NLP+CSS 2026

The Seventh Workshop on Natural Language Processing and Computational Social Science

Proceedings of the Workshop

July 3, 2026

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 979-8-89176-426-2

Introduction

Welcome to the Seventh Workshop on Natural Language Processing (NLP) and Computational Social Science (CSS)! This workshop continues a successful series, with many interdisciplinary contributions to make NLP techniques and insights standard practice in CSS research, as well as improve NLP through insights from the social sciences. We received 76 submissions and after a rigorous review process by our committee, we accepted 23 entries, 19 archival and 4 non-archival. We also hosted a shared task for the first time this year in partnership with the Opioid Industry Document Archive (OIDA), and we are delighted to include two shared task submissions in the proceedings. Our workshop program also includes keynote talks by three outstanding scholars: Philip Resnik, a professor in the Department of Linguistics and Institute for Advanced Computer Studies at the University of Maryland, Lucy Li, a professor in the Department of Computer Sciences at University of Wisconsin-Madison, and R. Stuart Geiger a professor at the University of California, San Diego with appointments in the Department of Communication and the Halıcıoğlu Data Science Institute.

We would like to thank the Program Committee members who reviewed the papers this year. They did a wonderful job providing high-quality reviews, and particularly helping with last minute emergency reviews. We would also like to thank the workshop participants for the opportunities to connect (or reconnect) and learn from each other.

Dallas Card, Anjalie Field, Katherine Keith, and Julia Mendelsohn (Co-Organizers)

Organizing Committee

Program Committee

Dallas Card, University of Michigan
Anjalie Field, Johns Hopkins University
Katherine Keith, Williams College / Cohere
Julia Mendelsohn, University of Maryland

Program Committee

Reviewers

Anurag Acharya
Daniel Acuna
Jisun An
Aparna Ananthasubramaniam
Maria Antoniak
Elliott Ash
Nicolas Audinet de Pieuchon
Blanca Calvo Figueras
Danqing Chen
Hong Chen
Hongyu Chen
Andres Cruz
Aida Davani
Felix Drinkall
Jonathan Dunn
Agnieszka Falenska
Mark Finlayson
Bijean Ghafouri
Andrew Halterman
Endre Hamerlik
Kevin S. Hawkins
Nathan Oken Hodas
Alexander Miserlis Hoyle
Weihang Huang
Abraham Israeli
Jonathan Ivey
Cheonkam Jeong
Connor Thomas Jerzak
Richard Johansson
Moa Johansson
Kristen Johnson
Kenneth Joseph
David Jurgens
Elma Kerz
Rasul Khanbayov
Philipp Koehn
Frauke Kreuter
Haewoon Kwak
Alexandria Leto
Carlo Lipizzi
Chang Liu
Li Lucy
Tessa Masis
Ben Miller
David Mimno
Joel Mire

Ashley Moran
Nazia Nafis
Jason K. Nam
Bill Noble
Brianna O'Boyle
Sebastian Padó
Thierry Poibeau
Prateek Puri
Sushrita Rakshit
Christopher Rashidian
Christoph Rauchegger
Manoel Horta Ribeiro
Anthony Rios
Martin Ruskov
Germans Savcisens
Djamé Seddah
Sadat Shahriar
Bangzhao Shu
Emily Silcock
Dan Simonson
Akshay Singh
Mingyao Song
Andreas Spitz
Ian Stewart
Martin Takáč
Samia Touileb
Takehito Utsuro
Gisela Vallejo
Vasudha Varadarajan
Rob Voigt
Charles Welch
Yinuo Xu
Jinghua Xu
Li Yue
Lechen Zhang
Yifan Zhang
Mian Zhong
Viktoria Zlomanova

Table of Contents

Prompt Perturbations Reveal Human-Like Biases in Large Language Model Survey Responses

Jens Rupprecht[1], Georg Ahnert[1], and Markus Strohmaier[1,2,3]
[1]University of Mannheim, Mannheim
[2]GESIS – Leibniz Institute for the Social Sciences, Cologne
[3]Complexity Science Hub, Vienna
[firstname.lastname]@uni-mannheim.de

Abstract

Large Language Models (LLMs) are increasingly used as proxies for human subjects in social science surveys, but their reliability and susceptibility to known human-like response biases, such as *central tendency, opinion floating* and *primacy bias* are poorly understood. This work investigates the response robustness of LLMs in normative survey contexts—we test 18 LLMs on questions taken from the World Values Survey (WVS), applying a comprehensive set of ten perturbations to both question phrasing and answer option structure, resulting in over 334,800 simulated survey interviews. In doing so, we not only reveal LLMs' vulnerabilities to perturbations but also show that almost all tested models exhibit a consistent *recency bias*, disproportionately favoring the last-presented answer option. While larger models are generally more robust, all models remain sensitive to semantic variations like paraphrasing and to combined perturbations. This underscores the critical importance of prompt design and robustness testing when using LLMs to generate synthetic survey data.

1 Introduction

Problem Large Language Models (LLMs) are increasingly being used as proxies for human subjects in social science research, particularly to generate synthetic responses to survey questions (Argyle et al., 2023; Bisbee et al., 2024, inter alia). This application holds promise in augmenting or replacing costly human data collection. Still, the reliability of these synthetic respondents and the extent of overlap with human responses and response biases remain open questions. In particular, research in survey methodology has found that human responses are sensitive to subtle variations in question and answer phrasing that lead to well-known **response biases** (Krosnick, 1991) and it remains unclear whether LLMs, trained on vast amounts of human text, exhibit the same vulnerabilities.

Approach We present a large-scale empirical study, with 334,800 survey interviews in total, investigating the response behavior and robustness of 18 different LLMs to normative questions derived from the World Value Survey (WVS; Haerpfer et al., 2022). By developing and applying a comprehensive set of ten perturbations in both the structure of the answer options (Table 1), such as typos or synonyms, and in the question phrasing (Table 4), such as, e.g., changes to response order or scale structure, we answer the following research questions.

1. Do prompt perturbations negatively affect the **robustness** of LLMs when answering closed-ended, normative survey questions?
2. Do LLMs exhibit **human-like response biases** when answering closed-ended, normative survey questions?

Contribution Next to the perturbation framework, we provide a detailed analysis of LLM response robustness, showing that while some models are more robust than others (e.g. `Llama-3.3-70B-Instruct` and `Gemini-1.5-Pro`), all are susceptible to specific perturbation types. Most notably, we find a consistent *recency bias* across almost all tested models, where the last-presented answer option is disproportionately favored by up to 20 times, and even the largest models remain sensitive to changes in question phrasing.

These findings underscore the importance of careful prompt and Q&A design when using LLMs for synthetic survey responses. The perturbation framework serves as a useful baseline for evaluating robustness in survey contexts, and we make the full Q&A dataset publicly available for benchmarking newer or other LLMs. We present an overview of which LLMs exhibited human-like survey response biases (Table 11).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 1–21

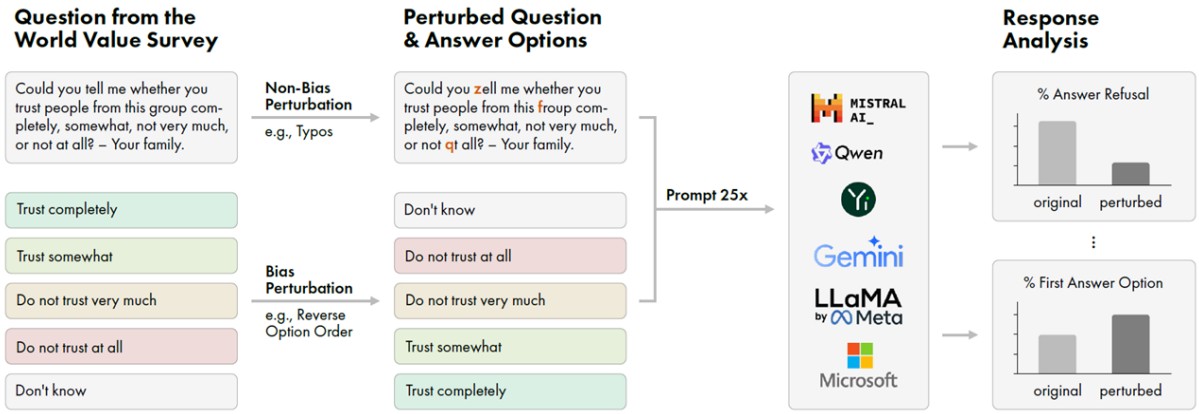


Figure 1: **The Interview Process.** The figure displays an example of a answer option perturbation (a *bias perturbation*, e.g. reversed option order) and an question perturbation (a *non-bias perturbation*, e.g. typos in the question). Each model is prompted 25 times with every perturbation as well as the original Q&A phrasing. All responses are collected, processed and statistically analyzed.

2 Related Work

Our work builds on two main streams of research: (1) survey methodology from the social sciences, which documents human response biases, and (2) recent studies in computer science on the robustness and biases of LLM's synthetic survey response generation.

Human Survey Response Biases Research in the social sciences has shown that how a survey question is asked can be as important as what is asked. Respondents often engage in "satisficing" rather than "optimizing", choosing a satisfactory answer with minimal cognitive effort instead of carefully formulating an optimal one (Krosnick, 1991). This can lead to systematic biases. For example, the order in which the answer options are presented can induce *primacy* (favoring early options in visual surveys) or *recency* (favoring later options in oral surveys) biases (Krosnick and Alwin, 1987). The presence or absence of a middle option or a "don't know" category can trigger a *central tendency bias* or *opinion floating*, respectively (Hollingworth, 1910; Koch and Blohm, 2016). In the first, if a central category is available on the answer scale, humans tend to choose the central category, whereas *opinion floating* indicates that responses are redistributed to central categories if a refusal category is missing (Tjuatja et al., 2024). In addition, *priming* effects, where the preceding context influences subsequent responses, are a well documented phenomenon (Bargh et al., 1996). We draw on these past findings to design perturbations testing whether LLMs exhibit similar human-like response patterns.

LLMs as Survey Respondents Recent studies explored LLMs as substitutes for human survey participants to generate synthetic data. They found that LLMs can replicate average public opinion on political topics, but often with less variance than human samples (Argyle et al., 2023; Bisbee et al., 2024; von der Heyde et al., 2025; Dominguez-Olmedo et al., 2024, inter alia). Others have found that LLM responses can be sensitive to prompting, revealing cultural and demographic biases (Geng et al., 2024). Laverghetta et al. (2022) found that LLMs can produce similar responses to human participants on diagnostic items, for example, in linguistic test. Contrary to this finding, Sühr et al. (2025) found that LLMs' responses to personality tests systematically deviate from human responses. This implies that the results of these tests cannot be interpreted in the same way. However, Huang et al. (2024) identified that LLMs have the potential to represent different personalities with specific prompt instructions. In addition, they found that response patterns of multiple LLMs showed consistency in responses to the Big Five Inventory, indicating a satisfactory level of reliability.

Our work is related to that of Tjuatja et al. (2024), who were among the first to systematically explore human-like response biases in LLMs. They investigated acquiescence, response order, opinion floating, and scale structure effects. Our study extends their work by: (1) using a different, globally diverse survey (the World Values Survey); (2) testing a wider range of LLMs, such as Gemini-2.5-Pro; and (3) incorporating a broader set of perturbations

on both answer and question phrasing, such as *keyboard typos, paraphrasing, synonyms, priming* as well as a combined *interaction* of two perturbations.

LLM Robustness to Perturbations Other researchers have evaluated the general robustness of LLMs to noisy or varied inputs on different tasks. They have shown that even state-of-the-art models can be sensitive to minor changes in the prompt. These perturbations range from the character level, such as typos created by swapping, inserting, or replacing letters (Moradi and Samwald, 2021; Gan et al., 2024), to word- or sentence-level, such as replacing words with synonyms or paraphrasing entire sentences (Qiang et al., 2024). A common finding is that character-level noise can significantly degrade performance, even in large models (Gan et al., 2024). The combination of multiple perturbations can even have a more negative effect (Dong et al., 2023). Although this research has primarily focused on knowledge-based or reasoning tasks, we adapt these perturbation techniques to the context of normative surveys to assess response stability where no single "correct" answer exists.

Evaluation and Prompting Finally, our work is guided by research on the identified ways for evaluating LLMs on multiple-choice tasks. Studies have shown that evaluation results can be highly sensitive to prompt format, e.g. if LLMs face an open- or closed-ended response, and forcing technique. However, forcing a model to choose from a predefined set of options is often necessary to obtain valid responses, as unconstrained responses can differ substantially (Röttger et al., 2024). The returned response labels might differ significantly when a LLM has the option to generate text output before returning the response label due to their auto-regressive nature. Furthermore, relying on the model's first predicted token can misrepresent its full textual output (Wang et al., 2024). Therefore, we include two LLMs with reasoning capabilities to compare their performance to non-reasoning models.

3 Methods

First, we select a subset of 62 questions representing a sample of different thematic categories, each question in the category sharing the same answer options. These normative, value-oriented Q&A pairs are taken from the WVS's core variables (Haerpfer et al., 2022), excluding all sociodemographic variables. Second, we perform the ten perturbations mentioned in Section 3.2 each Q&A pair.

Figure 1 illustrates two exemplary perturbations and the interview process. In total, we perform five perturbations on the answer options as well as five perturbations on the question phrasing of the chosen subset questionnaire. We further include one interaction of two perturbations, one on the answer option and one on the question.

Third, we carry out interviews with the original and each perturbed Q&A pair 25 times with 18 different LLMs. In total, we conducted 334,800 interviews, 18,600 with each model. Last, we compare the distributions of the responded labels for all perturbations response consistency, given the same interview setting, through entropy and calculate the Kullback-Leibler divergence (KL divergence) to compare the baseline response distribution on the original to the perturbed Q&A pairs. *Primacy bias* is further examined by comparing the response frequencies of the first and last answer options in the list, whereas *opinion floating bias* and *central tendency bias* are tested by checking the shift of responses toward or away from the center of the answer option scale.

3.1 Experimental Setup

Survey Data The questions and answer options are sourced from the WVS Wave 7 (2017-2022), a comprehensive cross-national survey on human beliefs and values (Haerpfer et al., 2022). The 259 core WVS Q&A pairs represent 10 distinct thematic categories, including *Trust in People*, *Confidence in Institutions*, *Moral Justifiability*, and *Perception of Democracy*, ensuring a diverse range of topics and answer scale formats (e.g., 3-point to 10-point scales). We used stratified sampling to select six to seven Q&A pairs per thematic category, resulting in a total of 62 Q&A pairs.

Models To ensure that our findings are not specific to a single model architecture or developer, we selected 18 instruction-tuned LLMs, varying in size, developer, and origin, including two with reasoning capabilities. This selection aims to establish external validity for our results and includes proprietary and open-source LLMs. The following models were interviewed:

- **Llama** (in the following tables abbreviated as L3.3-70B (Meta, 2025b), L3.1-8B (Grattafiori

et al., 2024), L3.2-3B, L3.2-1B (Meta, 2025a) represent their respective Instruct versions),

- **Qwen** (Q2.5-7B for Qwen-2.5-7B-Instruct; Q3-0.6B through Q3-32B for Qwen3 versions; Q3-30BT for Qwen3-30B-A3B-Thinking-2507, (Yang et al., 2025b,a)),
- **Gemini** (G1.5P for 1.5-Pro, G2.5P for 2.5-Pro, G2.5F for 2.5-Flash, (Georgiev et al., 2024)), and
- **Others** (M7B for Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), P3.5M for Phi-3.5-mini (Abdin et al., 2024), Y1.5-6B for Yi-1.5-6B-Chat, (Young et al., 2024)).

We listed the specific model IDs in Appendix Table 3.

3.2 Perturbation Design

We designed two categories of perturbations to test model robustness: (1) bias-inducing alterations to the answer options, based on survey methodology research that are known to induce biased responses in humans (Tjuatja et al., 2024), and (2) non-bias alterations to the question phrasing, mimicking common textual variations and errors. Table 1 and Table 4 provide examples as well as references for all perturbations. For each of the 62 Q&A pairs, we created the following ten perturbed versions.

Bias Perturbations These five perturbations manipulate the answer choices provided to test for known survey response biases identified in human subjects and presented in Section 2. Therefore, we call them *bias perturbations* (Tjuatja et al., 2024, p.3).

- **(1) Reversed Response Order:** The order of answer options is reversed (e.g., a scale from '1: Very important' to '5: Not important' becomes '1: Not important' to '5: Very important').
- **(2) Missing Refusal Option:** The "Don't know" or refusal option is removed from the list of choices.
- **(3) Odd/Even Scale Transformation:** For scales with an even number of options, we use `Gemini-1.5-flash` to generate a semantically appropriate middle category, transforming it into an odd-numbered scale (e.g., adding 'Neutral'). Conversely, for odd-numbered scales, we remove the middle category to create an even-numbered scale and adjust the integer label.
- **(4) Priming Suffix:** A sentence designed to elicit a response is appended to the prompt after the answer options: *'This is very important to my research! You better do not refuse the answer.'*

Non-Bias Perturbations These five perturbations modify the question text to assess robustness to stylistic variations and typos. Typically, humans are unaffected by such subtle changes in the question phrasing and are still able to understand the question's meaning (Tjuatja et al., 2024, p.3). Therefore, we call them *non-bias perturbations*.

- **Typographical Errors:** We introduce three types of typos: **(5) Key Typo** (replacing a character with a random one), **(6) Letter Swap** (swapping two letters in a random word), and **(7) Keyboard Typo** (replacing a character with an adjacent one on a QWERTY keyboard).
- **Semantic Variations:** We use `Gemini-1.5-flash` to create two semantic variations while preserving the original meaning: first, by **(8) Synonym Replacement** where five words in the original question are replaced with synonyms. Second, through **(9) Paraphrasing** the entire question is rephrased.

We manually validated all LLM-generated perturbations (paraphrases, synonyms, odd-scale options) on our 62-question subset to correct errors and ensure their semantic integrity.

Last, we introduce an **(10) Interaction Effect** to study the impact of not only one, but two perturbations. Thus, we created one additional condition that pairs a paraphrased question with reversed-order answer options.

3.3 Interview Procedure and Data Collection

Prompting To ensure internal validity, we used a single, consistent prompt structure for all interviews. The prompt was designed to be clear and direct, instructing the model to select a single option and return only its label.

Answer the following question: {question}
This is a list of possible answer options:
{answer_options}
You must pick one of the answer options. Only answer with the label.

Taking into account the findings of Röttger et al. (2024) we use *forced choice* prompts that provided the most valid responses in their experiments,

Type	Perturbation	Question	Answer Options	Bias and Reference
Original	**Original**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all', "-1=Don't know"]	(Haerpfer et al., 2022)
Bias Perturbations	**(1) Reversed Response Order**	For each of the following aspects, indicate how important it is in your life. Would you say it is very important, rather important, not very important or not important at all? Family	["-1=Don't know", '4=Not important at all', '3=Not very important ', '2=Rather important ', '1=Very important ']	**Primacy Bias** (Tjuatja et al., 2024; Krosnick and Alwin, 1987; Kampen, 2007; O'Halloran et al., 2014)
	(2) Missing Refusal Option		['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all']	**Opinion Floating Bias** (Schuman and Presser, 2000; Tjuatja et al., 2024)
	(3) Odd/Even Scale Transformation		['1=Very important ', '2=Rather important', '3=Neutral', '4=Not very important ', '5=Not important at all', '-1=Don't know']	**Central Tendency Bias** (Hollingworth, 1910; Cronbach, 1946; Aston et al., 2021; Crosetto et al., 2020)
	(4) Priming Suffix		[1=Very important , 2=Rather important , 3=Not very important , 4=Not important at all, -1=Don't know]This is very important to my research! You better do not refuse the answer.	**Priming Effect** (Bargh et al., 1996; Higgins, 1996; Weingarten et al., 2016; Li et al., 2023)

Table 1: **An exemplary perturbation scheme showing the original question, the bias perturbations and the the hypothesized human-like response bias.** The Q&A example is taken from the thematic category "Importance of Life Aspects" from the WVS wave 7. Bias perturbations have a constant question phrasing and varying answer options. The remaining *Non-Bias Perturbations* are listed in Table 4.

which means that the model correctly chooses one available option from the given answer options. In an unforced prompt setting, LLMs did not follow the instructions and only rarely returned valid response labels, making a statistical evaluation of the results infeasible. Preliminary tests revealed that especially smaller LLMs often fail to perfectly follow the instruction to "answer only with the label", as they return conversational filler or explanations alongside their choice.

Data Collection Each of the 18 models was presented with 12 experimental conditions (1 original + ten perturbations, where for the perturbation *Odd/Even Scale Transformation* one run was done for the odd and even scale scenario) for each of the 62 selected WVS questions. To obtain a stable distribution of responses and enable statistical analysis, we repeated each unique model-Q&A-perturbation combination 25 times. This resulted in a total of $18 \times 62 \times 12 \times 25 = 334,800$ interviews.

