that effectively captures these concepts is more likely to generalize to out-of-distribution data, making linguistic properties an essential aspect of synthetic data.

**Method.** Among various approaches to modeling linguistic features relevant to irony and sarcasm, we use the LIWC 2022 lexicon (Boyd et al., 2022). This choice is motivated by Sulis et al. (2016), who showed that user-marked figurative language (e.g., irony, sarcasm) exhibits distinct patterns detectable with closed-vocabulary methods. LIWC 2022 includes both basic linguistic categories (e.g., part-of-speech tags, word types) and more nuanced dimensions, such as emotions (anger, anxiety, sadness), psychological states (*Cognition, State*), and social behaviors (*Interpersonal conflict, Moralization*). It comprises 116 features, each representing the proportion of words in a text that match predefined lexical categories, yielding sparse feature vectors per sample and sparse matrices per dataset (ironic, sarcastic, or neutral). To compare LIWC features between original and synthetic data, we use the non-parametric Common Language Effect Size (CLES) (McGraw and Wong, 1992), which quantifies the

probability that a randomly selected instance from one group scores higher than one from another. Prior studies suggest CLES is more interpretable than traditional effect size metrics (Brooks et al., 2014). We correct for multiple comparisons across all configuration–feature pairs using the Benjamini–Hochberg procedure with a false discovery rate threshold of  $\alpha = 0.05$ .

**Original data.** Before comparing linguistic features of original and synthetic data, we first examine differences among ironic, sarcastic, and neutral instances in the original dataset to assess if prior findings (Sulis et al., 2016) hold. Since LIWC includes many features, we focus on a few discriminative categories – *Emotion*, *Social*, *Physical*, and *States* – based on definitions of irony and sarcasm. To better isolate patterns, we average only non-zero LIWC scores. For instance, we find that irony is more associated with anxiety, while neutral instances tend to reflect more positive emotion (see Figure 10 in the appendix).

**Synthetic data.** Figure 5 shows heatmaps of the CLES score for sarcasm across augmentation setups, with rows representing prompt configurations per model and columns representing LIWC feature categories sorted by average, ascending. After applying FDR correction, all configuration–feature comparisons were statistically significant. Red values indicate a lower presence of the feature in synthetic data compared to the original, while blue values indicate a higher presence of the feature. Paler squares indicate a higher overlap between synthetic and original features.

For both sarcasm and irony and across both models, similar patterns emerge. The category *Cognition*, which includes features such as *certitude*, *causation* and *memory*, was consistently underrepresented in synthetic data, while *Tone*, covering the degree of both positive and negative tone, was consistently overrepresented in synthetic data.

## 5.2 Human Judgement

To complement the plausibility evaluation, we conduct a human judgment study to assess the quality of the synthetic data from two aspects: *data detectability* and *data authenticity*, by comparing real and synthetic data. Data detectability assesses how well humans can distinguish the difference between irony and sarcasm for real and synthetic data. Data authenticity assesses the extent to which humans perceive synthetic data as believable or realistic

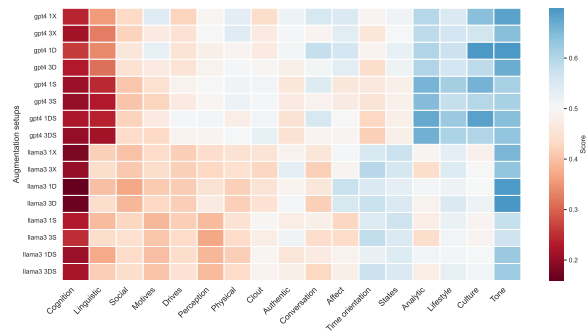


Figure 5: Heatmap of CLES scores for aggregated LIWC features across augmentation setups for sarcasm, lower values represent lower frequency in synthetic data compared to original.

compared to real data.

**Annotation setup.** Each aspect in the study is evaluated independently through two annotation tasks, with annotations provided by twelve volunteer participants. In both tasks, each example was annotated by three participants. Furthermore, we instructed participants to complete the data detectability task first and then proceed to the data authenticity task only after its completion. A strict task order was imposed to mitigate participant anchoring bias, as awareness of the different example origins (real or synthetic) could influence label judgments in the data detectability task (e.g., basing judgments on perceived origin rather than linguistic fit alone). To avoid this, participants were informed about the inclusion of the synthetic data only after completing the data detectability task. For both tasks, we randomly selected a balanced sample consisting of real and synthetic data, where synthetic data is selected from the prompt configurations that achieve the highest F1 scores for irony (gpt4 1X) and sarcasm (llama3 1X), respectively.

**Data detectability.** To evaluate whether humans can recognize irony and sarcasm equally well in real and synthetic data, we design an annotation task comparing how these categories are perceived across data sources. The intuition is that if synthetic data is linguistically and semantically plausible, then the ability of humans to identify irony and sarcasm should be comparable to that in real data. To this end, we devise a ternary classification task where annotators label each tweet as either *ironic*, *sarcastic*, or *neutral*, without knowing whether the tweet is real or synthetic. We compile a balanced dataset of 350 examples in total, uniformly sampled from five classes: real irony, real

sarcasm, real neutral, synthetic irony, and synthetic sarcasm. All irony and sarcasm examples, regardless of source, are grouped under their respective labels, while the neutral class (which only exists in real data) is treated as a separate category.

**Data authenticity.** To further inspect the quality of synthetic data compared to real data, we ask participants to label tweets as either real or synthetic, aiming to assess how convincing the synthetic data is from a human perspective. We uniformly sample a total of 280 ironic and sarcastic tweets from both the real and synthetic datasets. Additionally, we ensure that no examples are repeated from the data detectability setup.