Response Extraction and Validation To ensure accurate data for analysis, we developed a robust extraction pipeline. We compared two main approaches. First, `Gemini-1.5-Pro`, `Llama-3.1-8B`, and `Qwen2.5-7B` were prompted, and a regular expression was designed to extract the answer labels. Based on multiple conditions, e.g. if the given answer label is part of the original answer options or that only one response is provided, the methods should highlight which technique is the most promising in extracting valid responses and handling possible edge cases of model responses.

We manually labeled these extraction methods on a random sample of responses for validation. The LLM-based methods achieved accuracies between 77% and 97.5%, with the largest model `Gemini-1.5-Pro` performing best. However, our refined regular expression achieved the overall best extraction success on the validation set as it correctly extracted all responses in our validation set. Consequently, we used this regular expression to process all remaining 316,200 model responses.

4 Results

This section presents the results of our experiments, focusing on two key research questions: (1) LLMs' general robustness to various input perturbations, and (2) their susceptibility to human-like survey response biases revealed by interviews with bias prompt perturbations. In addition, we also investigate the models' general adherence to interview instructions.

4.1 Robustness to Question and Answer Perturbations (RQ1)

We distinguish between response robustness (the tendency to maintain a similar answer distribution under perturbation, measured by KL divergence) and response consistency (the tendency of a model to give the same answer to the same prompt, measured by entropy). A KL divergence of zero indicates a perfect match and thus full robustness against the input perturbation, whereas a high entropy value indicates very inconsistent response behavior.

Effect of Model Size on Robustness First, when assessing robustness to perturbations, we found a clear relationship with model size: *larger models tend to be more robust*. Table 2 and 5 show the percentage of questions for which the models produced a perfectly identical response dis-

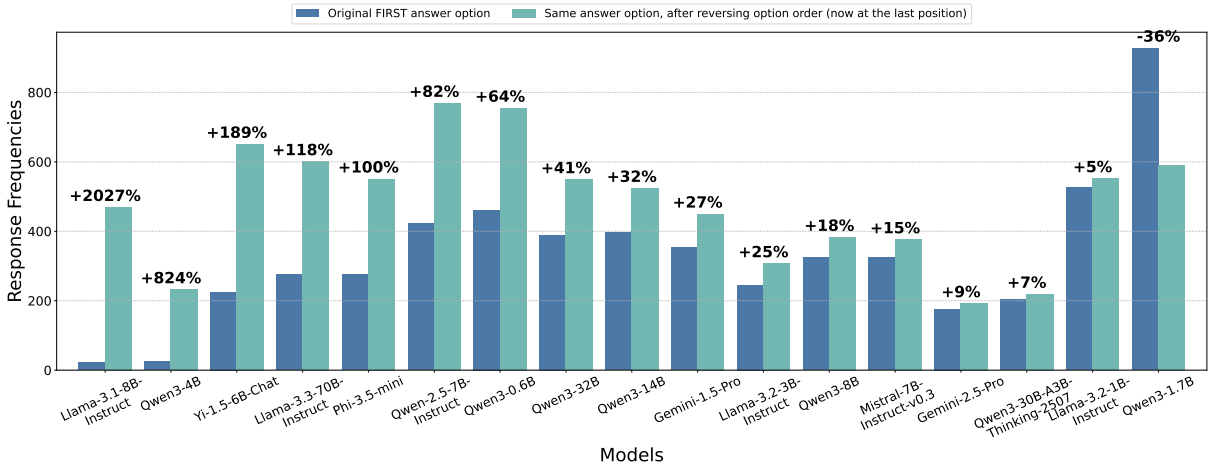


Figure 2: **Evidence of recency bias across all models.** The bars show the frequency of choosing the same answer option (e.g., "Very important") when it is presented first vs. last. Almost all models are more likely to select an option when it appears at the end of the list.

tribution (KL divergence = 0) despite perturbations. `Llama-3.3-70B` and `Gemini-1.5-Pro` were the most robust, often replicating their original answers in over 50% of cases. The smaller `Llama` models were the least robust, with `Llama-3.2-1B` perfectly replicating its answers in fewer than 5% of cases on average. This suggests that scale is a key factor in achieving stable response behavior in synthetic response generation. The reason why the responses for `Gemini-2.5-Pro` and `Flash` are not robust is most likely due to the fact that—despite the same experimental setup—the new Gemini series refuses value-oriented questions much more often than its previous series (see Table 7).

Second, we found that model size is in an inverse relationship with response consistency; smaller models exhibited higher entropy and standard deviation when asked the same question multiple times, indicating more random response behavior (Table 9).

Effect of Perturbation Type on Robustness Further, Tables 2 and 5 highlight the share of fully robust responses (KL divergence = 0) across all questions by perturbation and LLM. It shows that some perturbations had a greater impact on robustness across all models.

- **Combined Perturbations:** The interaction of two perturbations (paraphrased question + reversed answers) has the most bewildering effect on the responses, causing the lowest robustness scores for all models except `Phi-3.5-mini`.
- **Semantic vs. Lexical Changes:** Paraphrasing the question reduced robustness more than replacing individual words with synonyms in most LLMs. These findings are consistent with Moradi and Samwald (2021) who found that models trained on larger corpora are more robust when words are replaced by their synonyms.
- **Typographical Errors:** Randomly replacing characters (*Key Typo*) or using adjacent keys (*Keyboard Typo*) was more robustness-harming than simply swapping two letters within a word (*Letter Swap*). We assume that the training text corpora potentially contain more words with accidental letter swaps than random typos, and therefore LLMs might be more resilient against these perturbations.
- **Answer Option Changes:** Reversing the answer scale or changing it from odd to even (or vice versa) had a more negative impact on the robustness of responses than removing the refusal option or adding emotional priming.

Effect of Answer Option Scale Length We also observed that robustness is affected by the complexity of the task. For nearly all models, the share of fully robust responses decreased as the length of the answer scale, i.e. answer options, increased. For example, models were less likely to replicate their exact response distributions on a 10-point scale compared to a 4-point scale, indicating that a larger decision space can make LLMs more susceptible to perturbations. Figures 4 and 5 suggest that for most LLMs, except Gemini-1.5-pro, the size of the answer option scale has an impact on response robustness comparing the share of fully robust responses on e.g. the 4- and 10-point scale. This suggests

Model	(1) Reversed Answer	(2) Missing Refusal	(3) Even Scale	(4) Priming Suffix
Llama Family				
L3.3-70B	0.50	0.73	**0.60**	**0.82**
L3.1-8B	0.08	0.39	0.27	0.35
L3.2-3B	0.10	0.11	0.16	0.10
L3.2-1B	0.00	0.08	0.03	0.11
Qwen Family				
Q3-32B	0.27	0.35	0.19	0.31
Q3-30BT	0.32	0.23	0.15	0.23
Q3-14B	0.53	0.60	0.48	0.53
Q3-8B	0.34	0.53	0.26	0.39
Q2.5-7B	0.32	0.48	0.45	0.50
Q3-4B	0.24	0.50	0.35	0.31
Q3-1.7B	0.34	0.66	0.47	0.52
Q3-0.6B	0.02	0.03	0.02	0.00
Gemini Family				
G1.5P	**0.69**	0.76	0.55	0.74
G2.5P	0.32	0.11	0.26	0.00
G2.5F	0.15	0.16	0.16	0.00
Others				
M7B	0.68	**0.81**	0.53	0.74
P3.5M	0.53	**0.81**	0.45	0.79
Y1.5-6B	0.47	0.68	0.55	0.52

Table 2: **Share of Fully Robust Responses by Perturbation Type and Model** (↑). The models are grouped by model family.

that the larger the answer scale, the less likely models can reproduce the responses they gave in the original Q&A phrasing, under perturbed settings.

4.2 Evidence of Human-like Survey Biases (RQ2)

With many of the perturbations, we are able to go beyond LLMs robustness and consistency and also analyze whether LLMs exhibit human-like survey response biases. We find evidence for some human-like biases.

Recency Bias Contrary to the initial hypothesized primacy bias, we found *indications of a recency bias in 17 of the 18 models tested.* When we reversed the order of the answer scale, the probability to choose the first option plummeted, while the probability to choose the last option (which is the semantically identical first option in the original Q&A) increased strongly, *ceteris paribus*. As shown in Figure 2, this effect was substantial, with the selection frequency of the semantically same option increasing by more than 20 times for `Llama-3.1-8B` when moved to the last position, while all other configurations, such as question and prompt phrasing, were kept constant. This indicates that LLMs, similar to human respondents in oral surveys, might overemphasize the final options they process. However, LLMs with activated reasoning capabilities, such as `Gemini-2.5-Pro` and `Qwen3-30B.A3B-Thinking-2507`, mitigate this bias. The same answer option placed at the scale end is chosen just 0.07-0.09 times more often after reasoning.

Opinion Floating and Central Tendency The effects of removing the refusal option (*opinion floating*) or providing an explicit middle category (*central tendency*) were highly model-dependent, often correlated with model size (Tables 10a and 10b). For *opinion floating*, larger models like `Llama-70B`, `Gemini-1.5-Pro`, but also `Phi-3.5` were largely robust, showing minimal shifts in their response distributions. Smaller models, particularly `Qwen` and `Llama-8B`, showed a weak tendency to shift responses toward the scale's center when the refusal option was absent. Here, we expect that models redistribute their original non-responses to the center of the answer scale to maintain their indecisiveness, which is also known as opinion floating bias in humans.

Similarly, for *central tendency*, larger models (`Llama-70B`, `Gemini-1.5-Pro`, `Gemini-2.5-Flash`, `Mistral`) consistently shifted their mean response closer to the center across all scale types when an explicit middle option was provided compared to even answer option scales. However, smaller models, such as `Qwen3-0.6B`, `-1.7B`, `-4B` or `Phi-3.5-mini`, showed inconsistent effects or were completely unaffected. Further, we see an almost consistent response shift to the center across all except three models for medium-sized scales (four vs. five point Likert scales).

Binomial tests underlined that the middle option was selected significantly more often than expected under a uniform distribution, especially on larger scales (cf. Table 6). LLMs tend to choose the middle category significantly more often when the scale size increases from three to five and to 11 point Likert scales. All, except the two smallest Qwen3 models and `Llama-3.2-1B`, choose the middle category significantly more often than any other option when facing a eleven answer options.

Emotional Priming The impact of adding an emotional priming statement ("This is very important to my research!") was also model-dependent.

For larger models (Llama-70B, Gemini-1.5-Pro, Mistral), it either had no effect or slightly decreased the rate of refusal responses, suggesting they correctly interpreted the intent of the priming statement. Conversely, for the two Chinese models, Qwen2.5-7B and Yi-1.5-6B, the priming text even *increased* the share of refusal responses across most topics. No clear relationship can be drawn from these findings due to inconsistent behavior across models.

4.3 Interview Adherence and Refusal Rates

Overall, the models demonstrated high adherence to the prompt instructions, with an average of 96% of interviews yielding an extractable and valid answer that was part of the given answer options. However, performance varied significantly across models. Larger models such as Llama-3.3-70B and Gemini-1.5-Pro, but also Phi-3.5-mini and Mistral are very reliable response generators and followed the instructions well while returning little to no incorrect or no answer label. In contrast, other models, particularly smaller Llama models like Llama-3.2-3B (83.6%), Qwen2.5-7B, but also the two reasoning models, were more likely to produce invalid responses that did not follow instructions.

We combined invalid responses with explicit refusals (i.e., choosing the *Don't know* option) to measure overall non-response rates, as shown in Table 7. Llama-3.3-70B, Phi-3.5, and Mistral-7B consistently provided on-scale answers, with non-response rates typically below 10%. Conversely, Qwen2.5-7B and Llama-3.1-8B exhibited high non-response rates, often exceeding 30%.

In particular, we observed topic-specific sensitivity. For questions regarding the *Perception of Elections*, Qwen2.5-7B failed to provide a valid, on-scale answer in 91.3% of cases, even for the original, unperturbed questions. This might suggest the presence of strong content-based guardrails or restrictions in certain models (cf. Figure 8).

5 Discussion and Conclusion

Our experiments revealed that LLMs response robustness is negatively influenced by prompt perturbations when answering closed-ended survey questions (*RQ1*). the variety of perturbation allows us to gain insights into the robustness of LLMs as some models are more sensitive and some perturbations are more robustness-harming than others. For instance, swapping letters within a word has less negative impact than introducing random or keyboard-adjacent characters. This might be explained by the fact that letter swaps are more likely when typing and therefore might potentially take a greater part of training data (Dhakal et al., 2018). This possibly makes the LLM more resilient to this perturbation compared to exchanging characters with random others. Combining two types of perturbations has the strongest negative impact on robustness, whereas synonyms tend to be less confusing than paraphrasing.

Further, we found that perturbations can be an insightful approach to identify human-like survey response biases (*RQ2*). For example, the same answer option is more likely chosen if it is the last mentioned option than if it were the first answer option, holding all other specifications and phrasings constant. This consistent change in the response distribution to the last answer option suggests a *recency bias* rather than a primacy bias.

Although this is not valid across all inspected models, binomial tests reveal that most models choose the middle category more likely than the other categories. Thus, a *central tendency bias* could be identified for specific models across all scale types, whereas none of the LLMs consistently mirrors a *opinion floating bias*.

The findings emphasize the importance of the positioning of answer options when generating synthetic data. In addition, our results highlight the strong sensitivity of LLMs to simple prompt perturbation. Therefore, we strongly recommend researchers to consider prompt robustness checks when deploying closed-ended questions to LLMs. This is because (i) models show very different response behavior and robustness depending on their size, release date (e.g. Gemini-1.5-Pro vs. Gemini-2.5-Pro) and perturbation type, and (ii) LLM response biases are sometimes but not necessarily aligned with biases identified in humans.

Recommendations Based on our findings, we recommend researchers to:

- Use larger, non-reasoning, LLMs for overall better consistency and robustness in generating synthetic survey responses (cf. Tables 2 and 5)
- Reasoning seems to mitigate the *recency bias* identified in non-reasoning LLMs. However, it leads to more non-responses and refusals.
- Use smaller answer option scales for better reproducibility of results (cf. Figure 4).

- Reflect on the meaningfulness of adding a middle category. Including a middle category might steer some LLM responses to the center (cf. Table 10a).
- Reflect the meaningfulness of adding a refusal category. Adding a refusal category might highlight LLM guardrails or restrictions in some thematic areas, as the model can refuse to answer while still following the instructions as it returns a valid response label (cf. Figure 8).
- Use *forced-choice* prompts to generate high turnouts while also considering open-ended evaluation if sensible.

Limitations

This study investigates the robustness of LLM-generated survey responses when facing diverse prompt perturbations, but several methodological and conceptual limitations must be noted. The use of a multiple-choice format, originally designed for human respondents, imposes an artificial constraint on LLMs that typically work in open-ended contexts. As a result, the findings may not generalize to more naturalistic human-LLM interactions.

Although we constrained and validated the data augmentation process, relying on a LLM (`Gemini-1.5-flash`) for generating paraphrases risks semantic drift, as also noted by Qiang et al. (2024). More granular validation—e.g., with multiple human raters—could improve semantic reliability. In addition, perturbations were applied at a fixed intensity, limiting insight into how different degrees of linguistic noise affect model behavior.

Further constraints arise from our prompting and generation setup. The validation set for answer extraction was relatively small compared to the full dataset, so some extraction errors may remain. We also did not apply prompting strategies like persona prompting, shown to improve contextual consistency (Bisbee et al., 2024; Cho et al., 2024), nor used techniques such as *Chain of Thought* prompting. This could promote more deliberative responses instead of only the latent baseline behavior of LLMs. For example, one could use empirically grounded, survey-derived persona collections to gain more perspectivist synthetic survey responses (Rupprecht et al., 2026). Note that practitioners using demographic personas may observe different bias patterns, and that extending the perturbation framework to persona-conditioned settings is an important future direction. Moreover, our experiments focused exclusively on fine-tuned models, leaving open the question of how base models would behave under similar conditions. Additionally, a constant temperature setting restricted our ability to examine variability and creativity in the output.

Finally, reproducibility is another significant challenge. Closed-source LLMs can change without notice, altering response distributions over time and complicating replication efforts, as highlighted by Bisbee et al. (2024). This may have affected our `Gemini` results. Related work also shows that LLMs often offer contradictory answers to semantically equivalent questions when the format shifts from multiple choice, close-ended to an open-ended form (Röttger et al., 2024). Such response instability suggests that observed "attitudes" may be artifacts of prompt design rather than indicators of stable model beliefs or traits.

Ethical Considerations

Generating synthetic survey responses might be relevant in various domains and applied to different use cases, e.g. for pre-testing surveys. However, generating synthetic responses instead of the surveying a real, might result in over-reliance on synthetic responses. This can become risky when there is no ground truth data of the real target population available as the alignment of the artificial responses cannot be evaluated. Frequent reliance on artificial responses may normalize their use where human perspectives are irreplaceable (e.g. in policymaking or clinical trials). This risks sidelining real human voices in domains directly impacting human lives.

Researchers should also consider ethical evasion as one possible issue with synthetic survey responses. Synthetic respondents might be viewed as a way to bypass obligatory ethical review processes since no real human participants are involved. This might encourage under-regulated research practices and in the long run weaken ethical safeguards.

Running inference on the 18 LLMs required significant GPU hours, especially including the initial test phase before finalizing the interview pipeline, raising concerns about the environmental impact of experimenting with synthetic survey responses and the access disparities between well-funded and resource-constrained institutions.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, and Harkirat Behl. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Preprint*, arXiv:2404.14219.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Stacey Aston, James Negen, Marko Nardini, and Ulrik Beierholm. 2021. Central tendency biases must be accounted for to consistently capture Bayesian cue combination in continuous response data. *Behavior Research Methods*, 54(1):508–521.

John A. Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2):230–244.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4):401–416.

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. LLM-Based Doppelgänger Models: Leveraging Synthetic Data for Human-Like Responses in Survey Simulations. *IEEE Access*, 12:178917–178927.

Lee J. Cronbach. 1946. Response Sets and Test Validity. *Educational and Psychological Measurement*, 6(4):475–494.

Paolo Crosetto, Antonio Filippin, Peter Katuščák, and John Smith. 2020. Central tendency bias in belief elicitation. *Journal of Economic Psychology*, 78:102273.

Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878. Curran Associates, Inc.

Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wang, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, and Weiran Xu. 2023. Revisit Input Perturbation Problems for LLMs: A Unified Robustness Evaluation Framework for Noisy Slot Filling Task. In *Natural Language Processing and Chinese Computing*, pages 682–694, Cham. Springer Nature Switzerland.

Esther Gan, Yiran Zhao, Liying Cheng, Yancan Mao, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. Reasoning Robustness of LLMs to Adversarial Typographical Errors. *Preprint*, arXiv:2411.05345.

Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are Large Language Models Chameleons? An Attempt to Simulate Social Surveys. *Preprint*, arXiv:2405.19323.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and Soroosh Mariooryad. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Matthew Gereti, Alejandro Robinson, Sebastian Williams, Christopher Anderson, and Dominic Walker. 2024. Token-Based Prompt Manipulation for Automated Large Language Model Evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set.

Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, and Benno Stein. 2017. A Large-Scale Query Spelling Correction Corpus. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1261–1264, Shinjuku Tokyo Japan. ACM.

Edward Tory Higgins. 1996. Knowledge activation: Accessibility, applicability, and salience. In *Social Psychology: Handbook of Basic Principles*, pages 133–168. The Guilford Press, New York, NY, US.

H. L. Hollingworth. 1910. The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the Reliability of Psychological Scales on Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Jarl K. Kampen. 2007. The Impact of Survey Methodology and Context on Central Tendency, Nonresponse and Associations of Subjective Indicators of Government Performance. *Quality & Quantity*, 41(6):793–813.

Achim Koch and Michael Blohm. 2016. Nonresponse Bias (GESIS Survey Guidelines)Nonresponse Bias (GESIS Survey Guidelines). Technical report, GESIS - Leibniz Institute for the Social Sciences.

Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Jon A. Krosnick and Duane F. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201.

Antonio Laverghetta, Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2022. Predicting Human Psychometric Properties Using Computational Language Models. In *Quantitative Psychology*, pages 151–169, Cham. Springer International Publishing.

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. *Preprint*, arXiv:2307.11760.

Llama Meta. 2025a. Llama 3.2 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_-CARD.md.

Llama Meta. 2025b. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_-CARD.md.

Milad Moradi and Matthias Samwald. 2021. Evaluating the Robustness of Neural Language Models to Input Perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alissa O'Halloran, S. Sean Hu, Ann Malarcher, Robert McMillen, Nell Valentine, Mary A. Moore, Jennifer J. Reid, Natalie Darling, and Robert B. Gerzoff. 2014. Response order effects in the Youth Tobacco Survey: Results of a split-ballot experiment. *Survey practice*, 7(3):5.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt Perturbation Consistency Learning for Robust Language Models. *Preprint*, arXiv:2402.15833.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, 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.