**Annotation details.** We use the open-source platform Alanno (Jukić et al., 2023) to conduct both annotation tasks. All twelve participants are either at the Master’s (4), PhD (7), or postdoctoral level (1), working within the NLP domain. The group includes 11 computer scientists and 1 psychologist. The ages range from 24 to 34 years. All participants are fluent in English; two are native speakers, and ten demonstrate high proficiency.

**Results.** We assess how well annotator labels align with the gold labels for both real and synthetic data by calculating accuracy scores, which are reported for data detectability and authenticity in Table 2. In absolute terms, detectability scores fall below a random baseline of 0.5, highlighting the challenge of distinguishing irony from sarcasm – further supported by low annotator agreement (Fleiss’  $\kappa = 0.35$ ). Similarly, despite higher accuracy on the data authenticity task, poor agreement (Fleiss’  $\kappa = 0.14$ ) indicates difficulty and subjectivity in distinguishing real from synthetic data. Table 3 presents the confusion matrices for the original and synthetic datasets. The similar distribution of predictions across irony, sarcasm, and neutral labels suggests that the synthetic data exhibits class-level ambiguities and decision patterns comparable to those observed in original data. It is also a confirmation that LLMs generate neutral Detectability accuracy is higher on synthetic examples, whereas authenticity accuracy is higher on real examples. This suggests that synthetic data is easier to label correctly but harder to identify as artificial – likely due to being more straightforward or less nuanced, but still fairly realistic. This aligns with the finding that the majority of synthetic data falls under the easy-to-learn category.

|                      | Irony |           | Sarcasm |           |
|----------------------|-------|-----------|---------|-----------|
|                      | Real  | Synthetic | Real    | Synthetic |
| <b>Detectability</b> | 0.25  | 0.39      | 0.40    | 0.52      |
| <b>Authenticity</b>  | 0.77  | 0.51      | 0.81    | 0.54      |

Table 2: Detectability vs. authenticity accuracy scores of real and synthetic annotations for irony and sarcasm.

|           |                | Predicted |         |         |
|-----------|----------------|-----------|---------|---------|
|           |                | Irony     | Sarcasm | Neutral |
| Original  | <b>Irony</b>   | 19        | 25      | 26      |
|           | <b>Sarcasm</b> | 19        | 27      | 24      |
| Synthetic | <b>Irony</b>   | 27        | 15      | 28      |
|           | <b>Sarcasm</b> | 28        | 36      | 6       |

Table 3: Confusion matrices for detectability on original and synthetic data.

## 6 Discussion

We now address our research questions by examining results across multiple evaluation dimensions. **RQ1**, which concerns the impact of synthetic data on predictive performance and diversity, is best informed by the utility dimension. While synthetic data boosts performance, gains tend to plateau – likely due to its high proportion of easy-to-learn examples. Although diversity remains lower than in the original data, it is notable given the increased dataset size. For **RQ2**, the plausibility reveals that synthetic data often lacks variation in expression categories across setups but matches the original data in terms of linguistic markers. Human annotators’ difficulty in distinguishing synthetic from real examples further supports its plausibility. However, the clearer distinction between irony and sarcasm in synthetic data – reflected in higher labeling accuracy – suggests improved clarity of linguistic cues, although possibly at the cost of realism.

In contrast to **RQ1** and **RQ2**, which focus on individual dimensions, **RQ3** spans the entire evaluation framework, exploring correlations among evaluation aspects. All aspects, except for human judgment, are directly comparable across setups; we limited human judgment annotations to a single prompt configuration due to resource constraints. The other three aspects are standardized to values in the  $[0, 1]$  range: F1 score synthetic data of original size (predictive accuracy), intra-diversity (data diversity), and average CLES divergence from 0.5 (linguistic properties). We report the scores for sarcasm in Table 4; similar trends are observed for irony (see Table 9 in Appendix F).

| gpt4   |       |       |       | llama3 |       |       |       |
|--------|-------|-------|-------|--------|-------|-------|-------|
| Config | F1    | Div.  | Ling. | Config | F1    | Div.  | Ling. |
| 3DS★   | 0.754 | 0.837 | 0.085 | 1D     | 0.744 | 0.787 | 0.080 |
| 1DS    | 0.736 | 0.803 | 0.090 | 3D     | 0.708 | 0.811 | 0.080 |
| 1D     | 0.736 | 0.768 | 0.078 | 1DS★   | 0.699 | 0.840 | 0.069 |
| 3D     | 0.732 | 0.815 | 0.080 | 3X     | 0.696 | 0.826 | 0.077 |
| 3X★    | 0.719 | 0.820 | 0.068 | 1X     | 0.692 | 0.820 | 0.071 |
| 3S★    | 0.708 | 0.846 | 0.081 | 3DS★   | 0.658 | 0.840 | 0.070 |
| 1X     | 0.696 | 0.780 | 0.077 | 1S★    | 0.637 | 0.855 | 0.069 |
| 1S     | 0.686 | 0.823 | 0.086 | 1S★    | 0.637 | 0.855 | 0.069 |

Table 4: Score comparison across performance (*F1*), diversity (*Div.*), and linguistic properties (*Ling.*) for sarcasm. Cell shading reflects relative score quality per task per model (darker = better; for *Ling.*, lower is better), ★ = Pareto-optimal configuration.

**Is predictive accuracy robust enough for DA evaluation?** Ideally, synthetic data that improves predictive performance should be both diverse and linguistically consistent with the target phenomena. However, our results show that these qualities do not always align – for example, gpt4-generated data for sarcasm with high F1 scores had lower diversity and linguistic overlap, suggesting that increased variation in data can introduce noise. This suggests that increased variation or richness in language, while desirable in theory, may introduce noise or complexity. These findings highlight the need to evaluate model performance alongside explicit data quality metrics, rather than relying on it as a sole indicator of learning.

**Can a single prompt configuration cover all desired aspects?** No single prompt configuration optimizes performance, diversity, and linguistic consistency across models and tasks; instead, Pareto-optimal sets emerge for each model. Definition-only prompts (D) improve clarity with minimal complexity, while style-based prompts (S or DS) boost diversity but often hurt performance – especially with gpt4, whose outputs are more expressive. Few-shot settings also matter: 1-shot prompts reduce diversity but largely preserve consistency, indicating that more examples increase variation more than linguistic alignment. Overall, prompt design shapes synthetic data and often requires balancing expressiveness with model learnability. Comparing gpt4 and llama3 highlights the variability of LLMs as data generators. While gpt4 is more sensitive to prompt changes, exhibiting greater shifts in performance, diversity, and linguistic features, llama3 produces more consistent outputs with less variation.

## 7 Conclusion

In this paper, we introduced a multi-aspect evaluation framework for data augmentation (DA), targeting sarcasm and irony detection. While F1 score improvements are a standard measure of synthetic data utility, we argue that this metric is not sufficient. Our framework incorporates aspects that span the utility–plausibility spectrum and reflect diverse characteristics of the data. Results show that these aspects often do not correlate with F1 gains, revealing important gaps in conventional evaluation. Our findings demonstrate the potential of DA for complex language phenomena and emphasize the value of more nuanced, multi-dimensional evaluation in future work.

### Limitations

**Metrics.** The metrics used to evaluate different aspects may not perfectly capture the nuances of those aspects, as they are inherently relative and difficult to quantify. Nevertheless, to our knowledge, this remains the most comprehensive evaluation of DA data to date.

**Tasks** Our analysis has not been tested on other linguistically complex tasks other than irony and sarcasm, which may reveal different patterns or lead to alternative conclusions. Other tasks may also involve additional crucial aspects not covered here, particularly those that extend beyond standard metrics like F1. This only confirms the importance of multi-aspect evaluation.

**Models.** Lastly, we evaluated predictive performance using only a single model, which limits the generalizability of our findings. Different models may interact with the data in distinct ways, potentially leading to different outcomes, especially since predictive performance is an indirect measure that reflects how the data influences model learning.

### Acknowledgments

We thank our volunteer annotators for their valuable contribution to our work.

### References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. [SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic](#). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*,

- pages 802–814, Seattle, United States. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 model card](#).
- Joshua M. Averbeck. 2013. [Comparisons of ironic and sarcastic arguments in terms of appropriateness and effectiveness in personal relationships](#). *Argumentation and Advocacy*, 50(1):47–57.
- Ryan L. Boyd, Abhishek Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. [The development and psychometric properties of LIWC-22](#). Technical report, University of Texas at Austin, Austin, TX.
- M. E. Brooks, D. K. Dalal, and K. P. Nolan. 2014. [Are common language effect sizes easier to understand than traditional effect sizes?](#) *Journal of Applied Psychology*, 99(2):332–340.
- 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, et al. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- John D. Campbell and Albert N. Katz. 2012. [Are there necessary conditions for inducing a sense of sarcastic irony?](#) *Discourse Processes*, 49(6):459–480.
- Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. [MultiPICO: Multilingual perspectivist irony corpus](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.
- Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. [An empirical survey of data augmentation for limited data learning in NLP](#). *Transactions of the Association for Computational Linguistics*, 11:191–211.
- 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.
- Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. [Data augmentation using LLMs: Data perspectives, learning paradigms and challenges](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.
- Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. [A survey of data augmentation approaches for NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.
- Hyewon Jang and Diego Frassinelli. 2024. [Generalizable sarcasm detection is just around the corner, of course!](#) 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 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.
- Sullam Jeoung, Yueyan Chen, Yi Zhang, Shuai Wang, Haibo Ding, and Lin Lee Cheong. 2025. [Promptprism: A linguistically-inspired taxonomy for prompts](#). *Preprint*, arXiv:2505.12592.
- Josip Jukić, Fran Jelenić, Miroslav Bičanić, and Jan Snajder. 2023. [ALANNO: An active learning annotation system for mortals](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 228–235, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. [A large self-annotated corpus for sarcasm](#). *Preprint*, arXiv:1704.05579.
- Roger Kreuz and Sam Glucksberg. 1989. [How to be sarcastic: The echoic reminder theory of verbal irony](#). *Journal of Experimental Psychology: General*, 118:374–386.
- Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. [Data augmentation using pre-trained transformer models](#). In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.
- Christopher J. Lee and Albert N. Katz. 1998. [The differential role of ridicule in sarcasm and irony](#). *Metaphor and Symbol*, 13(1):1–15.
- John S Leggitt and Raymond W Gibbs. 2000. Emotional reactions to verbal irony. *Discourse processes*, 29(1):1–24.
- Kenneth O. McGraw and S. P. Wong. 1992. [A common language effect size statistic](#). *Psychological Bulletin*, 111(2):361–365.
- Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. [The parrot dilemma:](#)

- Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian’s, Malta. Association for Computational Linguistics.
- Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2024. [The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks](#). *Preprint*, arXiv:2304.13861.
- Sachi Nakamura. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. [BERTweet: A pre-trained language model for English tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.
- OpenAI et al. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Silviu Oprea and Walid Magdy. 2020. [iSarcasm: A dataset of intended sarcasm](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.
- Penny M. Pexman. 2008. [It’s fascinating research: The cognition of verbal irony](#). *Current Directions in Psychological Science*, 17(4):286–290.
- Frédéric Piedboeuf and Philippe Langlais. 2025. [On evaluation protocols for data augmentation in a limited data scenario](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3428–3443, Abu Dhabi, UAE. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. [Figurative messages and affect in twitter: Differences between irony, sarcasm and not](#). *Knowledge-Based Systems*, 108:132–143. New Avenues in Knowledge Bases for Natural Language Processing.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.
- Gemma Team et al. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Akira Utsumi. 2000. [Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony](#). *Journal of Pragmatics*, 32(12):1777–1806.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. [SemEval-2018 task 3: Irony detection in English tweets](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Shuangtao Yang, Xiaoyi Liu, Xiaozheng Dong, and Bo Fu. 2024. [Mini-DA: Improving your model performance through minimal data augmentation using LLM](#). In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 25–30, Mexico City, Mexico. Association for Computational Linguistics.
- Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. [GPT3Mix: Leveraging large-scale language models for text augmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. [Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614, Miami, Florida, USA. Association for Computational Linguistics.
- Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

## A Various irony and sarcasm definitions

| Citation                     | Summary                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (Kreuz and Glucksberg, 1989) | <i>Verbal irony, in contrast, need not refer explicitly to an ironic event or state. Instead, in verbal irony a speaker expresses an attitude toward some object, event, or person by saying something that is not literally true. Sarcasm is often characterized as a form of verbal irony and has been defined as "a sharp and often satirical or ironic utterance designed to cut or give pain."</i>                                                                                                                                                                                                                                                                                                                                                                                                                   |
| (Nakamura, 1995)             | <i>The allusional pretense theory of irony posits two necessary conditions for understanding ironic intent: allusion to failed expectation (FE) and pragmatic insincerity. An allusion to FE refers to a discrepancy between a certain expectation and the subsequent reality. Pragmatic insincerity, the second necessary component proposed by Kumon-Nakamura et al. (1995), follows from the felicity condition, originally discussed by Austin (1962) and later elaborated on by Searle (1979; see also Colston, 2000). When being ironic, the individual is not intending the compliment and is thus being insincere, violating the felicity condition. This insincerity is pragmatic, relating to use rather than semantics. From this view, both FE and pragmatic insincerity must be present to invite irony.</i> |
| (Utsumi, 2000)               | <i>Implicit display theory provides an additional contextual constraint for irony. A statement must be identified as being in an "ironic environment," which consists of: (1) an expectation by the speaker at time t, (2) a failure of that expectation, and (3) a negative emotional reaction to the incongruity. Ironic communication implicitly displays this environment by: (1) alluding to the expectation, (2) violating a pragmatic principle, and (3) expressing a negative attitude indirectly.</i>                                                                                                                                                                                                                                                                                                            |
| (Pexman, 2008)               | <i>Constraint satisfaction theories argue that no single cue is necessary to evoke sarcastic irony. Instead, irony is interpreted based on a set of probabilistic cues or hints evaluated in context. These might include who made the comment, the discourse or environmental support, and the manner of delivery. A single strong cue or multiple weaker cues may be sufficient to invite an ironic interpretation.</i>                                                                                                                                                                                                                                                                                                                                                                                                 |
| (Campbell and Katz, 2012)    | <i>Findings show that although several components have been claimed as "necessary" for sarcasm, none are truly required. Each factor independently predicts sarcasm ratings, and stylistic choices may play a significant role. Items with sarcastic instructions are rated more sarcastic than those without, even when other variables are controlled. Lexical features may contribute uniquely, outside traditional theory components.</i>                                                                                                                                                                                                                                                                                                                                                                             |
| (Lee and Katz, 1998)         | <i>This study supports that sarcasm often involves ridicule of a specific victim, differentiating it from irony. Sarcastic utterances bring to mind expectations associated with an identified person, while irony evokes broader, collective expectations.</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| (Averbeck, 2013)             | <i>Verbal irony is an intentionally inconsistent message that is usually benevolent and not directed at the listener. Sarcasm, however, is also inconsistent but malevolent and directed at the addressee. This distinction is based on intent and interpersonal direction.</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

Table 5: Various definitions in related work for phenomena covered by the term *ironic speech*.

## B Hyperparameters and Hardware Details

All models are fine-tuned using the HuggingFace Trainer<sup>1</sup> library on an NVIDIA RTX 3090 GPU (24GB RAM) with CUDA 12.9. We use the default hyperparameter settings provided by the Trainer library.