Jens Rupprecht, Leon Froehling, Claudia Wagner, and Markus Strohmaier. 2026. German General Social Survey Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies. pages 1761–1780, Palma, Mallorca, Spain.

Howard Schuman and Stanley Presser. 2000. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*, nachdr. edition. Sage Publ, Thousand Oaks, Calif.

Tom Sühr, Florian E. Dorner, Samira Samadi, and Augustin Kelava. 2025. Challenging the Validity of Personality Tests for Large Language Models. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '25, pages 74–81, New York, NY, USA. Association for Computing Machinery.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. *Preprint*, arXiv:2311.04076.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2025. Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. *Social Science Computer Review*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. *Preprint*, arXiv:2402.14499.

Evan Weingarten, Qijia Chen, Maxwell McAdams, Jessica Yi, Justin Hepler, and Dolores Albarracín. 2016. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, 142(5):472–497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao

Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2025b. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and Jing Chang. 2024. Yi: Open Foundation Models by 01.AI. *Preprint*, arXiv:2403.04652.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with Typos for BERT-based Passage Retrieval and Ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Reproducibility Materials

A.1 Infrastructure

The experiments were carried out on a high-performance computing cluster and a local server equipped with NVIDIA H100 (80GB) GPUs. The total runtime for one model's 18,600 interviews, e.g. `Llama-3.1-8B-Instruct` including all perturbed and original Q&As, was ca. 35 minutes with approximately 0.11 seconds per interview. To accommodate larger models on available hardware, we applied 8-bit quantization to `Llama-3.1-8B-Instruct` and `Llama-3.3-70B-Instruct`. Smaller models were run without quantization. Experiments with `Gemini` models were conducted on the Google Cloud Service, Vertex AI. The temperature in all models was kept at their default setting. However, varying the temperature could have a relevant impact and can be varied for robustness checks. The code is made available in an anonymous repository for replication at `https://shorturl.at/NJf6h`.

A.2 Models

This table provides an overview of the open-source LLMs used in the experiments including their Model ID on Huggingface.

Short Name	Huggingface Model ID
Llama 1B	`meta-llama/Llama-3.2-1B-Instruct`
Llama 3B	`meta-llama/Llama-3.2-3B-Instruct`
Llama 8B	`meta-llama/Llama-3.1-8B-Instruct`
Llama 70B	`meta-llama/Llama-3.3-70B-Instruct`
Qwen 7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen 0.6B	`Qwen/Qwen3-0.6B`
Qwen 1.7B	`Qwen/Qwen3-1.7B`
Qwen 4B	`Qwen/Qwen3-4B`
Qwen 8B	`Qwen/Qwen3-8B`
Qwen 14B	`Qwen/Qwen3-14B`
Qwen 32B	`Qwen/Qwen3-32B`
Qwen 30B (R)	`Qwen/Qwen3-30B-A3B-Thinking-2507`
Mistral 7B	`mistralai/Mistral-7B-Instruct-v0.3`
Yi 1.5B	`01-ai/Yi-1.5-6B-Chat`
Phi 3.5B	`microsoft/Phi-3.5-mini-instruct`

Table 3: **Language Models.** We evaluate all Survey Response Generation Methods on 18 open-weight LLMs. LLMs with activated reasoning capabilities are denoted with (R).

B Perturbation Scheme Summary

The following tables describe the remaining non-bias perturbations not introduced in Section 3.2.

The "Type" column categorizes the perturbations into two main classes: "Non-bias Perturbation" and

Type	Perturbation	Question	Answer Options	Bias and Reference
Non-bias Perturbation	**(5) Key Typo**	nor eaca jf the following aspecto, indicete how important it ir wn your liae. Would bou say it is very imporcant, rathes importano, not very imporgant ob not impodtant at all? Famizy	['1=Very important ', '2=Rather important ', '3=Not very important ', '4=Not important at all', "-1=Don't know"]	(Dong et al., 2023; Moradi and Samwald, 2021)
	(6) Letter Swap	For each of the following sapects, indicate how important it is in your life. uoWld you yas it is evry important, ratreh important, ton very important or not important ta all? Family		(Hagen et al., 2017; Moradi and Samwald, 2021; Zhuang and Zuccon, 2021)
	(7) Keyboard Typo	For esch of rhe following aspects, indicate how important ut is un your lide. Would you say it ia very imporrant, rather importamt, nit very important ir nor important ay all? Fanily		(Gan et al., 2024; Zhuang and Zuccon, 2021)
	(8) Synonym Replacement	Crucial in life: Family For each of the following aspects, indicate how significant it is in your life. Would you say it is very important, rather important, not very important or not at all important? Family		(Qiang et al., 2024; Gereti et al., 2024)
	(9) Paraphrase	How important is family to you? Please rate its significance in your life on a scale of "very important" to "not important at all".		(Dong et al., 2023; Qiang et al., 2024)
Interaction	**(10) Paraphrase x Reversal**	How important is family to you? Please rate its significance in your life on a scale of "very important" to "not important at all".	["-1=Don't know", '4=Not important at all', '3=Not very important ', '2=Rather important ', '1=Very important ']	(Dong et al., 2023)

Table 4: **An exemplary perturbation scheme showing non-bias and interaction perturbations.** The example is taken from the item set of category "Importance of Life Aspects". In the WVS wave 7 it is question Q1. Non-bias perturbations have variation in the question phrasing with constant answer options, while the interaction perturbation varies both.

"Interaction." The "Perturbation" column specifies the exact modification technique applied, which includes methods such as "Key Typo," "Letter Swap," "Keyboard Typo," "Synonym Replacement," and "Paraphrase." The "Question" column displays the resulting text after each specific perturbation is applied to the original question about the importance of family. The "Answer Options" column lists the response scale provided to the survey participant. Finally, the "Bias and Reference" column provides citations to relevant scientific literature for each perturbation type.

In the first five perturbations, the phrasing of the question is intentionally altered—for instance, by introducing typographical errors (e.g., "Key Typo," "Letter Swap"), substituting words with similar meanings ("Synonym Replacement"), or rephrasing the entire sentence ("Paraphrase"). While the question varies, the "Answer Options" remain constant, consistently ranging from "1-Very important" to "4-Not important at all.". In "(10) Paraphrase x Reversal," the question is paraphrased, and simultaneously, the order of the "Answer Options" is inverted, starting with "4-Not important at all" and ending with "1-Very important."

C Results

This section summarizes the findings discussed in the main part of the work and can serve as a reference to identify other response patterns as the heatmaps contain much information regarding LLMs, perturbation type as well as additional statistical tests on the response distributions.

C.1 Robustness against Non-Bias Perturbations

This heatmap highlights to which extent the LLMs are affected by *non-bias perturbations.* We can see large model-specific differences.

We see that especially the smallest `Llama` models are responding not in a robust way. These results are consistent across the different *non-bias perturbations*. Especially when facing more than one perturbation in the *Interaction* perturbation, where both answer option scale and question phrasing were altered, the robustness drops drastically across all models .

Moreover, we identify perturbations that are less robustness-harming than others. For example, swapping letters within a word does not impact response robustness as much as typos or introducing completely different words, synonyms, or rephrasing the whole sentence.

C.2 Distance Calculation for Central Tendency and Opinion Floating Bias

This section shows how the response shifts to the central category is measured for the bias perturbations *Odd/Even Scale Transformation* and *Priming Suffix*. By calculating the differences in distances to the central point of the answer option scale we try to identify if the average distribution shifts to the central scale point. The actual differences in distance for each answer option scale type and for the two perturbations are visualized in Table 10a and Table 10b.

To better understand how the shift towards the

Model	(5) Key Typos	(6) Letter Swap	(7) Keyboard Typos	(8) Synonyms	(9) Paraphrase	(10) Paraphrase x Reversed
Llama Family						
L3.3-70B	0.52	**0.76**	0.56	0.58	**0.61**	0.44
L3.1-8B	0.32	0.31	0.21	0.31	0.15	0.10
L3.2-3B	0.02	0.11	0.08	0.13	0.05	0.03
L3.2-1B	0.03	0.10	0.00	0.13	0.00	0.00
Qwen Family						
Q3-32B	0.31	0.34	0.31	0.34	0.19	0.23
Q3-30BT	0.19	0.31	0.19	0.31	0.21	0.23
Q3-14B	0.34	0.47	0.37	0.39	0.40	0.39
Q3-8B	0.29	0.35	0.32	0.37	0.37	0.19
Q2.5-7B	0.48	0.63	0.42	0.55	0.44	0.37
Q3-4B	0.39	0.44	0.47	0.50	0.34	0.19
Q3-1.7B	0.50	0.58	0.58	0.53	0.60	0.34
Q3-0.6B	0.03	0.05	0.10	0.13	0.06	0.00
Gemini Family						
G1.5P	**0.68**	0.73	**0.66**	0.58	0.53	0.24
G2.5P	0.35	0.34	0.35	0.31	0.32	0.24
G2.5F	0.21	0.16	0.27	0.10	0.15	0.13
Others						
M7B	0.58	0.65	0.60	**0.71**	0.53	0.45
Y1.5-6B	0.50	0.50	0.45	0.65	0.29	0.29
P3.5M	0.50	0.61	0.47	**0.71**	0.53	**0.48**

Table 5: **Share of Fully Robust Responses by Perturbation Type and Model (↑).** The models are ordered by model family and parameter size.

middle is measured, we present an anecdotal visualization in Figure 3 of the thought behind whether we observe a *central tendency bias* or an *opinion floating bias*.

In addition, Table 6 underlines that a middle category is significantly more often chosen than assumed under a uniform, or random, distribution of responses across the scale.

Thus, a twofold analysis of not only investigating the shift of average responses between response distributions in the perturbation and original setting is important, but also the statistically assessment to make grounded claims.

By conducting a statistical binomial test, we tried to account for that (Table 6).

C.3 Refusal and Invalid Responses

This section should give a broader overview of the refusal rates (LLM chose the "Don't know" answer option) and the invalid responses (e.g. a LLM did not return any valid response).

It is important to have inspect the overall return rates for all the models as these might have implications on the interpretability of the results. For example, when a model exhibits high refusal or invalid response rates, its results might be not very well interpretable as the main analysis only focused on the valid responses.

Model	3-pt Likert Scale	5-pt Likert Scale	11-pt Likert Scale
Llama Family			
L3.3-70B	1.00	1.00	**0.00**
L3.1-8B	1.00	0.07	**0.00**
L3.2-3B	**0.00**	**0.00**	**0.00**
L3.2-1B	0.31	**0.00**	1.00
Qwen Family			
Q3-32B	1.00	0.01	**0.00**
Q3-30BT	1.00	**0.00**	**0.00**
Q3-14B	0.85	0.95	**0.00**
Q3-8B	1.00	0.11	**0.00**
Q2.5-7B	1.00	0.76	**0.00**
Q3-4B	0.59	0.92	**0.00**
Q3-1.7B	**0.00**	1.00	1.00
Q3-0.6B	1.00	1.00	1.00
Gemini Family			
G1.5P	1.00	1.00	**0.00**
G2.5P	1.00	**0.00**	**0.00**
G2.5F	1.00	**0.00**	**0.00**
Others			
M7B	**0.00**	**0.00**	**0.00**
Y1.5-6B	1.00	**0.00**	**0.00**
P3.5M	1.00	1.00	**0.00**

Table 6: **P-Values of a Binomial Test on the Middle Category in an Odd Scale.** The p-values indicate a hypothesis test with a Null-Hypothesis stating that the middle category is not selected significantly more often assumed under a uniform distribution or completely random selection of answer options. However, for larger scale types, the middle category becomes much more relevant as it is significantly chosen more frequently than any other category.

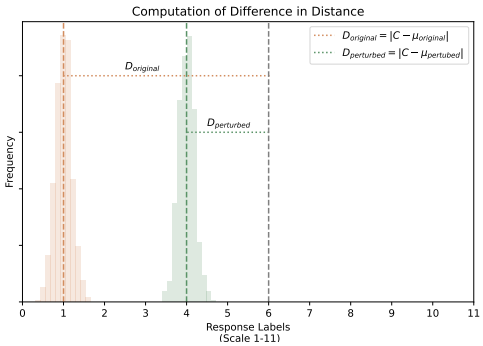


Figure 3: **Exemplary Difference in Distances to Scale Center of Responses to a Perturbed and Original Q&A Pair.** The absolute distance is measured between the scale center and the response mean. Then, $D = D_{perturbed} - D_{original}$. A negative result indicates that the mean response in the perturbed setting is closer to the ideal scale center.

Therefore, this analysis gives insights which results are more reliable than others as for some models there are more valid responses as for others. For example, for `Qwen` we can see large refusal and invalid response rates, generally, but especially in sensitive thematic areas, such as *Perceptions of Elections.*

C.4 Consistency of LLM Survey Responses

This section shows how consistent different LLMs respond to close-ended survey questions when facing the same Q&A pair multiple times. As explained in Section 3, we provide each model with the same Q&A pair in each perturbation stage 25 times and request a response. By running the same setting multiple times we try to identify how consistent LLMs respond generally and whether there are differences when facing the same, but syntactically incorrect (e.g. key typos, etc.), prompts.

Figure 9 highlights that the LLMs consistency in responding is not really affected by specific perturbation types. Thus, "incorrect, flawed prompts" do not increase the response inconsistency of LLMs.

However, there are again large differences between models. We see that especially the smallest `Llama` models and `Qwen` exhibit strong inconsistent responses given the same Q&A pair 25 times, whereas larger LLMs are very consistent. Nonetheless, the `Llama` model family seems to be more inconsistent, generally.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.01	**0.00**	0.02	0.06	0.03
L3.1-8B	0.19	0.02	**0.00**	0.14	0.23	0.19	0.25	0.23	0.22	0.25	0.20	0.10
L3.2-3B	0.23	0.23	0.15	0.28	0.22	0.22	0.30	0.37	0.27	0.24	0.30	0.25
L3.2-1B	0.15	0.28	0.07	0.09	0.32	0.19	0.11	0.16	0.15	0.16	0.19	0.24
Qwen Family												
Q3-32B	0.02	0.02	0.01	0.05	0.02	0.10	0.10	0.06	0.08	0.04	0.05	0.06
Q3-30BT	0.50	0.47	0.04	0.48	0.56	0.46	0.49	0.52	0.50	0.45	0.57	0.57
Q3-14B	0.10	0.08	**0.00**	0.07	0.11	0.09	0.22	0.12	0.19	0.14	0.21	0.11
Q3-8B	0.01	0.03	**0.00**	**0.00**	0.01	0.05	0.11	0.09	0.12	0.05	0.10	0.08
Q2.5-7B	0.40	0.08	0.06	0.31	0.43	0.42	0.49	0.46	0.50	0.37	0.38	0.10
Q3-4B	0.16	0.01	**0.00**	0.12	0.23	0.18	0.14	0.11	0.12	0.14	0.08	0.06
Q3-1.7B	0.02	0.03	0.02	0.02	0.07	0.04	0.08	0.02	0.07	0.06	**0.00**	0.05
Q3-0.6B	0.14	0.01	0.02	0.11	0.14	0.24	0.08	0.10	0.09	0.14	0.16	0.01
Gemini Family												
G1.5P	0.12	0.11	0.03	0.03	0.11	0.07	0.14	0.08	0.11	0.14	0.28	**0.00**
G2.5P	0.50	0.59	0.12	0.47	0.60	0.70	0.53	0.50	0.51	0.50	0.51	0.57
G2.5F	0.44	0.41	**0.00**	0.44	0.50	0.47	0.48	0.44	0.48	0.40	0.47	0.47
Others												
M7B	0.06	0.08	0.03	0.06	0.08	0.03	0.03	0.03	0.03	0.06	0.06	0.03
Y1.5-6B	0.06	0.05	**0.00**	0.16	0.18	0.11	0.06	0.02	0.10	0.06	0.35	0.05
P3.5M	0.05	**0.00**	**0.00**	**0.00**	0.10	0.08	0.10	0.06	0.10	0.05	0.02	0.02

Table 7: **Overall Share of Unsuccessful and Refusal Interviews across Perturbation Type and LLMs.** (↓) Especially the largest non-reasoning models, such as Llama-3.3-70B, Qwen3-32B, and Gemini-1.5-Pro do not refuse the answers or generate wrong interviews (e.g. wrong labels). Reasoning, however, drastically reduces the adherence to respond to a question with one of the answer option categories.

C.5 Comparison of Robustness against Perturbations by Scale Size

The following plots try to reveal in more detail the extent of robustness drop by perturbation depending on the answer option scale dimension. We identified a drop in robustness as the scale size became larger.

The responses of the smallest LLMs are generally not robust at all. However, when the scale has ten options the responses robustness plummets across all perturbations and not even a single response distribution returned in the original Q&A setting can be generated. This indicated that the smallest `Llama` as well as the chinese LLMs `Qwen` and `Yi` are not able to cope with perturbations, especially when facing a lot of options to choose from.

In all cases, the interaction perturbation with both question and answer option alterations leads to the largest drop in robustness across all models and scale sizes. It is striking, that larger models, especially state-of-the-art models like `Gemini-1.5-Pro`, cannot answer robustly given multiple perturbations.

Researchers should take into account the scale size when generating synthetic responses from close-ended survey Q&A pairs.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
L3.1-8B	0.51	**0.00**	**0.00**	0.39	0.54	0.43	0.71	0.50	0.39	0.72	0.25	**0.00**
L3.2-3B	0.39	0.41	**0.00**	0.46	0.38	0.48	0.45	0.40	0.38	0.39	0.23	0.13
L3.2-1B	0.03	0.02	0.07	0.07	0.07	**0.00**	0.03	**0.00**	0.01	0.06	0.13	0.13
Qwen Family												
Q3-32B	0.05	0.10	**0.00**	0.07	0.07	0.20	0.40	0.23	0.28	0.01	0.09	0.02
Q3-30BT	0.97	0.91	0.13	0.93	0.95	0.99	0.88	0.92	0.79	0.49	0.87	0.79
Q3-14B	**0.00**	**0.00**	**0.00**	0.01	**0.00**	0.21	0.12	0.07	0.05	**0.00**	0.17	0.07
Q3-8B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Q2.5-7B	0.91	0.34	**0.00**	0.69	0.93	0.50	0.83	1.00	0.83	0.83	0.83	0.67
Q3-4B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.13	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.03
Q3-1.7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Q3-0.6B	0.01	0.01	**0.00**	0.01	0.01	0.11	0.01	0.07	0.01	0.01	0.13	0.01
Gemini Family												
G1.5P	0.37	0.47	**0.00**	0.05	0.34	**0.00**	0.81	0.33	0.33	0.33	0.17	**0.00**
G2.5P	1.00	1.00	0.31	1.00	1.00	1.00	0.98	1.00	1.00	0.90	0.87	0.81
G2.5F	0.86	0.91	0.02	0.81	0.83	1.00	0.93	0.89	0.89	0.64	0.85	0.87
Others												
M7B	0.17	**0.00**	**0.00**	0.17	0.17	**0.00**	0.17	0.17	0.17	**0.00**	0.33	0.17
Y1.5-6B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.17	**0.00**	**0.00**	**0.00**	**0.00**	0.17	**0.00**
P3.5M	0.17	**0.00**	**0.00**	**0.00**	0.17	0.50	0.17	0.17	**0.00**	0.17	0.17	**0.00**

Table 8: **Perception of Elections: Share of Refusal & Unsuccessful Interviews.** (↓) The models are ordered by model family and parameter size.

Model	Orig.	Rev. Ans.	Miss. Ref.	Odd Scale	Even Scale	Emo. Prim.	Key Typos	Letter Swap	Keyb. Typos	Syn-onyms	Para-phrase	Para. x Rev.
Llama Family												
L3.3-70B	0.13	0.04	0.20	**0.00**	0.14	0.26	0.07	0.20	**0.00**	0.18	0.20	0.08
L3.1-8B	0.61	1.08	0.68	0.44	0.53	0.84	0.14	0.69	0.39	0.65	0.34	0.68
L3.2-3B	1.40	1.19	1.19	1.16	1.35	1.18	1.19	1.25	1.49	1.61	1.46	0.95
L3.2-1B	1.94	1.53	1.92	1.38	1.90	1.34	1.35	1.51	1.15	1.70	1.45	1.21
Qwen Family												
Q3-32B	0.73	0.62	0.66	0.17	0.73	1.00	0.79	0.83	1.01	1.12	0.39	0.49
Q3-30BT	0.39	0.42	0.21	0.32	0.40	0.54	0.28	0.46	0.31	0.43	0.56	**0.00**
Q3-14B	0.21	0.30	0.39	0.03	0.31	0.28	0.24	0.17	0.33	0.37	0.10	0.23
Q3-8B	0.41	0.33	0.46	0.26	0.50	0.33	0.26	0.62	0.27	0.80	0.28	0.44
Q2.5-7B	0.38	0.46	0.38	0.35	0.41	0.42	0.13	0.20	0.16	0.14	0.34	0.34
Q3-4B	0.79	0.69	1.29	0.50	0.83	0.31	0.50	0.44	0.19	0.59	0.67	0.54
Q3-1.7B	0.14	0.03	0.14	0.15	0.16	0.41	0.11	0.03	0.02	0.23	**0.00**	**0.00**
Q3-0.6B	0.68	0.74	1.43	0.91	0.58	2.00	0.58	0.42	0.31	0.58	0.55	0.52
Gemini Family												
G1.5P	**0.00**	**0.00**	0.07	**0.00**	0.02	0.05	0.22	0.17	**0.00**	0.39	**0.00**	**0.00**
G2.5P	0.48	0.50	0.57	0.39	0.47	0.69	0.38	0.61	0.51	0.68	0.33	0.50
G2.5F	0.85	0.64	0.85	0.54	0.86	0.83	0.82	0.83	0.63	1.27	0.50	0.46
Others												
M7B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Y1.5-6B	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
P3.5M	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 9: **Large model-specific differences in response entropy.** (↓) Little to no perturbation-specific differences. Each scale size subsumes all selected questions. This figure displays the mean entropy across all questions in that scale type for all perturbation and model combinations. Warmer colors indicate a higher average dispersion of the responses across the potential answer options. E.g., if a model answers always with the same label, the entropy is 0.

Model	3-pt Likert Scale	4-pt Likert Scale	5-pt Likert Scale	10-pt Likert Scale
Llama Family				
L3.3-70B	0.35	**-0.04**	0.04	**-0.12**
L3.1-8B		**-0.19**	**-0.15**	**-2.46**
L3.2-3B		**-0.47**	0.04	**-0.34**
L3.2-1B	**-0.04**	0.14	0.01	**-0.15**
Qwen Family				
Q3-32B	0.04	**-0.05**	0.03	**-0.37**
Q3-30BT	**-0.18**	0.81	0.07	**-0.15**
Q3-14B	**-0.04**	**-0.02**	0.03	**-0.71**
Q3-8B	0.01	**-0.11**	0.03	0.27
Q2.5-7B	0.19	**-0.49**	**-0.01**	0.28
Q3-4B	**-0.32**	**-0.12**	**-0.35**	0.24
Q3-1.7B	**-0.03**	**-0.12**	**-0.47**	0.70
Q3-0.6B	0.20	0.08	**-0.06**	1.49
Gemini Family				
G1.5P	0.06	**-0.26**		0.26
G2.5P	0.10	**-0.07**	**-0.07**	0.38
G2.5-F	0.09	**-0.07**	0.12	0.09
Others				
M7B	0.17	0.04		**-0.23**
Y1.5-6B	**-0.17**	**-0.08**	0.17	2.95
P3.5M	**-0.20**	**-0.05**	0.33	**-0.48**

(a) **Models adjust their answer behavior towards the middle when the refusal category is missing.**

Model	3-pt Likert Scale	4-pt Likert Scale	10-pt Likert Scale
Llama Family			
L3.3-70B	**-0.28**	**-0.22**	**-1.11**
L3.1-8B	**-0.50**	**-0.18**	**-0.13**
L3.2-3B	**-0.05**	**-0.28**	0.39
L3.2-1B	0.31	**-0.53**	0.78
Qwen Family			
Q3-32B	0.18	**-0.16**	**-0.43**
Q3-30BT	**-0.36**	0.78	**-1.06**
Q3-14B	**-0.31**	**-0.24**	**-0.38**
Q3-8B	**-0.28**	**-0.42**	**-1.83**
Q2.5-7B	0.42	**-0.23**	**-2.01**
Q3-4B	0.43	**-0.13**	1.43
Q3-1.7B	0.46	0.13	**-3.25**
Q3-0.6B	**-0.14**	**-0.43**	0.40
Gemini Family			
G1.5P	**-0.33**	**-0.47**	**-1.11**
G2.5P	**-0.61**	0.23	**-0.62**
G2.5F	**-0.21**	**-0.40**	**-0.45**
Others			
M7B	**-0.02**	**-0.28**	**-1.08**
Y1.5-6B	**-0.15**	**-0.16**	**-0.26**
P3.5M	0.30	**-0.26**	1.60

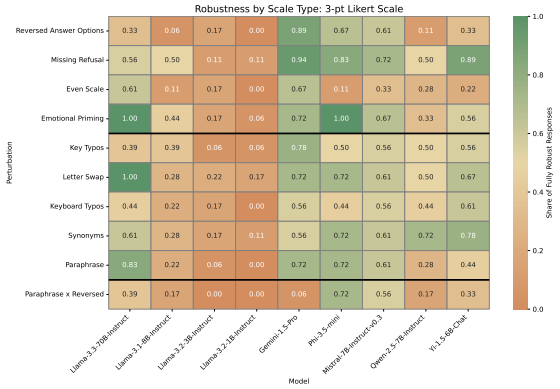
(b) **Models adjust their answer behavior towards the middle when a middle category is existent.**

Table 10: **The values display the difference in mean distance of the perturbed, (a) without refusal category and (b) with middle category to the scale center. Bold values indicate a shift towards the scale center.** For original even scales an artificial middle category is created and vice versa to be able to compare even and odd scales with one another for every question. Thus, in an original 5-pt Likert scale the middle category is removed, whereas in a 4-pt Likert scale a middle category is added. No changes are removed for better readability.

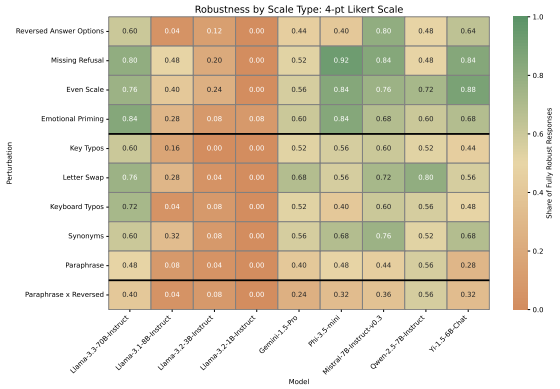
Family	Model	Params	Recency Bias	Opinion Floating	Central Tendency	Emotional Priming
Llama	Llama-3.3-70B-Instruct	70B	✓	×	∼	×
	Llama-3.1-8B-Instruct	8B	✓	∼	∼	×
	Llama-3.2-3B-Instruct	3B	✓	∼	✓	×
	Llama-3.2-1B-Instruct	1B	✓	∼	∼	×
Qwen	Qwen3-32B	32B	✓	∼	∼	×
	Qwen3-30B-A3B (Thinking)	30B	✓	∼	✓	×
	Qwen3-14B	14B	✓	∼	∼	×
	Qwen3-8B	8B	✓	∼	∼	×
	Qwen2.5-7B-Instruct	7B	✓	∼	∼	∼
	Qwen3-4B	4B	✓	∼	∼	×
	Qwen3-1.7B	1.7B	×	∼	∼	×
	Qwen3-0.6B	0.6B	✓	∼	×	×
Gemini	Gemini-1.5-Pro	n/a	✓	×	∼	×
	Gemini-2.5-Pro	n/a	✓	∼	✓	×
	Gemini-2.5-Flash	n/a	✓	∼	✓	×
Others	Mistral-7B-Instruct-v0.3	7B	✓	×	✓	×
	Phi-3.5-mini-Instruct	3.8B	✓	×	✓	×
	Yi-1.5-6B-Chat	6B	✓	∼	∼	∼

Notes: **Recency Bias**: disproportionate selection of the last-presented answer option when response order is reversed (17/18 models affected; increase ranges from $+5\%$ to $+2027\%$). **Opinion Floating**: shift of responses toward the scale centre when a "Don't know" refusal option is removed; most pronounced in smaller models. **Central Tendency**: over-selection of a middle category when one is explicitly provided; significant on larger scales (binomial tests, $p < .05$ for most models on 11-point scales). **Emotional Priming**: change in refusal rate after appending "This is very important to my research! You better do not refuse the answer."; inconsistent across models. Reasoning-capable models show the recency bias at greatly reduced magnitude (≈ 0.07–$0.09\times$ rather than $> 1\times$) but tend to generate more invalid/refusal responses overall.

Table 11: Human-Like Survey Response Biases Identified per LLM. **Legend:** ✓ human-like survey response bias; ∼ partial or inconsistent evidence; × not detected; Model sizes are approximate parameter counts.

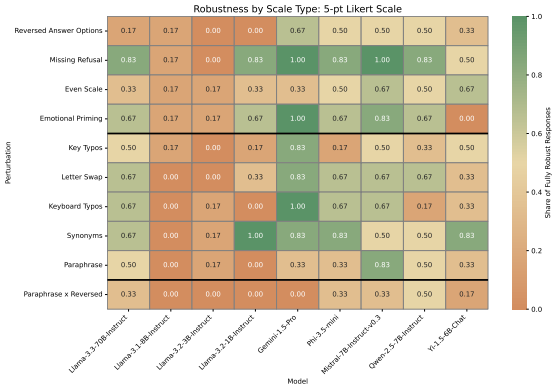


(a) 3-point Likert Scale

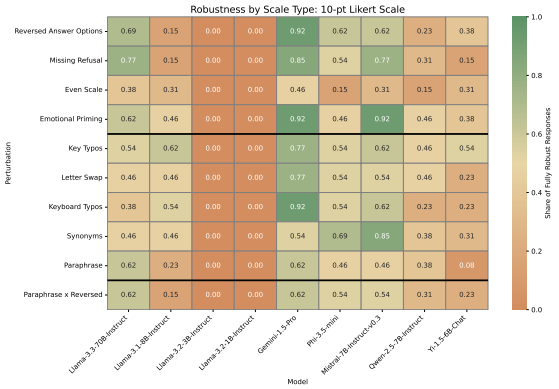


(b) 4-point Likert Scale

Figure 4: **Model-specific differences in fully robust responses on most perturbations on the 3 and 4-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 5 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform very poorly across all scales.



(a) 5-point Likert Scale



(b) 10-point Likert Scale

Figure 5: **Model-specific differences in fully robust responses on most perturbations on the 5- and 10-point scale.** This figure shows the share of fully robust response distributions given the original response distribution and the responses based on the specific perturbation on the y-axis. Compared to 4 the robustness of responses drops when the scale size becomes larger. The smallest `Llama` models perform poorly across all scales.

Borrowed Words, Borrowed Minds: Probing LLM Choice of English-Derived Loanwords in Japanese

Joseph James
Department of Computer Science, The University of Sheffield
Sheffield, United Kingdom
jhfjames1@sheffield.ac.uk

Abstract

The choice between English-derived loanwords (*gairaigo*) and native Japanese equivalents is a socially meaningful aspect of language use, carrying implications for register, style, and pragmatic interpretation. We introduce a controlled evaluation dataset probing how large language models encode this form of sociolinguistic variation. The dataset comprises 113 interchangeable lexical pairs embedded across six communicative contexts spanning formal and informal, spoken and written registers. We evaluate 16 Japanese-capable LLMs across three complementary tasks: sentence rating, pairwise choice, and masked word prediction. Although both lexical forms were generally rated as natural, models diverged substantially in their contextual sensitivity and lexical preferences, revealing architectural differences in how socially grounded lexical alternatives are represented. These findings suggest that surface fluency may mask instability in modeling pragmatic variation, with implications for socially aware language generation and evaluation. Dataset and prompts will be made publicly available upon publication to facilitate replication and further research.

1 Introduction

The Japanese lexicon is characterised by extensive borrowing, particularly from English (Irwin, 2011; Tomoda, 1999). English-derived loanwords, known as *gairaigo*, coexist with native (*wago*) and Sino-Japanese (*kango*) equivalents and are typically written in katakana. In many cases, loanwords overlap semantically with existing Japanese terms. For example, *henji* (返事) and *ripurai* (リプライ) both denote a reply, yet differ in communicative association and contextual framing. *Henji* functions as a broad native term used across a wide range of settings, whereas *ripurai* is more closely associated with digital communication and contemporary discourse. Such alternations are not neutral substitutions but carry contextual and social meaning.

Loanword usage in Japanese reflects multiple sociolinguistic dimensions. English-derived forms may signal modernity, international orientation, technological currency, or commercial branding, while native equivalents may evoke institutional authority, convention, or cultural continuity. At the same time, register distinctions in Japanese remain highly codified and play a central role in shaping lexical, grammatical, and pragmatic choices in both speech and writing (Liu and Allen, 2014; Dunn, 1999; Matsumoto, 1988). Formal and informal contexts often impose systematic constraints on lexical selection, making register a salient and empirically tractable dimension of sociolinguistic variation. Although lexical alternation cannot be reduced to formality alone, the formal–informal contrast provides a principled baseline for examining contextual sensitivity.

These issues are increasingly relevant in the context of large language models. Contemporary LLMs trained on large-scale Japanese corpora are exposed to diverse registers and communicative styles (Kuribayashi et al., 2021). However, exposure does not necessarily imply appropriate contextual differentiation. Models may generate fluent output while failing to distinguish subtle register constraints or socially appropriate lexical choices. For applications in education, translation, and writing assistance, such distinctions are consequential. If models treat near-equivalent loanword and native forms as interchangeable across contexts, they risk obscuring sociolinguistic nuance.

Despite growing interest in stylistic control and sociolinguistic evaluation of LLMs, systematic examination of context-sensitive lexical alternation in Japanese remains limited. In this paper, we investigate how LLMs select between English-derived loanwords and native Japanese equivalents across structured communicative settings. We introduce a

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 22–36

dataset of 113 interchangeable lexical pairs embedded across six contexts that vary by register, mode, and discourse function. Because each sentence pair differs only in lexical form, the design enables controlled testing of contextual differentiation while holding semantic content constant.

We evaluate a closed-source model (GPT-5) alongside state-of-the-art open-source Japanese-capable LLMs using three complementary tasks: sentence-level rating, pairwise comparison, and masked prediction. Together, these tasks allow us to examine both surface judgements of naturalness and token-level lexical preferences. Our study provides a reusable evaluation resource and empirical insight into how contemporary LLMs encode sociolinguistic variation in Japanese, contributing to ongoing efforts to develop culturally informed and pragmatically appropriate NLP systems. Our objective is not to determine which lexical form is correct in a given context, but to probe whether different architectures exhibit systematic and context-sensitive differentiation under controlled conditions.

2 Related Work

2.1 Loanwords in Japanese

The tripartite structure of the Japanese lexicon, *wago*, *kango*, and *gairaigo*, is well established (Shibatani, 1990; Irwin, 2011). While borrowing from Chinese has shaped the lexicon for centuries, English-derived loanwords have expanded rapidly in recent decades and now occupy a prominent role across communicative domains. Beyond filling lexical gaps, *gairaigo* frequently serve stylistic and pragmatic functions. Sociolinguistic accounts emphasise their role in indexing modernity, cosmopolitan identity, technological innovation, and global orientation (Stanlaw, 2004; Loveday, 1996; Takashi, 1990).

Loanword choice is therefore not merely semantic substitution but a socially meaningful choice. In advertising and professional discourse, English borrowings can signal Western affiliation or prestige (Takashi, 1990), while native equivalents may evoke institutional authority or tradition. Loanword choice is therefore not merely semantic substitution but a socially meaningful one. This view follows a long tradition in variationist sociolinguistics establishing that lexical and phonological alternation is socially stratified and stylistically conditioned (Labov, 1973), and that variants carry not fixed categorical meanings but a field of context-dependent social associations (Eckert, 2008). In advertising and professional discourse, English borrowings can signal Western affiliation or prestige (Takashi, 1990), while native equivalents may evoke institutional authority or tradition. Register distinctions further shape lexical selection, as formal and informal contexts systematically influence lexical, grammatical, and pragmatic choices (Liu and Allen, 2014; Dunn, 1999; Matsumoto, 1988). These findings establish lexical alternation as a context-sensitive phenomenon embedded in broader sociocultural systems.

Semantic divergence between loanwords and their English source forms has also been documented. Using distributional embeddings, Takamura et al. (2017) demonstrate measurable shifts in meaning, reinforcing the need for careful consideration of semantic equivalence when constructing interchangeable pairs. Research on English-derived words coined within Japan further highlights divergence in usage and interpretation (Hatanaka and Pannell, 2016). Such work underscores the complexity of treating loanwords and native equivalents as fully interchangeable forms.

2.2 Loanwords in Language Acquisition

Loanwords play a documented role in second language acquisition. Daulton (2008) describes English-derived vocabulary in Japanese as a "built-in lexicon" that facilitates lexical access while potentially encouraging assumptions of cross-linguistic equivalence. Classroom studies report that both learners and teachers perceive loanword-based transfer as a source of facilitation as well as confusion, particularly where form and meaning diverge (Spring, 2018). Attitudinal research further suggests ambivalence toward loanwords, balancing perceived usefulness against concerns about clarity or appropriateness (Daulton, 2011).

Empirical studies demonstrate both benefits and risks of loanword reliance. Aizawa et al. (2024) show that Japanese learners perform better on English vocabulary tests when target words correspond to familiar loanwords, while Ferries (2022) find evidence of loanword-influenced semantic transfer in learner English writing. Comprehension studies also indicate that understanding of loanwords varies depending on speaker background (Alharaki et al., 2023). Together, this literature suggests that near-equivalent loanword–native pairs may not be interpreted uniformly across audiences, and that

contextual appropriateness remains a pedagogically relevant concern.

2.3 LLM Evaluation and Sociolinguistics

Large language models have demonstrated strong performance in Japanese–English translation and related generation tasks (Yan et al., 2024; Jiao et al., 2023). Benchmarks such as the Open Japanese LLM Leaderboard evaluate translation, summarisation, and dialogue performance, indirectly reflecting stylistic competence, though without explicit focus on lexical alternation.[1]

Beyond translation quality, research has begun probing LLMs for sociolinguistic sensitivity. Controlled prompting studies show that persona and role instructions can shift output style (Salewski et al., 2023). Dialect-sensitive evaluation reveals disparities across language varieties (Deas et al., 2023; Tjuatja et al., 2024), while multilingual analyses of politeness and formality report partial but inconsistent alignment with human norms (Srinivasan and Choi, 2022). In Japanese NLP, corpora for spoken-to-written style conversion and text simplification further highlight the importance of modelling register and pragmatic variation (Ihori et al., 2020; Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018; Hatagaki et al., 2022; Nagai et al., 2024; Urakawa et al., 2024).

However, systematic evaluation of context-sensitive lexical choice between loanwords and native equivalents remains limited. Our work addresses this gap by introducing a structured evaluation for analysing contextual differentiation in Japanese lexical choice across multiple LLM architectures.

3 Data processing

3.1 Loan word extraction

To extract loanwords, we leveraged several large-scale Japanese-English datasets (JMdict Project, 2025; range3, 2023; Maruyama and Yamamoto, 2018). We filtered the lexicon to identify all entries designated as loanwords (katakana written terms) and selected those with a corresponding native Japanese synonym. This procedure provided an initial list of pairs. This list was further supplemented with manually selected pairs from semantic domains in which loanwords are especially prevalent (e.g., colour terminology), ensuring broader coverage across frequently occurring lexical categories. The resulting list was then curated to retain only those pairs where the terms were commonly interchangeable in modern usage. In total, we extracted 221 candidate loanwords.

3.2 Sentence Generation

Using the curated list of loanword–native pairs, we constructed a sentence-level evaluation dataset. To ensure systematic and context-controlled generation, we employed the generative model Gemini 2.5 Pro (Comanici et al., 2025) to produce paired sentences for each lexical item. The aim was to embed each pair within communicative settings that vary along dimensions of register, mode, and discourse function, enabling controlled testing of contextual sensitivity. Prompt provided in Appendix B in Tab 5.

For each lexical pair, we generated sentences across six predefined communicative contexts:

- **Formal Conversation:** A short, polite dialogue typical of a business or service interaction.
- **Formal Written:** A sentence resembling a report, academic paper, or official correspondence.
- **Formal Explanation:** A definition or technical description, as found in a textbook or manual.
- **Informal Conversation:** A casual exchange between friends or peers.
- **Informal Written:** A sentence similar to a personal message, email, or social media post.
- **Informal Explanation:** A casual explanation directed at a peer.

For each context, the model was prompted to generate two sentences that were semantically equivalent and differed only in lexical choice, specifically the use of the loanword versus its native counterpart. Because some loanwords exhibit polysemy, prompts explicitly constrained generation to senses in which both forms were contextually interchangeable. This ensured that lexical form remained the sole manipulated variable, isolating contextual preference rather than semantic divergence.

3.3 Automatic Quality Check

We implemented an automatic quality check to filter the generated dataset for semantic consistency. The core of this check was a back-translation workflow designed to detect potential meaning shifts between the loanword and native sentence variants.

[1] https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard

Japanese	English
これは成功するチャンスです。	This is a **chance** to succeed.
これは成功する機会です。	This is an **opportunity** to succeed.