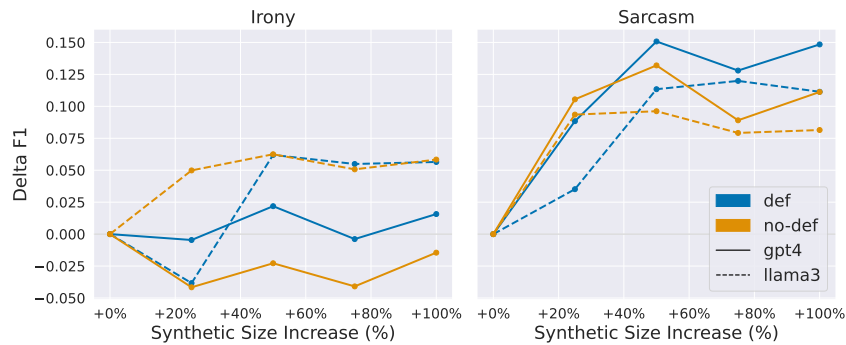
## C Utility – additional experiments and information

### C.1 Baseline prompts

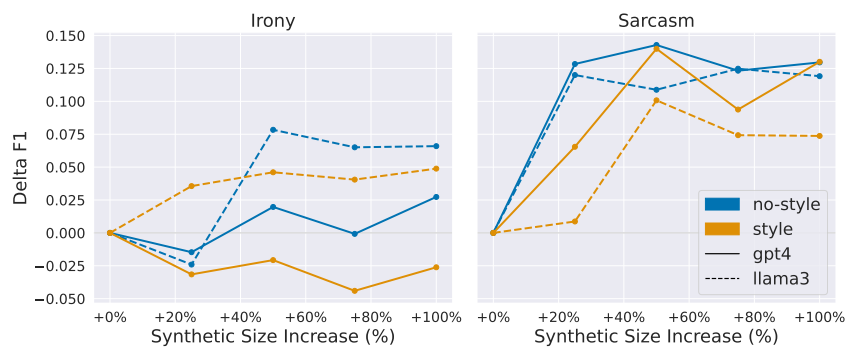
| Task    | System prompt                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sarcasm | You are a helpful assistant. Classify the following tweet into sarcastic or not sarcastic. Sarcasm is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Sarcastic tweets contradict the state of affairs, are directed towards an addressee and express a critical attitude. Output only the classification result. |
| Irony   | You are a helpful assistant. Classify the following tweet into ironic or not ironic. Irony is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Ironic tweets contradict the state of affairs but are not obviously critical toward any particular addressee. Output only the classification result.                |

Table 6: System prompts used for classification tasks.

### C.2 Predictive accuracy



(a) Grouped by definition setting.



(b) Grouped by style setting.

Figure 6: Change in F1 scores relative to baseline across increasing synthetic data sizes.

<sup>1</sup>[https://huggingface.co/docs/transformers/main\\_classes/trainer](https://huggingface.co/docs/transformers/main_classes/trainer)

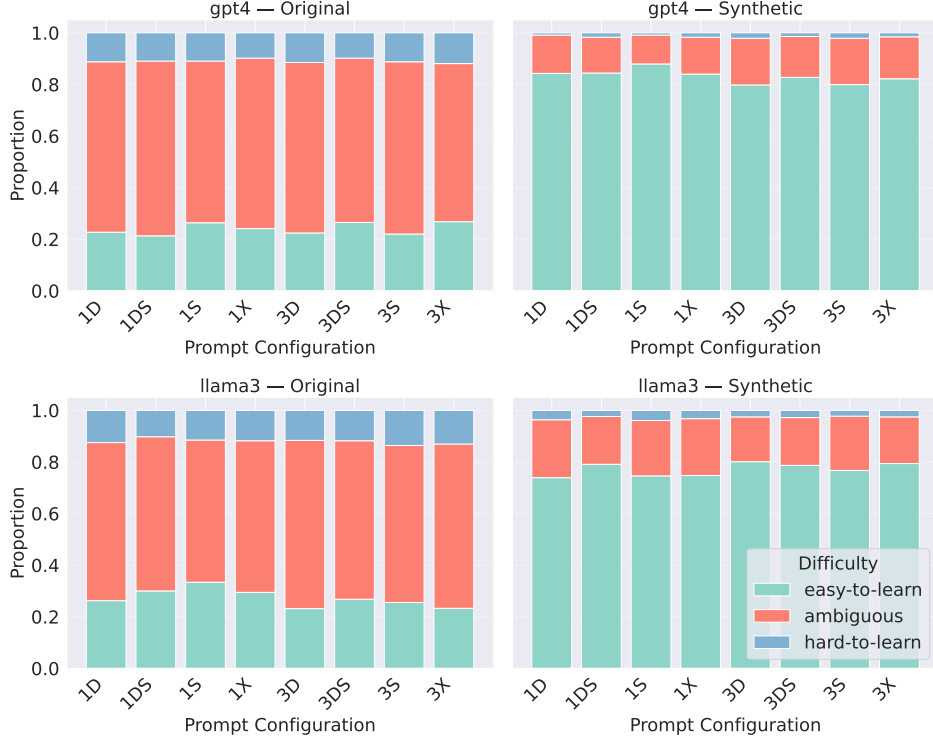


Figure 7: Comparison of the distribution of difficulty levels (easy-to-learn, ambiguous, and hard-to-learn) across different prompt configurations for sarcasm.

### C.3 Data diversity

Given a real dataset partitioned into  $R$  calls  $\{X_1, X_2, \dots, X_R\}$ , and a semantic similarity function  $\text{sim}(\cdot, \cdot)$  between real instances  $x \in X_r$  and their corresponding synthetic augmentations  $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$ , we compute *inter-diversity* as:

$$d_{\text{inter}} = 1 - \frac{1}{R} \sum_{r=1}^R \left( \frac{1}{k \cdot |X_r|} \sum_{x \in X_r} \sum_{i=1}^3 \text{sim}(x, \tilde{x}_i) \right)$$

This metric captures how distinct the synthetic data is from its corresponding real data across all calls.

## D Plausibility – additional experiments and information

### D.1 Linguistic properties

**CLES.** To focus on the data and avoid speculating about relevant linguistic categories, we use Common Language Effect Size (CLES) to select the most relevant ones for distinguishing between ironic, sarcastic, and neutral instances.