Table 1: Example illustrating lexical pair: “チャンス” (chansu) vs. “機会” (kikai).

For each Japanese sentence pair, both versions were translated into English using the DeepL API[2] and the Google Translate API,[3] producing corresponding English sentences. To quantify semantic similarity, we generated sentence embeddings for each translation using a pre-trained multilingual Sentence-BERT model (Reimers and Gurevych, 2019).[4] Cosine similarity was then computed between the embedding vectors.

Sentence pairs with a similarity score below a threshold of 0.95 were automatically flagged for review. Terms for which the majority of sentence pairs were flagged were removed from the dataset. For terms with only one or two flagged instances, new sentences were generated.

The use of round-trip translation as a proxy for semantic equivalence follows established practice in machine translation evaluation, where back-translation and semantic comparison are used to detect meaning divergence and verify consistency between parallel sentence pairs (Jia et al., 2025; Edunov et al., 2020; Federmann, 2012).

3.4 Manual Quality Check

The quality and consistency of the generated dataset were ensured through manual verification by a fluent bilingual (English–Japanese) evaluator. The evaluator confirmed the semantic integrity and contextual interchangeability of each sentence pair. Specifically, both the loanword and native term had to convey the same propositional meaning and function as valid substitutes within the specified context (e.g., Formal Written). Grammatical well-formedness under substitution was also verified.

Table 1 illustrates a case in which lexical alternation may introduce subtle shifts in connotation without altering overall intent. Although *chansu* (チャンス) and *kikai* (機会) both refer to the possibility of doing something, chansu is often rendered as “chance,” carrying a slightly casual or motivational tone, whereas kikai is more frequently translated as “opportunity,” particularly in formal or institutional contexts. Substituting one for the other does not substantially change the underlying meaning of the sentence, but it may affect perceived register or formality. Pairs exhibiting such shifts in domain specificity, connotation, or pragmatic scope were excluded from the dataset. This filtering criterion was applied consistently across all candidates to ensure that retained pairs were genuinely interchangeable across the six communicative contexts. After automatic and manual quality checks, 113 lexical pairs remained. Each pair was embedded across six contexts with two variants per context, yielding 12 sentences per pair and a total of **1,356** sentences in the final dataset.

4 Experimental Setup

4.1 Task Definition

We evaluate lexical selection using three complementary tasks designed to probe contextual sensitivity at both sentence and token levels.

In the **Rating Task**, the model is presented with a single sentence containing either the loanword or its native equivalent, together with the specified communicative context. The model assigns a naturalness score on a five-point scale. These ratings allow comparison of perceived acceptability across lexical forms and contexts.

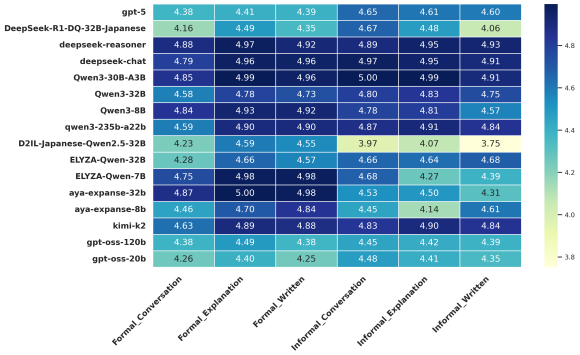
In the **Comparison Task**, the model is given two sentences that are identical except for the target word and must select the more natural option within the given context. Each pair is presented twice with reversed order. The **Self** score in this task measures order consistency, calculated as the proportion of instances in which a model selects the same lexical option regardless of presentation order. Pairwise agreement across models is used to assess cross-model convergence in lexical preference.

In the **Masked Prediction Task**, the target word is removed and the model is asked to generate the most appropriate lexical item. We record whether the model produces the loanword, the native equivalent, or an alternative form. Two variants are implemented. In the **WITH** condition, the model selects between the original pairs. In the **OPEN** condition, no restriction is imposed and the model freely generates a lexical item. The **Self** score in this task measures agreement between a model’s WITH and OPEN predictions for the same item.

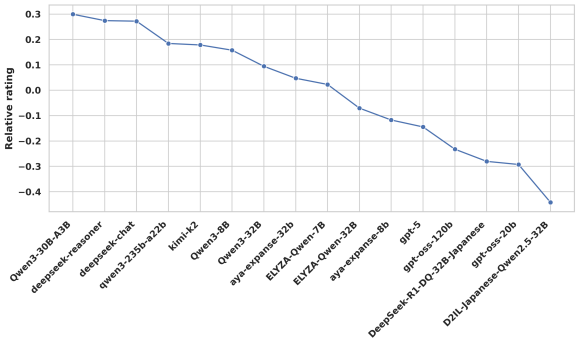
[2] https://www.deepl.com/en/pro-api

[3] https://cloud.google.com/translate/docs/reference/rest

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2



(a) Average ratings heatmap.



(b) Model rating sensitivity.

Figure 1: Results of the Rating Task. (a) Average naturalness ratings (1–5) across models and contexts. (b) Context sensitivity of models, shown as relative differences between model scores and the overall average.

Importantly, our evaluation does not treat either lexical form as inherently correct within a given context, nor does it assume a fixed normative standard of speaker preference. Rather than measuring alignment with human speaker norms, our analysis focuses on relative architectural divergence across models under controlled contextual conditions. The goal is to examine how different LLMs distribute lexical choices given identical inputs, not to determine which model best reflects contemporary usage.

4.2 Models

We evaluate a broad set of Japanese-capable LLMs. The closed-source group is represented by `GPT-5` (Singh et al., 2025), while the open-source group spans several model families: `GPT-OSS` (20B, 120B)(OpenAI, 2025), `DeepSeek` (Reasoner, Chat) (DeepSeek-AI, 2024), `CyberAgent` (DeepSeek-R1 Distill Qwen-32B Japanese) (Ishigami, 2025), `Deep Analysis Research` (Japanese Qwen2.5-32B), the `ELYZA Shortcut series` (7B, 32B) (Hirakawa et al., 2025), the `Qwen-3 family` (8B, 32B, 30B-A3B, 235B-A22B) (Team, 2025), `Kimi-K2` (Team et al., 2025), and `Cohere Aya-Expanse` (8B, 32B) (Dang et al., 2024). All prompts and model specifications are provided in Appendix A and B.

5 Results

5.1 Rating Task

Although both loanword and native equivalents are valid, one may be more appropriate in context. The Rating Task evaluates whether the constructed pairs are perceived as interchangeable by measuring overall naturalness without separating scores by lexical type. Across all models, ratings for sentences containing either form clustered toward the upper end of the scale (see Figure 1a). This concentration suggests that the majority of generated pairs were judged natural regardless of whether the loanword or native variant was used, supporting the semantic equivalence of the dataset.

Within this generally compressed range, variation is more strongly attributable to model calibration than to lexical form. Stability differs by architecture rather than uniformly by size. DeepSeek-Reasoner exhibits one of the flattest rating profiles across communicative settings, with minimal separation between formal and informal contexts, and DeepSeek-Chat shows similarly limited spread. GPT-5, by contrast, demonstrates a systematic uplift in informal settings relative to formal ones, indicating a consistent context effect rather than complete uniformity. DeepSeek-R1-DQ-32B-Japanese displays the largest within-model variation, largely driven by a lower score in Informal Written. Family-level tendencies are also visible: Qwen and DeepSeek models generally assign higher average ratings overall, whereas GPT-OSS variants apply comparatively stricter evaluations, as illustrated in Figure 1b.

These differences primarily reflect evaluative calibration rather than strong lexical discrimination. More generous systems may present a broad range of lexical choices as equally acceptable, potentially obscuring finer register distinctions. Conversely, stricter systems may assign lower scores even when both variants are contextually legitimate. The Rating Task confirms that both variants are generally accepted as natural within their contexts. This ensures that subsequent tasks examine differences in preference rather than problems of semantic mismatch.

Comparative research in Japanese sociolinguistics shows that lexical preferences vary systematically across contexts. Native terms are generally favoured in formal settings, reflecting expectations of careful or official language use (Hashimoto, 2019). In informal contexts, distinctions are less stable, and conversational settings tend to allow greater variability (Stanlaw, 2004; Loveday, 1996). Divergence between lexical alternatives is often greater in interactionally sensitive contexts and lower in informational or technical discourse (Stanlaw, 2004; Loveday, 1996). Sensitivity to register and context is therefore central to sociolinguistic norms of lexical selection.

The relative generosity of Qwen and DeepSeek models may reflect differences in training objectives and instruction tuning. Systems optimised for conversational helpfulness or safety alignment may be less inclined to assign low ratings in the absence of clear grammatical errors. Training data composition may also play a role, as exposure to web or media corpora, where katakana loanwords are frequent, could increase tolerance towards lexical variation. Sentence length likewise affects ratings (Appendix C). Overall, rating behaviour appears to reflect calibration and optimisation choices in addition to sociolinguistic sensitivity, underscoring the importance of recognising model-specific tendencies when interpreting LLM feedback.

5.2 Comparison Task

When comparing sentence pairs, models exhibited systematic but non-convergent preferences. Agreement rates were consistently above chance yet moderate overall (see Table 2). Larger models tended to align more closely with one another, with GPT-5 showing the highest agreement with GPT-OSS-120B and DeepSeek-Reasoner. In contrast, smaller Aya, Kimi, and ELYZA variants demonstrated lower cross-system alignment.

Order bias was evaluated by reversing sentence presentation (Pezeshkpour and Hruschka, 2024). Most models maintained high internal consistency under this manipulation, particularly larger systems, indicating that preferences were stable and unlikely to arise from superficial prompt artefacts. Model disagreement therefore appears to reflect genuine differences in lexical judgement rather than task instability.

Lexical preferences further revealed a contextual divide (Figure 2). Native equivalents were more frequently selected in formal registers, whereas

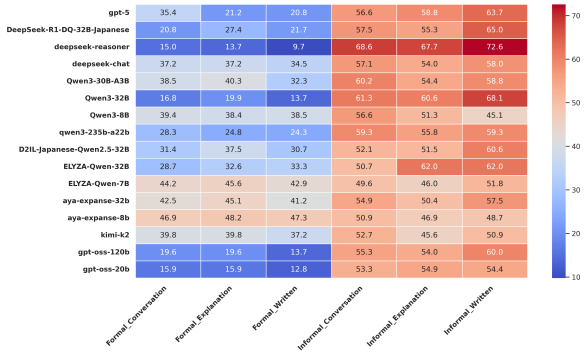


Figure 2: Heatmap of loanword selection percentages across six communicative contexts. Values indicate how often models chose the loanword sentence. Darker/red = higher loanword selection; lighter/blue = native selection

loanwords appeared more often in informal settings. DeepSeek and GPT-OSS variants demonstrated clearer contextual sensitivity, adjusting preferences across registers. Aya variants showed weaker differentiation and tended to favour loanwords more uniformly. We do not establish a human normative baseline in this study; our analysis therefore describes divergence in model-internal lexical distributions rather than verified alignment with contemporary speaker behaviour.

The contrast with the Rating Task is notable. While scalar ratings suggested broad acceptability of both forms, direct comparison exposed sharper architectural divergence. By requiring an explicit binary choice, the Comparison Task reveals lexical preferences that remain hidden in gradient evaluations.

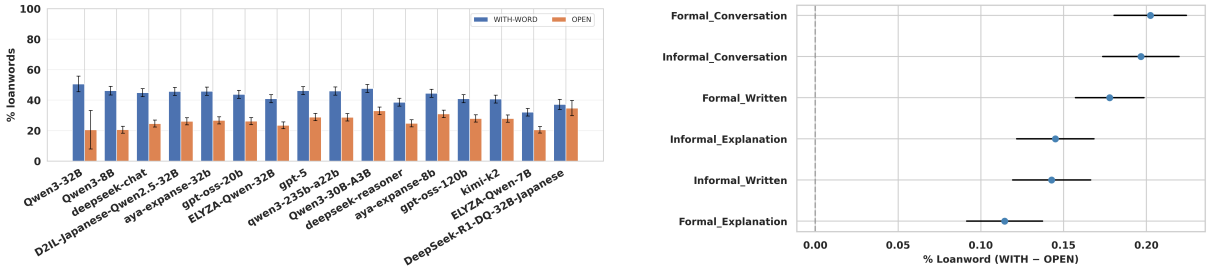
5.3 Masked Word Prediction Task

Within-model "Self" scores measured consistency between the WITH condition, in which the original loanword–native pair was provided, and the OPEN condition, which allowed unconstrained generation. Most systems demonstrated moderate to high internal stability, particularly ELYZA, Kimi and GPT variants. However, internal consistency did not necessarily translate into cross-model agreement. ELYZA models were stable within themselves yet systematically distinct from other systems, whereas Qwen models showed lower internal stability but aligned more closely with peers.

Pairwise comparisons under both WITH and OPEN exceeded chance levels but varied across model families as shown in Table 3. DeepSeek-Chat, Aya-Expanse-32B and Qwen3-32B exhibited relatively stronger agreement with other systems,

Model	Self (%)	Pairwise agreement (%)				
		Avg	Min	Model	Max	Model
gpt-5	91.7	68.5	56.8	aya-expanse-8b	78.5	gpt-oss-120b
gpt-oss-120b	85.3	66.6	55.5	aya-expanse-8b	78.5	gpt-5
deepseek-reasoner	85.1	64.6	57.2	kimi-k2	75.8	gpt-5
gpt-oss-20b	83.3	65.3	55.1	aya-expanse-8b	75.4	gpt-5
Qwen3-32B	82.3	64.4	54.9	aya-expanse-8b	73.0	gpt-5
DeepSeek-R1-DQ-32B-Japanese	74.0	63.0	56.4	ELYZA-Qwen-7B	69.0	deepseek-reasoner
ELYZA-Qwen-32B	68.0	61.5	55.1	ELYZA-Qwen-7B	65.0	gpt-5
Qwen3-8B	63.0	60.0	54.1	aya-expanse-8b	64.6	gpt-5
aya-expanse-32b	62.4	60.4	55.0	aya-expanse-8b	65.9	gpt-5
qwen3-235b-a22b	61.8	61.3	53.8	aya-expanse-8b	68.0	gpt-5
Qwen3-30B-A3B	59.1	59.3	53.6	aya-expanse-8b	65.2	gpt-5
D2IL-Japanese-Qwen2.5-32B	59.1	60.1	53.6	aya-expanse-8b	65.6	gpt-5
deepseek-chat	55.2	60.5	53.9	aya-expanse-8b	65.5	gpt-5
ELYZA-Qwen-7B	46.5	57.2	53.5	kimi-k2	60.3	gpt-5
kimi-k2	33.8	56.6	52.2	aya-expanse-8b	58.8	gpt-5
aya-expanse-8b	32.9	54.6	52.2	kimi-k2	56.8	gpt-5

Table 2: Pairwise agreement. "Self" is to check order bias. Avg/Max/Min computed excluding self.



(a) Loanword selection rates under WITH vs. OPEN.

(b) Contextual differences between WITH and OPEN.

Figure 3: Loanword usage patterns in the Masked Prediction Task. (a) shows aggregated loanword proportions across models under WITH vs. OPEN with order from left to right showing the difference between the two tasks; (b) shows context-specific gaps with 95% confidence intervals.

while ELYZA-Qwen-7B and DeepSeek-R1-DQ-32B showed weaker alignment. Loanword selection patterns clarify these dynamics (see Figure 3). Under WITH, loanword usage approached balance between alternatives. Under OPEN, loanword rates decreased and cross-model divergence increased. Constraining the candidate set therefore promotes convergence, whereas unconstrained generation exposes underlying distributional tendencies. Task design thus meaningfully shapes lexical outcomes.

Taken together, masked prediction shows that models converge at the level of broad stylistic orientation but diverge at the level of specific lexical realisation. Structured prompting encourages agreement, whereas open generation surfaces architectural differences in lexical priors.

5.4 Discussion

Our results show that sentence-level ratings conceal substantial variation in lexical preference. Across contexts, models consistently assigned high naturalness scores to both loanword and native alternatives, creating the impression of broad acceptability. However, comparison and masked prediction tasks revealed clearer divergence: models differed in loanword frequency and in the degree to which they tracked contextual cues. In several cases, architectural divergence exceeded shifts across communicative contexts. Rating-based evaluation therefore appears relatively flat, whereas token-level probing exposes finer lexical tendencies. This pattern aligns with findings that generative models can exhibit socially patterned behaviour even when surface-level evaluations suggest neutrality (Hu et al., 2025).

Agreement patterns further clarify this distinction. Models often converged on broad stylistic orientation while diverging in specific lexical realisations. Masked OPEN prediction tended to favour native forms, whereas constrained WITH prediction produced more balanced distributions. Pairwise comparisons also revealed clearer register-based shifts than scalar ratings (Figure 4). These findings support the view that language models encode distributions of socially meaningful styles and

Model	Self (%)	WITH (%)					OPEN (%)				
		Avg	Min	Model	Max	Model	Avg	Min	Model	Max	Model
ELYZA-Qwen-7B	73.7	74.3	71.7	gpt-5	78.1	ELYZA-Qwen-32B	75.8	71.5	DeepSeek-R1-DQ-32B-Japanese	80.0	deepseek-chat
kimi-k2	72.5	79.5	76.2	ELYZA-Qwen-7B	83.6	deepseek-chat	76.5	68.6	DeepSeek-R1-DQ-32B-Japanese	87.2	Qwen3-32B
gpt-oss-120b	72.3	79.2	74.4	ELYZA-Qwen-7B	83.0	deepseek-chat	75.7	67.5	DeepSeek-R1-DQ-32B-Japanese	81.0	deepseek-chat
gpt-5	71.8	78.9	71.7	ELYZA-Qwen-7B	86.2	Qwen3-32B	73.5	64.7	DeepSeek-R1-DQ-32B-Japanese	77.9	gpt-oss-120b
deepseek-reasoner	71.3	78.3	75.6	ELYZA-Qwen-7B	83.4	deepseek-chat	76.3	69.9	DeepSeek-R1-DQ-32B-Japanese	85.4	Qwen3-32B
ELYZA-Qwen-32B	70.4	80.2	76.7	gpt-oss-20b	82.1	D2IL-Japanese-Qwen2.5-32B	78.0	72.1	DeepSeek-R1-DQ-32B-Japanese	84.6	Qwen3-32B
deepseek-chat	69.8	81.4	74.5	ELYZA-Qwen-7B	85.4	aya-expanse-32b	78.0	69.7	DeepSeek-R1-DQ-32B-Japanese	81.7	ELYZA-Qwen-32B
qwen3-235b-a22b	69.6	79.3	73.1	ELYZA-Qwen-7B	83.3	aya-expanse-32b	75.4	68.8	DeepSeek-R1-DQ-32B-Japanese	79.0	deepseek-chat
aya-expanse-8b	67.8	79.1	74.2	ELYZA-Qwen-7B	83.9	aya-expanse-32b	72.9	67.2	DeepSeek-R1-DQ-32B-Japanese	79.2	aya-expanse-32b
Qwen3-30B-A3B	66.9	78.5	74.0	ELYZA-Qwen-7B	80.6	ELYZA-Qwen-32B	72.3	66.2	DeepSeek-R1-DQ-32B-Japanese	76.7	Qwen3-32B
aya-expanse-32b	66.7	81.5	75.7	ELYZA-Qwen-7B	85.4	deepseek-chat	77.3	70.4	DeepSeek-R1-DQ-32B-Japanese	84.2	Qwen3-32B
gpt-oss-20b	66.6	77.2	72.1	ELYZA-Qwen-7B	81.8	Qwen3-32B	73.5	68.1	DeepSeek-R1-DQ-32B-Japanese	81.4	Qwen3-32B
DeepSeek-R1-DQ-32B-Japanese	66.3	79.6	77.5	ELYZA-Qwen-7B	82.0	ELYZA-Qwen-32B	69.9	64.7	gpt-5	85.7	Qwen3-32B
D2IL-Japanese-Qwen2.5-32B	66.1	76.8	72.2	ELYZA-Qwen-7B	82.1	ELYZA-Qwen-32B	74.4	67.7	DeepSeek-R1-DQ-32B-Japanese	81.4	Qwen3-32B
Qwen3-8B	61.6	77.4	72.6	ELYZA-Qwen-7B	80.3	Qwen3-30B-A3B	74.6	70.3	DeepSeek-R1-DQ-32B-Japanese	79.1	deepseek-chat
Qwen3-32B	60.0	80.7	73.3	ELYZA-Qwen-7B	86.2	gpt-5	79.5	72.5	gpt-oss-120b	87.2	kimi-k2

Table 3: Pairwise agreement summary by condition (WITH / OPEN). "Self" is WITH vs. OPEN self-consistency. All values shown as percentages. Avg/Max/Min computed excluding self.