CLES is defined as:

$$\text{CLES} = P(X > Y) + 0.5P(X = Y)$$

where  $X$  and  $Y$  are randomly chosen values from the distribution of the two datasets.

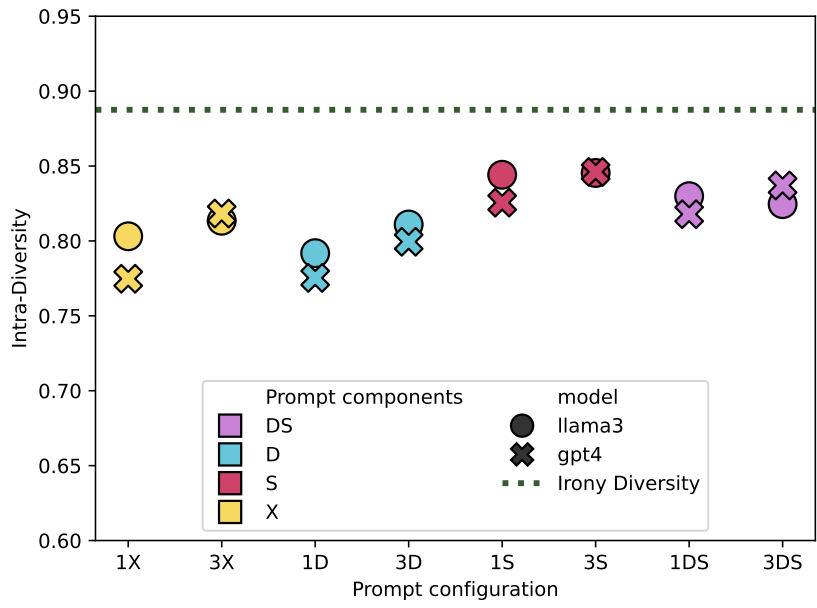
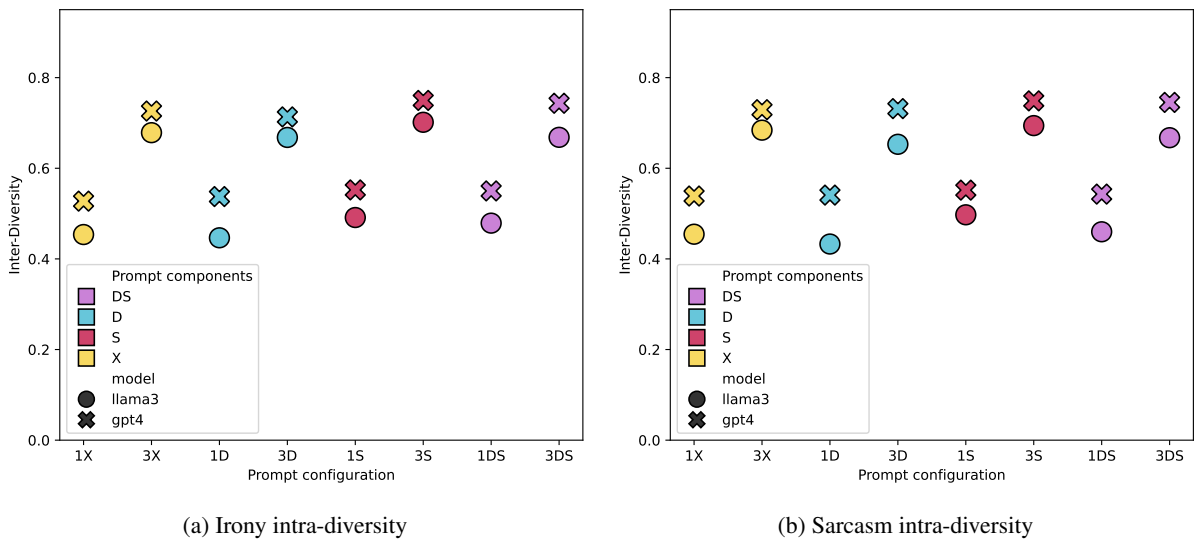


Figure 8: Intra-diversity plot for irony synthetic datasets across prompt setups (shown in color) and models (shown in symbols).



(a) Irony intra-diversity

(b) Sarcasm intra-diversity

Figure 9: Inter-diversity plots for irony and sarcasm synthetic datasets across prompt configurations (shown in color) and models (shown in symbols).