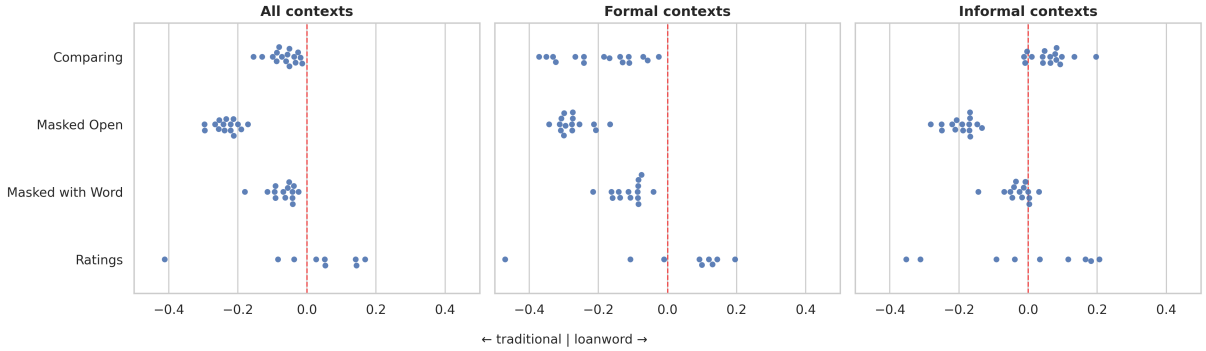


Figure 4: Bias scores across tasks (ratings, comparisons and masked predictions) in formal and informal contexts. Points represent model averages; values to the left indicate preference for native forms, and values to the right indicate preference for loanwords.

registers alongside grammatical structure (Grieve et al., 2025).

At the same time, register alone does not determine loanword usage. English-derived forms in Japanese frequently index modernity, global orientation, technical precision, or euphemistic framing independently of formal–informal contrasts (Stanlaw, 2004; Loveday, 1996; Takashi, 1990). In explanatory or technical contexts, katakana forms may signal domain alignment rather than informality. While our evaluation operationalises register within structured settings, the observed variation likely reflects interaction between contextual cues and broader sociocultural meanings. The results should therefore be interpreted as evidence of register-sensitive differentiation within this design rather than as a comprehensive account of the indexical functions associated with loanword usage.

These findings also have implications for LLM-assisted language learning. Scalar judgements may present alternatives as equally natural, obscuring contextual differentiation, whereas comparison and generative tasks reveal architecture-dependent lexical tendencies that vary with prompt framing. Effective deployment of LLMs in educational contexts therefore requires sociolinguistically informed evaluation and careful calibration (Nguyen, 2025). This dynamic is reminiscent of how language attitudes are socially conditioned: perceptions of appropriateness and prestige are shaped by social experience rather than being fixed properties of forms (Garrett, 2010; Eckert, 2008). In contemporary Japanese discourse, loanwords often function as markers of modernity and professional identity (Takashi, 1990). If training data disproportionately reflect media-heavy corpora, similar distributional pressures may influence model outputs, consistent with research on implicit bias formation in humans and language models (Greenwald and Banaji, 1995; Caliskan et al., 2017). While prompting methods can alleviate some of these tendencies by constraining output space or explicitly foregrounding register, they do not eliminate underlying distributional biases inherited from training data and model optimisation.

Although designed for controlled evaluation, the dataset makes explicit how lexical alternation interacts with communicative context by embedding interchangeable pairs across six register conditions. This structure isolates register-sensitive variation rather than inferring it from heterogeneous corpora.

More broadly, the results underscore that surface-level naturalness does not guarantee contextual appropriateness, highlighting the importance of task design in probing socially conditioned lexical choice in LLMs.

6 Conclusion

Our findings demonstrate that LLMs do not encode a stable contextual rule for loanword versus native lexical selection, despite producing generally fluent outputs. While both forms are often judged acceptable at the sentence level, cross-architectural differences reveal uneven sensitivity to socially conditioned register cues. This highlights a broader limitation in socially grounded language modeling: fluency does not guarantee consistent representation of pragmatic variation. For learner-facing applications, this means outputs should be interpreted cautiously and framed appropriately. More broadly, our dataset provides a controlled benchmark for evaluating context-sensitive lexical generation and offers insight into how contemporary LLMs model socially meaningful linguistic alternation.

Limitations

Our study isolates lexical choice by constructing semantically equivalent sentence pairs. While this provides experimental control, it overlooks document context, which can strongly influence lexical decisions in communication. The process of selecting lexical pairs and generating sentences with LLM assistance, followed by back-translation using commercial MT systems, may introduce uneven coverage of semantic domains and artifacts from machine translation. Finally, our focus on English-derived loanwords in general-domain contexts limits the scope of our findings, they may not generalise to loanwords from highly specialised terminology.

Japanese presents a particularly demanding test case for lexical modelling because sociolinguistic contrasts are partially encoded orthographically (e.g., katakana vs. kanji), historically layered (wago/kango/gairaigo), and pragmatically dependent on discourse setting. A model that succeeds in distinguishing these layers demonstrates sensitivity not only to lexical frequency but to socially structured linguistic choice. This makes Japanese a high-resolution benchmark for stylistic control in LLMs more broadly.

Acknowledgements

Joseph James was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, and 1 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Sura Alharaki, Muhammad Alif Redzuan Abdullah, and Syed Nurulakla Bin Syed Abdullah. 2023. Comprehension of english loanwords in japanese by japanese and english speakers. *World*, 13(5).

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, 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*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Frank E Daulton. 2008. *Japan's built-in lexicon of English-based loanwords*, volume 26. Multilingual Matters.

Frank E Daulton. 2011. On the origins of gairaigo bias: English learners' attitudes towards english-based loanwords in japan. *The Language Teacher*, 35:7.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Cynthia Dickel Dunn. 1999. Coming of age in japan: Language ideology and the acquisition of formal speech registers. In *Language and ideology: selected papers from the Sixth International Pragmatics Conference*, volume 1, pages 89–97. International Pragmatics Association Antwerp.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *Prague Bull. Math. Linguistics*, 98:25–36.

Jonathan Ferries. 2022. A corpus analysis of loanword effects on second language production. *Englishes in Practice*, 5(1):107–132.

Peter Garrett. 2010. *Attitudes to language*. Cambridge University Press.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Daiki Hashimoto. 2019. Sociolinguistic effects on loanword phonology: Topic in speech and cultural image. *Laboratory Phonology*, 10(1).

Koki Hatagaki, Tomoyuki Kajiwara, and Takashi Ninomiya. 2022. Parallel corpus filtering for japanese text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 12–18.

Mariko Hatanaka and Justin Pannell. 2016. English loanwords and made-in-japan english in japanese. *Hawaii Pacific University TESOL Working Paper Series*, 14:14–29.

Masato Hirakawa, Tomoaki Nakamura, Akira Sasaki, Daisuke Oba, and Shoetsu Sato. 2025. elyza/elyza-thinking-1.0-qwen-32b.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Mana Ihori, Akihiko Takashima, and Ryo Masumura. 2020. Parallel corpus for japanese spoken-to-written style conversion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6346–6353.

Mark Irwin. 2011. Loanwords in japanese.

Ryosuke Ishigami. 2025. Deepseek-r1-distill-qwen-32b-japanese.

Yepai Jia, Yatu Ji, Xiang Xue, Lei Shi, Qing-Dao-Er-Ji Ren, Nier Wu, Na Liu, Chen Zhao, and Fu Liu. 2025. A semantic uncertainty sampling strategy for back-translation in low-resources neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 528–538, Vienna, Austria. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

JMdict Project. 2025. JMdict Japanese–English Dictionary (Yomitan distribution). `https://github.com/yomidevs/jmdict-yomitan`. Accessed 2025-02-16.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. 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 5203–5217, Online. Association for Computational Linguistics.

William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Xiangdong Liu and Todd James Allen. 2014. A study of linguistic politeness in japanese. *Open Journal of Modern Linguistics*, 4(05):651–663.

Leo J Loveday. 1996. *Language contact in Japan: A sociolinguistic history*. Clarendon Press.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of pragmatics*, 12(4):403–426.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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

range3. 2023. Japanese Wikipedia Dump (2023-01-01 version). https://huggingface.co/datasets/range3/wikipedia-ja-20230101. Accessed 2025-02-16.

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.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Mark Spring. 2018. Unconscious gairaigo bias in efl: A case study of japanese teachers of english. *Shinshu University Journal of Arts and Sciences*, 12:166–181.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Stanlaw. 2004. *Japanese English: Language and culture contact*, volume 1. Hong Kong University Press.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Kyoko Takashi. 1990. A sociolinguistic analysis of english borrowings in japanese advertising texts. *World Englishes*, 9(3):327–341.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Takako Tomoda. 1999. The impact of loan-words on modern japanese. In *Japan Forum*, volume 11, pages 231–253. Taylor & Francis.

Toru Urakawa, Yuya Taguchi, Takuro Niitsuma, and Hideaki Tamori. 2024. A Japanese news simplification corpus with faithfulness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 659–665, Torino, Italia. ELRA and ICCL.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

A Models

Open-source models were either run locally on an H100 80GB GPU or accessed via LiteLLM[5], while closed-source models were accessed through their official APIs. All evaluations were conducted under default settings. Full model list is provided in Table 4.

[5] https://docs.litellm.ai/docs/project

Family	Size	Model ID
OpenAI (Closed)	–	gpt-5-2025-08-07
GPT-OSS	120B	openai/gpt-oss-120b
	20B	openai/gpt-oss-20b
DeepSeek	685B	deepseek/deepseek-reasoner
	685B	deepseek/deepseek-chat
Cohere Aya-Expanse	32B	CohereLabs/aya-expanse-32b
	8B	CohereLabs/aya-expanse-8b
ELYZA Shortcut	7B	elyza/ELYZA-Shortcut-1.0-Qwen-7B
	32B	elyza/ELYZA-Shortcut-1.0-Qwen-32B
Qwen-3 Family	30B	Qwen/Qwen3-30B-A3B-Instruct-2507
	235B	qwen/qwen3-235b-a22b-instruct-2507
	4B	Qwen/Qwen3-4B-Thinking-2507
	32B	Qwen/Qwen3-32B
	8B	Qwen/Qwen3-8B
Kimi	1T	kimi-k2-instruct
Deep Analysis Research	32B	deep-analysis-research/D2IL-Japanese-Qwen2.5-32B-Instruct-v0.1
CyberAgent	32B	cyberagent/DeepSeek-R1-Distill-Qwen-32B-Japanese

Table 4: Models evaluated.

B Prompts

You are a data generation bot for a linguistics research project. Your task is to take a Japanese loanword and its traditional counterpart and generate a complete dataset entry.

Your entire output must strictly follow the format specified below. Do not include any introductory text, explanations, headers, or anything else outside of this format.

1. First Line: Provide the single English word that is the common translation for the input pair.
2. Subsequent Six Lines: For each of the six "sentence topics" below, generate one complete data row.
   - Crucially, you must generate a pair of sentences for each topic: one using the loanword and one using the traditional word.
   - Ensure the two sentences are semantically identical and the words are as interchangeable as possible within that context. The only difference should be the target words.
   - Provide a single, shared English translation for the pair.

The six sentence topics are:
- Formal_Conversation
- Formal_Written
- Formal_Explanation
- Informal_Conversation
- Informal_Written
- Informal_Explanation

Output Format:

Your output must be structured in exactly 7 lines: one line for the translated word, and six lines for the data, formatted as follows. Use | as the delimiter.

[Translated Word]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]
[Sentence Topic]|[Sentence with Loanword]|[Sentence with Traditional Word]|[English Sentence]

Input Word Pair:
* Loanword: {loanword}
* Traditional Word: {traditional_word}

Table 5: System prompt for dataset generation using Gemini 2.5 Pro.

You are an AI Japanese language model. Given a single Japanese sentence and the **sentence type** specified by the user (e.g., casual statement, polite request, formal announcement), assess how well the sentence:

1. follows Japanese grammar,
2. sounds natural to native speakers, and
3. suits the intended type in terms of register and word choice nuance.

Respond in exactly two sections:

1. Overall Rating (1-5): Place the single integer score (1-5) inside <score>(1-5)</score> tags.
2. Brief Analysis (1 sentence)
 • State the main issue (grammar, unnatural wording, register).
 • Mention any words you would replace (if any).
 • If the rating is 5, simply note that the sentence is well-formed and appropriate.

Keep the analysis concise; do not add extra sections or explanations.

Table 6: System prompt used for the rating task.

You are an AI Japanese language model. You will be given a context and two Japanese sentences, labeled 'a' and 'b'.

Your task is to determine which sentence sounds more natural and is more appropriate for the given context.

Respond with your choice, 'a' or 'b', inside <choice> tags.

For example:

<choice>[choice]</choice>.

Do not provide any other text, explanation, or punctuation.

Table 7: System prompt used for the comparison task.

You are an AI Japanese language model. Your task is to predict the most likely word to fill the blank space marked with [MASK] in the provided sentence. Based on the context, provide the most probable word that could complete the sentence. Respond ONLY with the word. The word must be enclosed in <option> tags. Do not add any other text, explanations, or numbering. Example format: <option>[option]</option>

Table 8: System prompt used for the open masked prediction task.

You are an AI Japanese language model. Your task is to find the single best Japanese word to fill the [MASK] in a sentence. You will be given the sentence and the target English word that the mask represents. Based on the context of the sentence and the meaning of the English word, provide the single most probable Japanese word. Respond ONLY with the Japanese word, enclosed in <option> tags. Do not add any other text or explanations. Example format: <option>[option]</option>

Table 9: System prompt used for the masked prediction task.

C Rating score based on sentence length

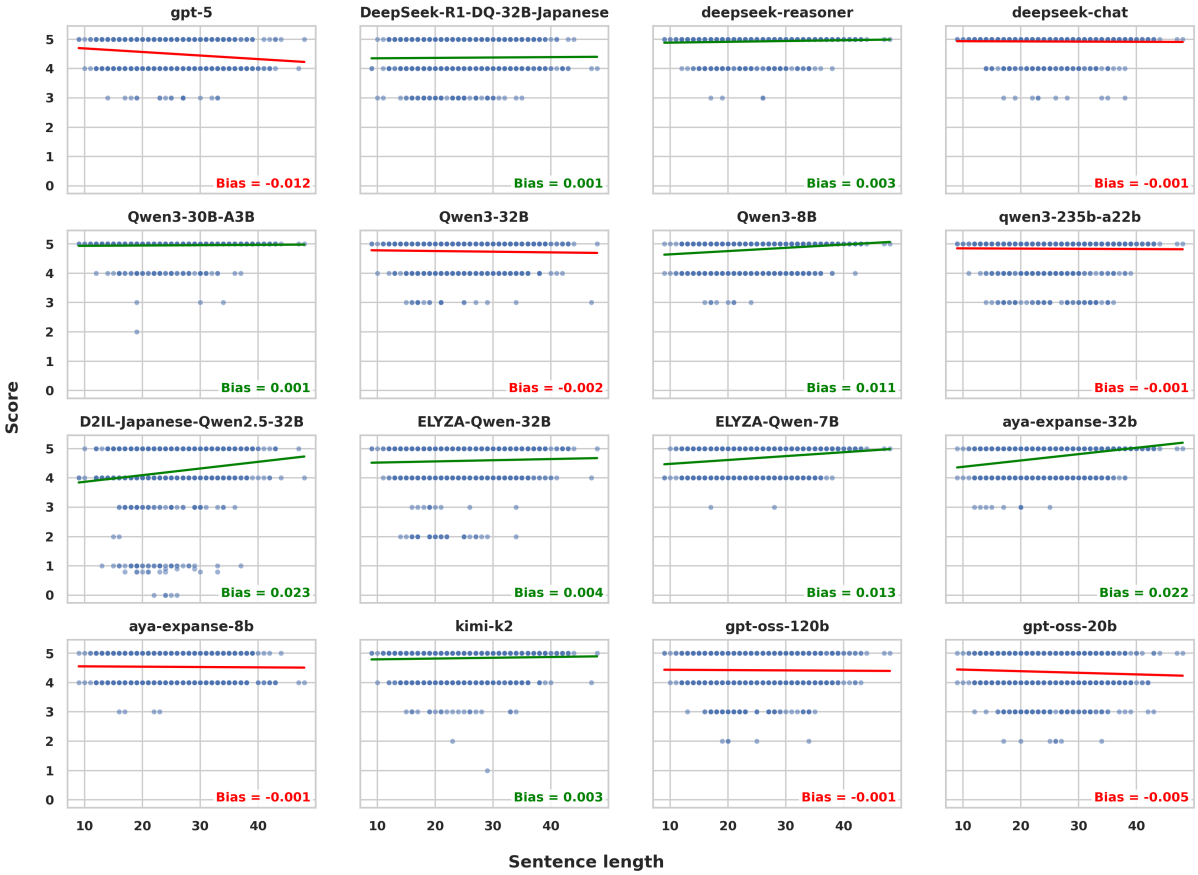


Figure 5: Scores by sentence length (number of characters) across models, where the slope indicates length bias.

Does Local News Stay Local?: Online Content Shifts in Sinclair-Acquired Stations

Miriam Wanner*, **Sophia Hager***, **Anjalie Field**
{mwanner5, shager2, anjalief}@jhu.edu
Johns Hopkins University

Abstract

Local news stations are often considered to be reliable sources of non-politicized information, particularly local concerns that residents care about. The Sinclair Broadcast group is a broadcasting company that has acquired many local news stations in the last decade. We investigate the effects of local news stations being acquired by Sinclair: how does coverage change? We analyze YouTube content put out by local news stations through topic modeling, log-odds ratios, and word embedding analyses to investigate changes after being acquired by Sinclair. We find evidence that local news stations report more frequently on national news at the expense of local topics, and that their coverage of polarizing national topics increases. These findings associate acquisition by Sinclair with increasing polarization and nationalization of news content, which in-turn risks increasing political polarization of local news viewers.

1 Introduction

Historically, local news outlets have played a vital role in the news ecosystem for many Americans by providing information that is community-focused with less perceived partisanship than national outlets. Viewers find local news topics like weather, local crime, and traffic reports important to know about for daily life (Pew Research Center, 2019). American adults also tend to view local news positively regardless of political affiliation, whereas there are stark political divides in opinions about national news (Pew Research Center, 2024). Furthermore, local news consumption has been associated with greater knowledge of local election candidates and increased likelihood of voting for candidates from different political parties for state governor and U.S. president rather than solely along party lines (Moskowitz, 2021).

The Sinclair Broadcast Group, one of the largest broadcasting companies in the United States, owning or operating 185 stations,[1] has acquired a number of local news stations, with purchases primarily concentrated around 2000, 2012-14, and 2016-17. These acquisitions and subsequent observations of news coverage have raised concerns around ways Sinclair is influencing local news. Outside reporters have exposed Sinclair for requiring stations to run specific video segments or to deliver the same scripted speech, and they accused the company of right-wing bias.[2] Researchers have similarly identified conservative bias (Tryon, 2020), and demonstrated that Sinclair stations produce more stories with dramatic elements, commentary, and partisan sources than non-Sinclair stations (Hedding et al., 2019). Concerningly, there is also evidence that Sinclair takeovers actually influenced viewers perceptions of politicians (Levendusky, 2022).

Given the importance of local news and the growing Sinclair influence, we investigate the effect that acquisition by Sinclair has on the content of local news stations. We compare content in news stations before and after Sinclair purchases, and we further draw comparisons with national news outlets. We focus on two levels of analysis:

1. How does overall news differ after purchase?
2. How does coverage of politicized topics differ after purchase?

While a small amount of prior work has compared broadcasts in Sinclair-owned and non-Sinclair stations (Martin and McCrain, 2019; Hedding et al., 2019) or news station websites (Blankenship and Vargo, 2021), Americans are increasingly viewing digital local news, rather than

*Equal contribution.

[1] https://sbgi.net

[2] https://www.nytimes.com/2018/04/02/business/media/sinclair-news-anchors-script.html?searchResultPosition=16

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 37–82

TV Channel	City	Purchased	Affiliation	Youtube	#Videos
TV Channels Purchased by Sinclair					
WSBT-TV	South Bend, IN	02/12/16	Fox	@wsbttv	10624
KECI/KCFW/KTVM	Missoula, MT; Kalispell, MT; Butte, MT	09/01/17	NBC	@NBCMontana	3314
WCTI-TV	Greenville, NC; New Bern, NC; Morehead City, NC	09/01/17	ABC	@WCTI	3320
WCYB	Bristol, VA; Greenville, TN; Johnson City, TN; Kingsport, TN	09/01/17	NBC/The CW	@wcyb5	4620
WLUK-TV	Green Bay, WI	12/19/14	Fox	@Fox11online	19774
WJAR	Providence, RI; New Bedford, MA	12/19/14	NBC	@NBC10WJAR	5044
WGXA	Macon, GA	09/03/14	Fox/ABC	@WGXA	2063
WJLA-TV	Washington, DC	08/01/14	ABC	@7NewsDC	19425
Left- and Right-Wing TV Channels for Comparison					
CNN				@CNN	27560
Fox				@FoxNews	29986

Table 1: Summary data statistics. We collected transcripts from 8 geographically diverse local news YouTube channels that were purchased by Sinclair, as well as transcriptions from YouTube channels for two national outlets.

obtaining it through broadcast television or radio (Pew Research Center, 2024). Thus, we focus on a novel data source: news station YouTube channels, allowing us to uniquely examine the content that news stations choose to highlight on social media and if it reflects trends in broadcast data. Our dataset contains data from eight stations over sixteen years of publishing videos. This construction allows us to examine differences in coverage within the same station before and after acquisition as well as between the larger group of Sinclair-affiliated and non-affiliated stations at any particular point in time. We further include two national news outlets (Fox News and CNN), enabling direct comparisons of Sinclair-owned local news and national news.

We use a combination of corpus analysis methods to examine overall shifts in content and target politicized topics, including comparisons of word choice (Monroe et al., 2008), topic modeling with covariates (Roberts et al., 2013, 2019), and word embeddings analyses (Mikolov et al., 2013; Garg et al., 2018). We find compelling evidence that after purchase, news channels move from covering mostly local topics to politicized national topics. Overall our work offers insight into the content changes associated with Sinclair purchases, thus contributing understanding of how the purchases may influence viewers and highlighting the urgent decline of community-focused news.

2 Related Work

Sinclair Broadcast Group While the Sinclair Broadcast Group's takeover of local news stations attracted public interest and journalism, analysis of content differences in news coverage have been limited to a few prior studies. Martin and McCrain (2019) use topic modeling and comparisons of phrases with U.S. Congressional Record (Gentzkow and Shapiro, 2010) to show Sinclair ownership is associated with a drastic increase in national political coverage over local political coverage and a right-wing shift in ideology. Blankenship and Vargo (2021) similarly find decreased coverage of local news in their analysis of locations mentioned in news stories on Sinclair-owned station websites, though decreasing local news coverage predates Sinclair ownership and coincides with reposted content. In a content-focused analysis, Hedding et al. (2019) find that Sinclair stations produce more stories with dramatic elements, commentary, and partisan sources. None of these studies focus on YouTube data or conduct an in-depth language analysis focusing on politicized topics.

A related line of work has focused on the effects of Sinclair purchases on viewers, without examining content changes in news coverage. Miho (2018) examines the effect of Sinclair ownership on election results. Using event study methodology, they find a 2.5%-point increase in the Republican vote of the 2008/2012 elections, with double the increase in the 2016/2020 elections as a result from exposure to Sinclair content starting in 2004.

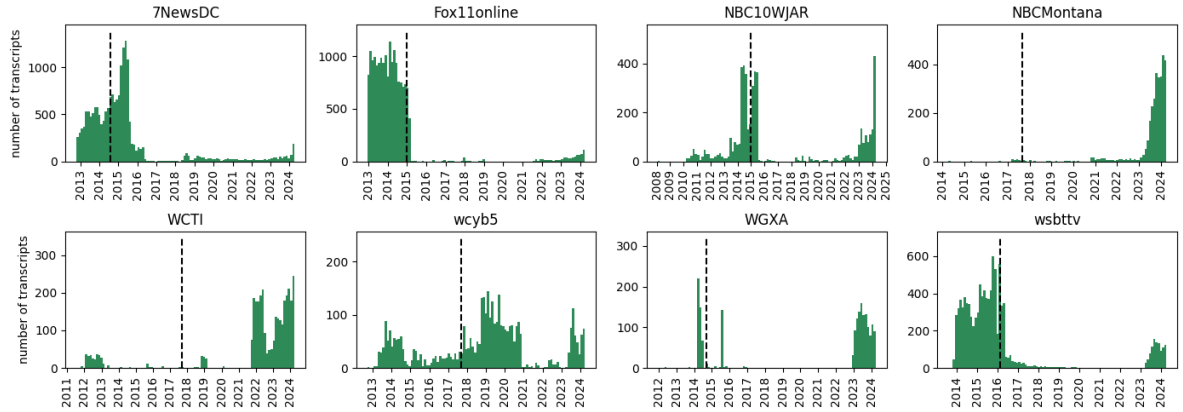


Figure 1: The distribution of the data by year. Vertical lines denote the date that the station was purchased by Sinclair.

Levendusky (2022) use statistical methods to find that living in an area with a Sinclair-owned TV station reduces viewers' approval of President Obama. They find lower approval during his time in office, and additional evidence that viewers are then less likely to vote for the presidential Democratic nominee. These findings that Sinclair purchases are associated with observable changes in preferences of viewers motivate our investigation into understanding the language and content changes that may be driving them.

U.S. Local news Concerning local news more generally, there has been a documented decline in local news organizations, leading to growing "news deserts": areas without consistent news coverage (Abernathy, 2016, 2018). There is evidence that declining local news is contributing to political polarization. Local news consumption is associated with decreased voting exclusively along party lines (Moskowitz, 2021), while increased coverage of local content is associated with lower feelings of political divide (Darr et al., 2021). These factors add further motivation to understanding content and language changes in Sinclair-owned stations. In-depth text analyses of local news have focused coverage of the COVID-19 pandemic (Horne et al., 2022) and the creation of datasets for further investigation (Joseph et al., 2022). These studies are less related to our work but generally validate interest in understanding local news coverage.

U.S. Media polarization Analyses of a variety of text data, including social media posts (Demszky et al., 2019) and political speeches (Card et al., 2022) has uncovered evidence of increasing polarization in the U.S. Despite early evidence of media slant in news articles (Gentzkow and Shapiro, 2010) and many anecdotes about media bias, fewer quantitative analyses have focused on evidence of polarization from video footage. The Stanford Cable TV NewsAnalyzer (Hong et al., 2021) offers an extensive dataset for examining content in three U.S. cable news networks (CNN, Fox, and MSNBC). Ding et al. (2023) use this data to evaluate the semantic polarization in online public discourse, finding that CNN and Fox News cover similar topics, however with varying, distinct contexts, reflecting the polarization between the political leaning of these news stations. They also show that polarization sharply increases around 2016, with its highest peak in 2020, aligning with the death of George Floyd and following Black Lives Matter demonstrations. These findings motivate our use of CNN and Fox News as comparisons datasets in our analysis of local news.

3 Dataset

Collection We construct a new dataset consisting of automated closed captions from YouTube channels of news stations. First, we identified news stations that were acquired Sinclair by starting from an initial list[3] and retaining only stations that (1) have a YouTube channel and (2) began posting videos before they were purchased by Sinclair. We identified 8 local news stations for analysis, as well as Fox News and CNN for comparison. For each station, we download YouTube closed captions for all videos on the channel. We preprocessed this data by converting all transcripts into lowercase

[3] https://en.wikipedia.org/wiki/List_of_stations_owned_or_operated_by_Sinclair_Broadcast_Group

and removing common non-speech tokens or utterances unlikely to provide meaningful signal.[4]

Table 1 reports the full list of stations, their purchase date, and the number of identified videos. Four stations were purchased in 2014, one was purchased in 2016, and three were purchased in 2017. The stations are geographically diverse, reflecting various cities in the east and central U.S. While there is variance in the amount of data from each each station, our data contains at least 2,000 videos for each station. In Figure 1, we further show how the transcripts for each local news station are distributed over time, relative to the data of Sinclair purchase. For some stations (e.g., 7NewsDC, NBC10WJAR) there is a concentration of data just before and just after purchase. For other stations (wcyb5) the data is more dispersed over time.

4 RQ1: How does overall news differ after purchase?

We first use exploratory text analysis methods to broadly examine how news coverage differs before and after Sinclair purchase, as well as in comparison to the two national outlets.

4.1 Methods

We use two primary methods for examining overall news coverage. First, we examine words that are overrepresented in data before purchase as compared to after purchased using log-odds ratio with a Dirichlet prior (method referred to as "Fightin' Words"; from Monroe et al. (2008)). We preprocess the transcripts by filtering out words that do not appear at least ten times in transcripts from every station. A potential confounder is that news changes over time, and our data tends to have more Sinclair-owned stations as time goes on. To mitigate this, we stratify our data by year and only compare log-odds over the years where we have a reasonable amount of paired data (2014, 2015, and 2016). The year with the most paired data, 2014, contains 19,936 videos, 16,640 from stations before they are purchased, and 3,296 from already purchased stations. Paired data decreases in the following couple years, as many stations are purchased during this time period. Purchased stations become more prevalent in 2015 data with 8,641 videos from after purchase, and 4,281 before. This imbalance is more pronounced in 2016 with 696 videos before purchase, and 1,788 after.

[4]">>," "[music]," "[applause]," "uh"

Second, we use topic models to examine coverage changes in clusters of co-occurring words, rather than just individual words. We specifically use the Structured Topic Model (STM) (Roberts et al., 2013, 2019), which is an extension of the popular Latent Dirichlet Allocation (LDA) (Blei et al., 2003), that flexibly incorporates document metadata as covariantes. We chose this model for this property, as well as based on evidence that classical LDA-style models achieve better stability and alignment with human annotations than more recent neural alternatives (Hoyle et al., 2022).

We train five STM models on the transcripts of Sinclair-owned stations, non-Sinclair affiliated stations, Fox, and CNN, in order to determine how the topics discussed by these stations change.[5] For the first four models, we use news-affiliation (which includes four options: Before Sinclair purchase, After Sinclair purchase, CNN, Fox), and date as covariates. We use the convenience function to select a flexible b-spline basis for the date covariate. These four models only differ in the subset of data used. First, we use data from all dates collected. Models 2-4 use data only from 2014, 2015, and 2016, respectively. With these models, we evaluate the topic prevalence, and plot the difference of prevalence along two axes: (1) Before Purchase - After Purchase, and (2) CNN - Fox, in order to highlight topic relationships between Sinclair owned stations and political leaning. We remove the 1% most sparse and common words, and use 30 topics for all models, which we found to have the most coherent topics.

A shortcoming of the STMs introduced thus far is that they assume a fixed vocabulary distribution within topics. If we are interested in comparing the how discourse differs for the same topic, we need to allow these distributions to vary within topics. To study this, we train a 5th model across all the data where we let the influence of Sinclair-affiliated versus non-Sinclair-affiliated (excluding CNN and Fox) be a topical content covariate. We can then then look at the difference in prevalence of words between the content covariate for a given topic. We use the same data filtering and number of topics as in the previous models.

[5]In appendix section A.1 we further conduct a controlled pairwise comparison between two Sinclair-purchased stations and two non-purchased stations to isolate the effect of purchase.

Non-Sinclair	Sinclair
2014	
so, little, it's, it, okay, green, really, bit, bay, nice,yeah, then, we're, going, great, just, can, snow, fun, kind	7, he, virginia, **president**, washington, police, matthew, who, his, was, jury, that, wilson, live, williams, united, robert, said, quarterback, thank
2015	
south, 22, patrick, jennifer, st., kelly, football, james, desk, st, accurate, season, first, watching, downtown,play, says, at, year	you, i, that, 7, okay, **government**, washington, what, we, **trump**, going, island, **president**, virginia, think, **federal**, let's, sam, bay, of
2016	
school, snow, tonight, ice, animals, girls, cold, coach, st., submit, chevy, morning, sale, home, kids, church, temperatures, at, 22, lead	that, **trump**, think, **government**, **federal**, **president**, of, **republican**, i, **sanders**, states, going, security, **campaign**, sort, there's, is, terms, **voters**, **political**

Table 2: Fightin' words results broken down by year. Words that are likely relevant to broader political concerns in the United States have been bolded. Sinclair-purchased stations tend to have more of these words, while stations that are not owned by Sinclair tend to discuss more local concerns.

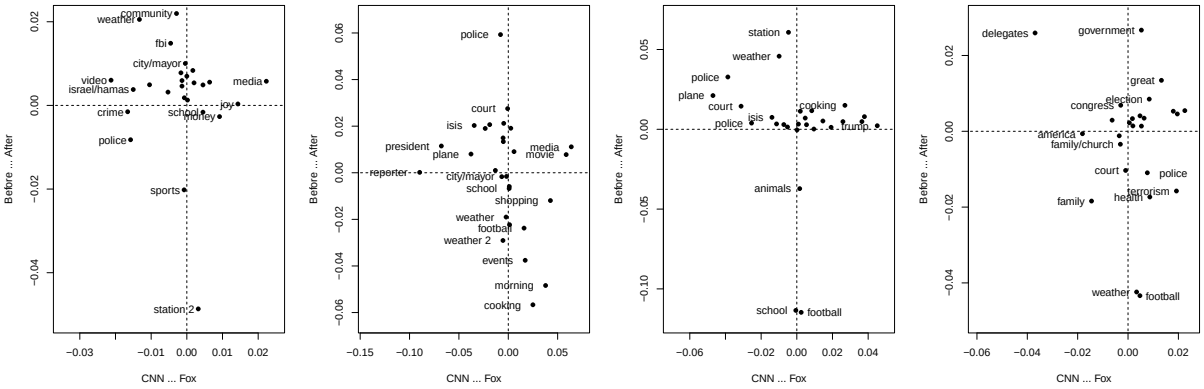
4.2 Results

The Fightin' Words analysis is shown in Table 2. Stations that have been purchased by Sinclair are more likely to use words relevant to national politics, rather than local concerns, using words such as "president," "government," or "federal." Non-Sinclair (pre-purchase) stations were more likely to discuss events that were relevant to local viewers, such as weather, sports, and school, using words like "snow," "downtown," or "school." Some of this may be due to the timing of the purchase coinciding with events occurring in the year (particularly 2016, as an election year). Notably, however, Sinclair-owned stations in 2015 were more likely to discuss the national election than non-Sinclair stations in the election year of 2016, demonstrating that the timing of purchase by Sinclair cannot fully explain the shift to discussing national topics.

Topic Models STM results for selected topics are shown in Figures 2a-2d, where we manually assign representative names for each topic. In the appendix, we report results for all topics (Figures 5-8) and complete lists of probable words for each topic (Tables 6-9). These plots display change in topic proportion between CNN and Fox data, and before and after Sinclair purchase.

Coverage by stations acquired by Sinclair becomes more politicized compared to their reporting before purchase. Topics most widespread on stations not Sinclair affiliated include local topics about the local community, including school (2b, 2a), family (2d), and local events (2b). Discourse on the local environment is also prevalent before purchase with topics including weather (2b, 2d), animals (2c), and health (2d). Topics of local interest include football (2b, 2c, 2d), other sports (2b, 2a), and cooking (2b), and are also mostly featured before purchase. Finally, we see topics around local station small talk, including morning conversation (2b) and station specific language (2a). After stations are acquired, topics predominantly covered shift, with a more political and national focus. Discourse on presidential candidates, including Clinton and Trump are covered at a higher rate on Sinclair-affiliated stations (6), in addition to other national topics like the FBI (2a). In additional to national news, these Sinclair-owned stations also discuss US-relevant international news, including ISIS (2b) and terrorism (7). These Sinclair purchased stations still cover local news including family/church (6, 5), community, city/mayor information, and football (5), indicating that although these stations appear to increasingly report on politicized topics, there is still some local news coverage. A few topics are widely covered and do not tend to align consistently with either before/after purchase. Police and crime are often reported on, and are prevalent both before (2d) and after (2b) coverage, and are particularly aligned with CNN (2a, 2c). This trend is also reflected in proportion of weather discourse, prevalent both before (6) and after (2a) Sinclair purchased stations.

Overall, the difference in topic proportion is less pronounced on the CNN/Fox axis. In fact, topics reported on pre-purchased stations tend to be not clearly aligned with either CNN or Fox, indicating these stations are reporting equally, and possibly very little, on these topics. However, there exists more variation on the CNN/Fox axis in station coverage after Sinclair purchase, and a few are more topics predominantly covered by Fox or



(a) All years (b) 2014 (c) 2015 (d) 2016

Figure 2: Results for STM on all, 2014, 2015, and 2016 data subsets. Change in topic proportion shifting from CNN to Fox on the x-axis and before Sinclair purchase to after Sinclair purchase on the y-axis. Coverage becomes more national and political in stations purchased by Sinclair, with stations before purchase discussing local topics.

CNN. For instance, the conflicts between Israel and Hamas appeared to be reported on more by CNN (2a), in addition to other politicized topics like the president (2b) and delegates in congress (2d). In 2014, Fox reports more on media and movies (2b), which are not inherently political topics. This balance shifts, when in 2015 Trump is a more prevalent topic reported by Fox (2c). This topic shift is further seen in 2016, with topics like ISIS, Cuba, immigration, and presidential candidates being covered by Fox more (8).

Not only can we observe these interactions and changes of topic proportion on the two axes, but we can also see topics shift and appear over time. The Ebola outbreak of 2014 emerges as a topic (6) in the 2014 STM (2b), mostly covered by stations purchased by Sinclair. Leading up to the election, presidential candidates, debates, debate topics (e.g. immigration, china) emerge as topics in 2015 and 2016, similarly covered by Sinclair-owned stations (2c, 2d).

From our STM with non- and Sinclair owned as a topical context covariate, we can observe differences in the language used to discuss certain topics. The results can be found in tables 10-11, and figure 9 in the appendix. Discourse on certain topics differs between stations. Coverage of disasters and violence are covered by before-purchase stations in discussions of healthcare, and after-purchase stations in discussions of topics such as war and terrorism (9b). Similarly, discussions of crime appear typical in stations not affiliated with Sinclair, but purchased stations include discussions of protest and protesters (11). A topic on general government infrastructure (9c) uses language about money (taxes, dollars, etc.) in pre-purchase coverage, and is more national after purchase, using words like pentagon and intelligence. Discourse on legislature differs, with non-Sinclair owned stations talking about local politics including the mayor or local candidates, and Sinclair stations using words like immigration and shutdown, seemingly more national (9d). Discourse on photos and cameras shifts towards a theme of surveillance after purchase, with pre-purchase stations using language like pictures (9a).

5 RQ2: How does coverage of politicized topics differ after purchase?

5.1 Methods

In order to examine shifts in coverage of politicized topics, we train word embedding models and examine properties of embeddings for selected keywords, using similar methodology as prior word embedding analyses (Garg et al., 2018; Rodriguez and Spirling, 2022). We train separate Word2Vec models (Mikolov et al., 2013) for all transcripts of stations before Sinclair purchase and all transcripts after purchase. We additionally train embedding models for data from CNN and Fox News, for left- and right-wing news station comparison.[6]

We curate a set of keywords related to politicized issues for analysis, following Rodriguez and Spirling (2022). We start with their set of words pertaining to policy issues that are debated by po-

[6]For all embedding models we use the following hyperparameters: `window=50`, `min_count=10`, `seed=42`, `workers=16`, `vector_size=100`. Prior work has shown exact parameter settings have little impact on analysis results (Rodriguez and Spirling, 2022; Joseph and Morgan, 2020).

white		black		bias	
Before	After	Before	After	Before	After
black	supremacist	white	africanamerican	chiffonade	implicit
red	supremacy	tan	racial	basil	racial
yellow	secret	hoodie	africanamericans	greens	perpetuate
blue	house	sweatshirt	color	noir	racist
velvet	presidents	wearing	blacks	vinaigrette	discrimination
burgundy	clancy	dark	colored	italian	racially
roses	obamas	colored	racism	mince	quote
orange	obama	stripes	brown	riesling	shameful
colored	pierson	bandana	african	baguette	prejudice
chardonnay	omar	yellow	movement	seedless	language
climate		**equality**		**abortion**	
Before	After	Before	After	Before	After
growth	fuels	rights	freedom	abortions	abortions
industry	emissions	gay	rights	parenthood	roe
uncertainty	global	marriage	prolife	privileges	reproductive
algae	fossil	religious	dignity	admitting	wade
economy	everglades	democracy	equal	ultrasound	prolife
regionally	sustainability	moral	freedoms	gyn	prochoice
ratings	pollution	civil	unborn	prolife	overturning
potential	droughts	marriages	racism	clinic	marriage
consumption	environmental	supreme	democracy	clafer	affirming
economic	impacts	samesex	lgbt	pregnancies	incest

Table 3: Ten nearest neighbors to the query word (bold) for stations that were not owned by Sinclair (Before) and Sinclair-owned stations (After). Sinclair-owned stations are more likely to have nearest neighbors that are politically charged.

litical parties and motivate voting: "immigration," "abortion," "welfare," "taxes." We add words relating to policy issues not covered in their original set, including words related to racial bias ("racism," "bias," "black," "white"), "climate," "police," "military," and "guns." We further include words that Rodriguez and Spirling (2022) curate as expected to solicit different response in different people: "democracy," "freedom," "equality," "justice," "republican," and "democrat," though they are less relevant to our focus on politicized topics. We identify the 10 nearest neighbors for each keyword in the before-purchase and after-purchase embedding models using cosine similarity.