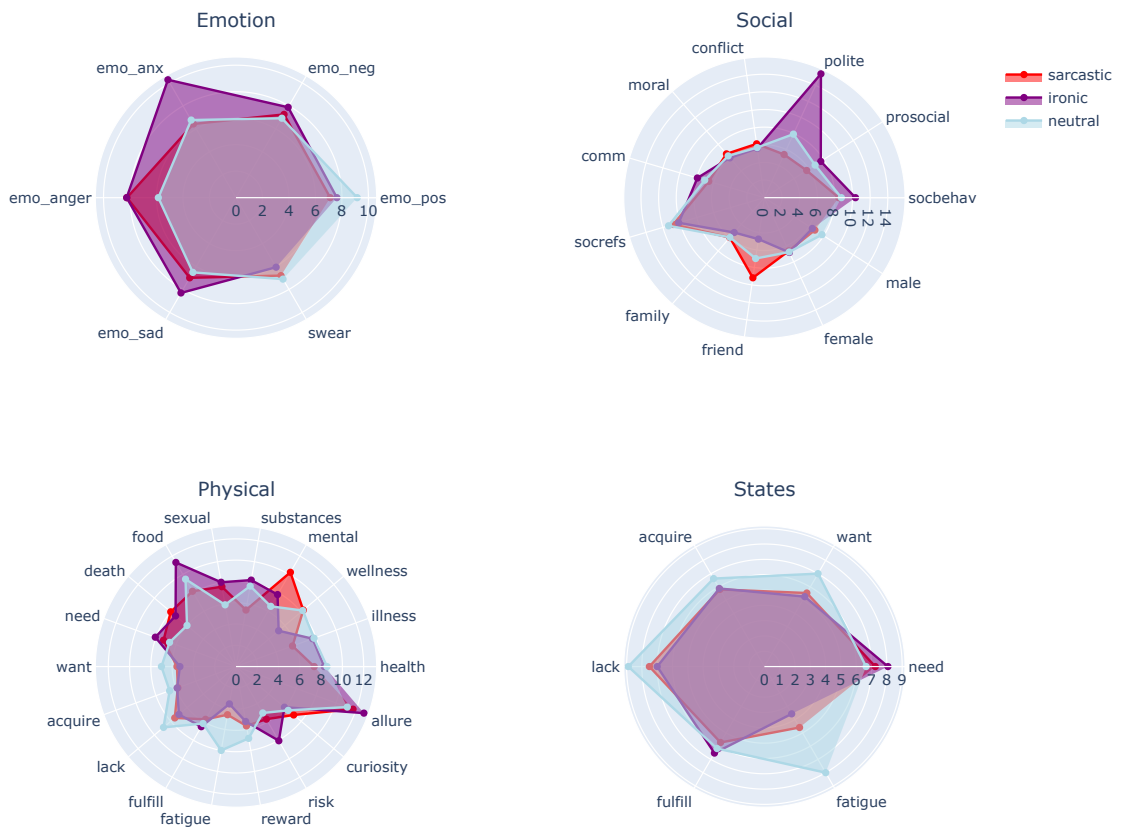


Figure 10: Spider plot for average non-zero LIWC values across notable categories for neutral, ironic, and sarcastic datasets.

## E Augmentation setup

### E.1 Prompt examples

Table 7 shows the prompts used for the data augmentation.

| Task    | Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Irony   | <p>Your task is to generate 9 ironic Tweets.</p> <p><b>Irony is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Ironic tweets contradict the state of affairs but are not obviously critical toward any particular addressee.</b></p> <p><i>The Tweets should be written in unstructured Twitter style — no formal grammar, proper capitalization, or structured sentences required. Focus on mimicking the style of the given tweet examples, such as mixing formal and informal grammar, structured and unstructured text, and occasionally using emojis and URLs. Pay attention to syntactic properties like ellipses and punctuation. Minimize the use of interjections and conversational phrases (e.g., 'love', 'nothing like', 'just', 'woke up', 'amazing', etc.).</i></p> <p>Separate each new instance with ordinal numbers.</p> <p>Examples:<br/>           - &lt;example tweet&gt;<br/>           - &lt;example tweet&gt;<br/>           - &lt;example tweet&gt;</p> <p>Ironic Tweets:<br/>           1.</p>                |
| Sarcasm | <p>Your task is to generate 9 sarcastic Tweets.</p> <p><b>Sarcasm is a rhetorical device where the intended meaning is opposite to the literal meaning, often used to mock, criticize, or highlight contradictions in a humorous or satirical way. Sarcastic tweets contradict the state of affairs, are directed towards an addressee and express a critical attitude.</b></p> <p><i>The Tweets should be written in unstructured Twitter style — no formal grammar, proper capitalization, or structured sentences required. Focus on mimicking the style of the given tweet examples, such as mixing formal and informal grammar, structured and unstructured text, and occasionally using emojis and URLs. Pay attention to syntactic properties like ellipses and punctuation. Minimize the use of interjections and conversational phrases (e.g., 'love', 'nothing like', 'just', 'woke up', 'amazing', etc.).</i></p> <p>Separate each new instance with ordinal numbers.</p> <p>Examples:<br/>           - &lt;example tweet&gt;<br/>           - &lt;example tweet&gt;<br/>           - &lt;example tweet&gt;</p> <p>Sarcastic Tweets:<br/>           1.</p> |

Table 7: Prompts for 3-shot setup used to generate ironic and sarcastic tweets. The **bolded** text provides the **definition**, while the *italic* portion covers *stylistic guidance*. For the 1-shot setup, the only difference is the requested number of generated tweets in the output (3 instead of 9) and the number of example tweets in the prompt. In ablated versions, definition, style, or both were omitted by removing the respective parts.

## E.2 Human judgment study

Table 8 shows the guidelines used in the annotation.

| Task               | Guidelines                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data detectability | <p><b>IRONIC</b></p> <p><b>Definition:</b><br/>Irony occurs when the tweet’s intended meaning is opposite to its literal meaning. It often emphasizes contradictions or absurdities in a humorous or satirical way.</p> <p><b>Usually, there is no clear target or addressee</b> of the criticism. If a target exists, the tone remains lighthearted and satirical.</p> <p><b>Examples of ironic tweets:</b></p> <ol style="list-style-type: none"> <li>1. <i>See Brexit is going well.</i></li> <li>2. <i>I just absolutely LOVE how I’ve got to work outside for the next 3 days in the heatwave.</i></li> <li>3. <i>My eldest is having a wild Friday night out. She’s going to bingo.</i></li> </ol> <p><b>SARCASTIC</b></p> <p><b>Definition:</b><br/>Sarcasm is a form of irony used to mock, ridicule, or convey contempt. While sarcastic tweets also contradict reality, they are marked by a sharper tone and are typically directed at a specific individual or entity (e.g., a person, brand, or group).</p> <p><b>Examples:</b></p> <ol style="list-style-type: none"> <li>1. <i>Loving season 4 of trump does America. Funniest season yet #DonaldTrump #Trump #MAGA #MAGA2020</i></li> <li>2. <i>Wouldn’t it be cool if I could spontaneously combust.</i></li> <li>3. <i>When your @Apple delivery gets stolen by a UK Mail delivery driver and the police know who it is but I still don’t get taken seriously</i></li> </ol> <p><b>NEUTRAL</b></p> <p><b>Definition:</b><br/>Tweets that are literal, informative, express emotion without irony, or do not involve contradiction, criticism, or mockery.</p> <p><b>Examples:</b></p> <ol style="list-style-type: none"> <li>1. <i>Going to the gym later today.</i></li> <li>2. <i>Thanks to everyone who came to the event!</i></li> <li>3. <i>It’s hot outside, but I’m staying hydrated.</i></li> </ol> |
| Data authenticity  | <p>Your task is to read each text sample and label it as either:</p> <ul style="list-style-type: none"> <li>• <b>Original</b> – Written by a human (obtained via Twitter/X)</li> <li>• <b>Augmented</b> – Generated by a language model (LLM)</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |

Table 8: Annotation guidelines for human judgement study.

## F Multi-aspect score comparison

| gpt4   |       |       |       | llama3 |       |       |       |
|--------|-------|-------|-------|--------|-------|-------|-------|
| Config | F1    | Div.  | Ling. | Config | F1    | Div.  | Ling. |
| 1X     | 0.720 | 0.775 | 0.085 | 3X     | 0.780 | 0.813 | 0.075 |
| 3D     | 0.719 | 0.799 | 0.073 | 1D     | 0.758 | 0.792 | 0.068 |
| 1D     | 0.713 | 0.775 | 0.085 | 3D     | 0.749 | 0.811 | 0.076 |
| 3X     | 0.703 | 0.818 | 0.078 | 3S     | 0.753 | 0.845 | 0.070 |
| 1DS    | 0.696 | 0.818 | 0.099 | 3DS    | 0.741 | 0.825 | 0.073 |
| 3DS    | 0.680 | 0.837 | 0.085 | 1X     | 0.722 | 0.803 | 0.079 |
| 3S     | 0.671 | 0.846 | 0.098 | 1DS    | 0.723 | 0.830 | 0.076 |
| 1S     | 0.594 | 0.826 | 0.102 | 1S     | 0.724 | 0.844 | 0.070 |

Table 9: Score comparison across performance ( $F1$ ), diversity ( $Div.$ ), and linguistic features ( $Ling.$ ) for irony. Cell shading intensity reflects relative score quality per task per model (darker = better; for  $Ling.$ , lower is better).

# Author Index

- Ahmad, Irfan, 52  
Alhetelah, Bushra, 52  
Alvisi, Lorenzo, 13  
Aranberri, Nora, 123
- Barić, Ana, 305  
Bezobrazova, Anastasiia, 107  
Bizzoni, Yuri, 146
- Cao, Rui, 289  
Cima, Lorenzo, 13  
Cresci, Stefano, 13
- Deng, Zhenyun, 289  
Ding, Zifeng, 289
- Enevoldsen, Kenneth, 146
- Feldkamp, Pascale, 146  
Filatova, Elena, 1, 60  
Flek, Lucie, 222
- Gangopadhyay, Surupendu, 252  
Greschner, Lynn, 265
- Hoste, Veronique, 161  
Hovy, Dirk, 83
- Iaroshenko, Polina V., 96
- Jafari, Sadegh, 161  
Jain, Shweta, 1, 60  
Jakob, Charlott, 172
- Karimi, Akbar, 222  
Klapper, Emily, 60  
Klinger, Roman, 73, 265
- Lefever, Els, 161  
Levitan, Sarah Ita, 1, 60  
Loukachevitch, Natalia V, 96  
Lyngbaek, Laurits, 146
- Majer, Laura, 305  
Markov, Ilia, 134  
Marreddy, Mounika, 37  
McMurry, Ian W., 24
- Mehler, Alexander, 37  
Mehrabani, Mahnoosh, 252  
Monfared, Mohammad Hossein Akbari, 222  
Münker, Simon, 208
- Neves, Pia Wenzel, 172  
Nielbo, Kristoffer, 146  
Nozza, Debora, 83
- Ollagnier, Anaïs, 235  
Orasan, Constantin, 107
- Plaza-del-Arco, Flor Miriam, 83  
Plum, Alistair, 208  
Puvača, Bojan, 305
- Rettinger, Achim, 208  
Rezapour, Rezvaneh, 278  
Rooein, Donya, 83
- Sandalj, Florijan, 305  
Schlichtkrull, Michael Sejr, 289  
Schmitt, Vera, 172  
Schouten, Stefan F., 134  
Schwager, Nils, 208  
Schäfer, Johannes, 73  
Sharma, Vivek, 1, 60  
Shokri, Mohammad, 1, 60  
Sokova, Daria, 107  
Šnajder, Jan, 305
- Tardelli, Serena, 13  
Teo, Joycelyn, 289  
Tesconi, Maurizio, 13
- Unković, Ivan, 305
- Verma, Bhuvanesh, 37  
Vlachos, Andreas, 289  
Vossen, Piek, 134
- Wagner, Janne, 73  
Weber, Sabine, 265  
Wen, Ximing, 278