Embedding Similarity We conduct an analysis of whether increased politicization can be noted in our learned word embeddings when put in context of national news stations. We examine how similar word embeddings are for the seed words described in subsection 5.1 in comparison to two national news sources, Fox News and CNN. We choose Fox News and CNN in particular as they are considered to be politically polarized (Ding et al., 2023) and thus are likely to be talking about politically polarizing issues. In this experiment, we train embedding models on data from stations before purchase and after purchase specifically between the years 2014-2016 and evaluate similarity between embeddings from before/after purchase trained models with models trained on Fox News/CNN transcripts from the same time period. We query these models with the mentioned seed words, and align the embedding spaces and their vocabularies using the Procrustes transformation. We then calculate cosine similarity between embedding vectors for the same word, to determine whether local news outlets tend to be more likely to discuss these words in similar contexts to the polarized national news outlets.

5.2 Results

Nearest Neighbor Analysis We show the six most interpretable nearest neighbor results in Table 3, with the remaining twelve less-interpretable results presented in Table 4. We find that the embedding model trained on transcripts from stations after being purchased by Sinclair tends to have nearest neighbors to our query words that are more overtly politically charged. The first three words we show,"white," "black," and "bias," demonstrate the clearest movement towards polarizing rhetoric. Before purchase, "white" and "black" are generally associated with other colors and patterns (e.g."red," "stripes") or items that might be that color (e.g."chardonnay," "roses"). The embedding model trained on data after Sinclair

acquisition associates black mostly with words pertaining to race ("africanamerican," "racism"), as well as "movement," likely relevant to protest movements such as Black Lives Matter. While "supremacist"/"supremacy" are the nearest neighbors to "white," indicating increased use of white as a racial descriptor, many of the nearest neighbors appear to be relevant to the presidency ("house," "obama") indicating an increased discussion of White House policy and national political news. We note a similar result with the query word "bias." The nearest neighbors before acquisition associate "bias" predominantly with cooking, and with cutting in particular ("chiffonade," "mince"), likely due to the phrase "cutting on the bias" being frequently used as an instruction in cooking videos. The model trained on data post-acquisition associates "bias" with words more evocative of societal bias ("implicit," "prejudice").

Table 3 also displays results for "climate," "equality," and "abortion," which demonstrate some signs of increased politicization, but may be influenced by the confounding factor of time. For instance, the post-acquisition embedding model was more likely to associate "climate" with words referencing climate change ("emissions," "pollution"), while the pre-acquisition model generally referenced other topics ("growth," "economy,"). This may indicate increased discussion of global warming, but may also be influenced by increased discussion of climate change in recent years. The pre-acquisition model associates "equality" with an assortment of words which suggest discussions of Obergefell v. Hodges (2015)[7], such as "marriages" and "supreme." After, some neighbors are relevant to reproductive justice ("prolife," "unborn"), possibly due to conversations about Dobbs v. Jackson Women's Health Organization (2022)[8]. While this may indicate increased discussion of reproductive justice, it also may demonstrate confounds of this data. This case may also explain the shift in nearest neighbors to "abortion"; the pre-acquisition model associates "abortion" with more healthcare-related words (e.g."admitting"+"privileges," "ultrasound") while the post-acquisition embedding model associates it more with politicized rhetoric around abortions (e.g."prolife," "prochoice").

In summary, the analysis of the differences in embedding models trained on these transcripts offers clear evidence that stations owned by Sinclair discuss polarizing political issues more than ones that have not been purchased. While we cannot necessarily attribute Sinclair purchase as the sole cause of this coverage shift, it nevertheless indicates increasing politicization of local news.

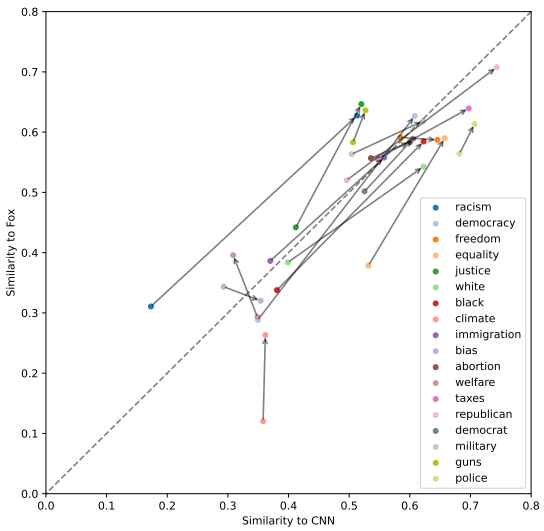


Figure 3: Comparing embedding similarity for our target words to embeddings for CNN and Fox News between the years 2014-2016. Arrows show the shift for embeddings trained on data before acquisition to embeddings trained on data after acquisition. There is a clear trend towards increasing nationalization– similarity tends to increase to both national news outlets.

Embedding Similarity Figure 3 shows how similarity with our seed words changes from before purchase to after purchase. We find that there is a general trend towards increasing similarity between our polarized news outlets and local news stations after acquisition by Sinclair, with the majority of our seed words showing increased similarity to both outlets. Almost all the arrows direct up and to the right along the central diagonal line. This shift indicates that the usage of these words becomes more similar to both Fox and CNN after Sinclair purchase. Only three words, ("freedom," "welfare," and "bias"), decrease noticeably in similarity to one of the national broadcasters, and each increases in similarity to the other broadcaster.

[7] https://www.justice.gov/sites/default/files/crt/legacy/2015/06/26/obergefellhodgesopinion.pdf, a Supreme Court decision which legalized gay marriage in the United States

[8] https://www.supremecourt.gov/opinions/21pdf/19-1392_6j37.pdf, a Supreme Court decision which overturned Roe v. Wade by asserting there was no constitutional right to abortion

Thus, there is a clear shift towards increasingly national language as opposed to local language. Unlike the nearest neighbor results, we limit both the national and local station data to 2014 through 2016, minimizing the effect that different news events during different time periods has on our analysis. We do not observe a shift towards either national outlet in particular. While prior work has observed right-wing slant in Sinclair-owned stations (Tryon, 2020), we suspect word embeddings are not sufficient to capture this trend, as they have a limited context window size. Fox and CNN have also been measured as less polarized before 2020 (Ding et al., 2023), suggesting that comparisons against these outlets may not be sufficient for capturing slant. Nevertheless, this figure demonstrates that after purchase by Sinclair, stations tend to use these words in contexts similarly to polarized national news stations.

6 Discussion

Connection to Communications Theory Communications scholars have identified agenda setting and framing as tools for influencing public opinion, which can be used to characterize media bias (Entman, 2007). While agenda-setting broadly encompasses ways the media reports on some events at the exclusion of others (e.g., *what* topics are covered), framing involves highlighting specific aspects of a topic or event in order to promote a particular interpretation (e.g., *how* topics are covered) (McCombs and Shaw, 1972; Entman, 2007), though these two mechanisms are not always distinct (McCombs and Ghanem, 2001). When we consider our analysis through this lens, we note second-order agenda-setting strategies in our results. Increasing national focus, as well as focus on polarizing political issues, primarily occurs through agenda-setting, which is evident in topic-level changes (Figure 2a-Figure 2d) and politicized word usage (Table 3). Our results reveal some evidence of framing changes in vocabulary shifts within topics (Table 11), but future work targeting framing specifically is needed to fully explore these trends. In contrast to prior work (Tryon, 2020), we do not find clear evidence of right-wing slant: our results do not consistently show Sinclair ownership is associated with more similarity to Fox than CNN. This may be explained by several factors, such as differences in what content news stations highlight on YouTube, limited polarization in CNN and Fox in the years we focus on (Ding et al., 2023), or inability of our methods to capture nuanced coverage differences over broad changes in topics. Future work targeting framing could aid in disentangling these factors (Entman, 2007).

Implications for viewers In addition to framing and agenda setting, a third tenet of media's distribution of power and influence over public opinion is *priming*: the effects agenda-setting and framing have on the audience (Entman, 2007). While our study does not measure priming effects, previous work has connected the influence of Sinclair to material changes in partisan voting and to changed opinions of politicians (Miho, 2018; Levendusky, 2022). The evidence that we uncover of shifts in digital content after purchase by Sinclair offers insight into the possible mechanisms leading to public opinion changes associated with Sinclair ownership. Our establishment of YouTube videos as a data source for analyzing content shifts also offers opportunities to studying priming. In future work, comments on YouTube videos could offer a way to directly examine viewers responses to specific content. This data could also be crossed with other social media sources, such as what links are shared on other platforms.

7 Conclusion

Across all text analysis methods, we find consistent evidence that acquisition by Sinclair is associated with increased coverage of national and political news, often at the expense of conventional local news topics such as cooking or local sports. Combined with the increasing closure of local news outlets, our results offer a grim picture of the decline of community-focused news in the U.S. We further demonstrate the usefulness of YouTube data in measuring and understanding this trending, thus highlighting opportunities for follow-up research.

8 Limitations

While we target a causal question (the impacts of Sinclair purchase), our work is observational, which reduces our ability to draw causal conclusions. There are potential unobserved confounders, including general shifts in coverage over time across all news outlets, possibly driven by industry-wide efforts to attract views and increase engagement. We do take steps to correct for industry trends over time, including segmenting data by time in Table 2 and constructing a tightly controlled

paired analysis (in appendix section A.1). However, these settings require restricting to smaller subsets of the data, which limits the analyses we can conduct. Our dataset generally contains imbalances which could impact results, e.g., the distribution across stations and time is uneven. Overall, while the consistency of content shifts coinciding with the timing of Sinclair purchases strongly suggests a causal relationship, we cannot definitively rule out the possibility of further confounders. Finally, although we draw from previous work, our interpretations somewhat rely on subjective and personal judgments about what words are politicized. These choices have an impact on our conclusions.

References

Penelope Muse Abernathy. 2016. *The rise of a new media baron and the emerging threat of news deserts*. Center for Innovation and Sustainability in Local Media, University of North ....

Penelope Muse Abernathy. 2018. *The expanding news desert*. Center for Innovation and Sustainability in Local Media, School of Media and ....

Justin C. Blankenship and Chris J. Vargo. 2021. The effect of corporate media ownership on the depth of local coverage and issue agendas: A computational case study of six sinclair tv station websites. *Electronic News*, 15(3-4):139–158.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Joshua P Darr, Matthew P Hitt, and Johanna L Dunaway. 2021. *Home style opinion: How local newspapers can slow polarization*. Cambridge University Press.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaohan Ding, Michael Horning, and Eugenia Rho. 2023. Same words, different meanings: Semantic polarization in broadcast media language forecasts polarity in online public discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:161–172.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Kylah J Hedding, Kaitlin C Miller, Jesse Abdenour, and Justin C Blankenship. 2019. The sinclair effect: Comparing ownership influences on bias in local tv news content. *Journal of Broadcasting & Electronic Media*, 63(3):474–493.

James Hong, Will Crichton, Haotian Zhang, Daniel Y. Fu, Jacob Ritchie, Jeremy Barenholtz, Ben Hannel, Xinwei Yao, Michaela Murray, Geraldine Moriba, Maneesh Agrawala, and Kayvon Fatahalian. 2021. Analysis of faces in a decade of us cable tv news. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery.

Benjamin D Horne, Maurício Gruppi, Kenneth Joseph, Jon Green, John P Wihbey, and Sibel Adalı. 2022. Nela-local: A dataset of us local news articles for the study of county-level news ecosystems. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1275–1284.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kenneth Joseph, Benjamin D Horne, Jon Green, and John P Wihbey. 2022. Local news online and covid in the us: relationships among coverage, cases, deaths, and audience. In *Proceedings of the International AAAI Conference on Web and social media*, volume 16, pages 441–452.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Matthew S. Levendusky. 2022. How does local tv news change viewers' attitudes? the case of sinclair broadcasting. *Political Communication*, 39(1):23–38.

Gregory J. Martin and Joshua McCrain. 2019. Local news and national politics. *American Political Science Review*, 113(2):372–384.

Maxwell McCombs and Salma I Ghanem. 2001. The convergence of agenda setting and framing. In *Framing public life*, pages 83–98. Routledge.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Antonela Miho. 2018. Small screen, big echo? estimating the political persuasion of local television news bias using the sinclair broadcasting group as a natural experiment. *SSRN Electronic Journal*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Daniel J. Moskowitz. 2021. Local news, information, and the nationalization of u.s. elections. *American Political Science Review*, 115(1):114–129.

Pew Research Center. 2019. Older americans, black adults and americans with less education more interested in local news. Technical report, Washington, D.C.

Pew Research Center. 2024. Americans' changing relationship with local news. Technical report, Washington, D.C.

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An r package for structural topic models. *Journal of statistical software*, 91:1–40.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, and 1 others. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Pedro L. Rodriguez and Arthur Spirling. 2022. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):101–115.

Chuck Tryon. 2020. Sinclair broadcasting as minimedia empire: Media regulation, disinfomercials, and the rise of trumpism. *Media, Culture & Society*, 42(7-8):1377–1391.

A Appendix

A.1 Paired Analysis

This paper has considered data from the same stations before and after purchase by Sinclair, allowing for control over confounding variables such as differing political content between stations. However, this also introduces time as a significant confounder, as stations' coverage of various news stories will obviously vary over time as current events unfold. We conduct an additional analysis in which we more tightly control for possible news variance over time, aiming to isolate the effects of purchase. We choose two stations in our dataset with similar nearby stations which were never Sinclair affiliated. We choose KECI/KCFW/KTVM (YouTube @NBCMontana) with KPAX-TV (YouTube @kpaxmissoula) as the non-Sinclair affiliated channel, both in western Montana (MT), and WCYB (YouTube @wcyb5) with WJHL-TV (YouTube @WJHLtv11), both in the middle of the Virginia-Tennessee (VA-TN) border. For the non-Sinclair affiliated stations, we scrape the same number of videos as in the respective Sinclair affiliated channel, scraping videos closest to the videos in the Sinclair affiliated channels. We also subsample the CNN and Fox data in the same way, selecting the subset in each closest to the videos in the Sinclair affiliated channels.

A drawback of analyzing these paired stations is that local news stations may copy each other's content. It is therefore possible that Sinclair's purchase of a local station also impacts content posted on other local stations, violating the condition of no interference. Regardless, this analysis provides additional insight: in preceding analyses, interference is less of a concern, but controls for time are less strict. Similar trends in both settings would lend support to the conclusion that Sinclair purchases impact coverage of local topics.

Methods We train STMs, as in **RQ1**. We train two new STMs on the transcripts from the two local MT stations, two local VA-TN stations, and subsampled CNN and Fox data. We used time, and the before/after and Sinclair affiliated/non-affiliated subsets as covariates. The first STM is trained with all data, and the second with data from before 2020, as we aimed to study the Sinclair effect without the dominant pandemic topic.

Results STM results for selected topics are shown in Figures 4 and 10. As before, we manually assign representative names for each topic based on the full topics listed in Tables 12-13. We additionally assigned local/national/unclear topic labels. These plots display change in topic proportion between non-Sinclair affiliated channels and Sinclair affiliated channels before and after the purchase date of the Sinclair affiliated channel.

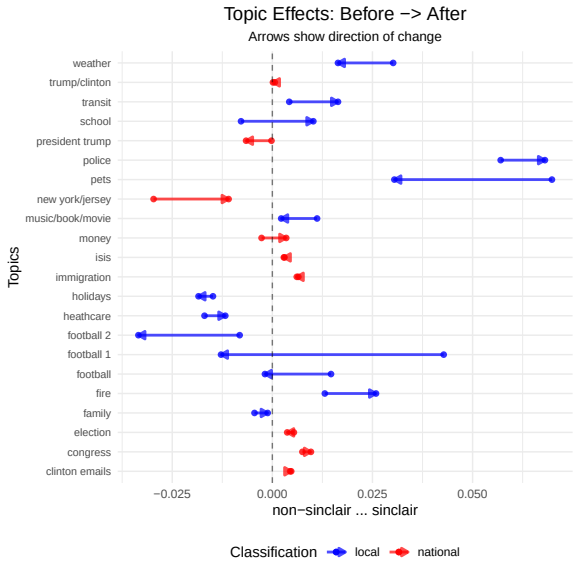


Figure 4: Results for STM on **paired data before 2020**. Change in topic proportion shifting from non-Sinclair to Sinclair affiliate on the x-axis, and shift before and after purchase date is shown with arrows. Red denotes national topics and blue denotes local topics. Topic list is shown on the y-axis. Topics with unclear national/local interpretation are omitted here, and included in Figure 12.

Topic modeling on all data demonstrates the prominence of pandemic coverage, which seems to dominate the topic proportion for dates after purchase, and is disproportionately covered by Sinclair owned channels. We also observe a shift away from local coverage for Sinclair-owned stations, which is likely due to Sinclair stations reporting less on local issues. STM results on paired data before 2020 demonstrate this same shift towards local coverage for non-Sinclair affiliated stations, especially seen in topics covering football and pets. We see less shift in national topics between channels, but do observe a small shift towards Sinclair-owned stations.

A.2 Additional nearest neighbor results

Table 4 shows results for our remaining query words. We find that these results demonstrate less interpretable shifts in ideology or framing.

racism		**democracy**		**freedom**	
Before	After	Before	After	Before	After
racist	racial	equality	extremism	freedoms	democracy
alleges	injustice	minister	freedom	sacrifice	freedoms
appointed	racist	civil	freedoms	equality	equality
dismissed	equality	unrest	nation	pride	sacred
consulted	rhetoric	america	values	faithful	liberation
pape	africanamerican	conflict	america	nation	slavery
vulgar	hatred	islamic	radical	civil	symbol
derogatory	movement	humanitarian	equality	1963	birthright
goodell	protesting	muslim	ideology	democracy	nation
judgment	muslim	nations	defend	rights	ideals
justice		**immigration**		**welfare**	
Before	After	Before	After	Before	After
prosecution	judicial	immigrants	undocumented	claims	services
argued	injustice	reform	immigrants	petition	leno
innocent	criminal	undocumented	reform	misuse	westmoreland
civil	prosecute	congress	deportation	claiming	mandated
prosecutor	innocence	citizenship	obamacare	deny	systematically
selfdefense	accountable	nra	repeal	abuse	lenor
actions	prejudice	conservatives	deport	seek	medicaid
conviction	collective	unaccompanied	repealing	fraud	stamps
testify	dignity	border	enact	status	cheri
judicial	accountability	bipartisan	latinos	coverup	care
taxes		**republican**		**democrat**	
Before	After	Before	After	Before	After
tax	tax	democrat	democratic	republican	democratic
income	income	democratic	democrats	candidate	delegate
taxpayers	debt	candidate	gop	democratic	republican
budgets	corporations	grothman	republicans	incumbent	incumbent
debt	costs	representative	democrat	reelection	partisan
rates	revenue	senator	candidate	campaigning	reelection
pension	wealthy	congressman	conservative	primary	congressman
paying	taxpayers	reelection	caucus	nomination	hollen
fees	burden	politics	partisan	romney	democrats
fiscal	deferral	democrats	electorate	mitt	candidate
military		**guns**		**police**	
Before	After	Before	After	Before	After
troops	army	weapons	firearms	authorities	mpd
iraq	soldiers	gun	gun	investigators	officers
afghanistan	afghanistan	firearms	handgun	witnesses	authorities
civilian	combat	rifles	firearm	detectives	detectives
combat	troops	ammunition	semiautomatic	deputies	juveniles
navy	marines	firearm	illegal	mpd	cops
forces	armys	handguns	weapons	sources	suspects
soldier	soldier	semiautomatic	rifles	officers	patrols
iraqi	forces	concealed	handguns	suspects	deputies
soldiers	overseas	rifle	criminals	suspect	suspect

Table 4: Nearest neighbors of other words. These words demonstrated no significant interpretable changes.

A.3 Dataset example

Table 5 shows an example of how our data is structured.

A.4 Topic Modeling Details

Tables 6, 7, 8, and 9 contain columns with words from various methods for evaluating top words for topics from a structured topic model. Highest Prob denotes the highest probability words. FREX denotes the highest ranking FREX (FRequency and EXclusivity) words. Lift denotes the highest scoring words by lift, which weights words and divides by their frequency in other topics. This gives higher weight to words less frequent among other topics. Score denotes the best words by score by dividing the log frequency of the word in the topic by the log frequency of the word in other topics. The "Label" column contains our manually written labels for each topic.

Channel	Title	URL	Date	Transcript
wsbttv	Cold temperatures could impact fruit production	https://www.youtube.com/watch?v=6zEWpCJ95hc	2024-03-24T16:00:11Z	typically Mother's Day is when apple orchards across the area start to see their crops in full bloom at ker Sunrise Orchards though they're already a month ahead of schedule with...

Table 5: Example from the scraped YouTube data

Table 6: Topics for All-Time STM. See section appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**police**	police, fire, car, say, officers, officer, scene	fire, officer, scene, car, police, crash, officers	submit, firefighters, gunman, crash, transported, driver, flames	police, submit, officers, officer, car, scene, fire
2	**law/ crime**	law, people, gun, crime, enforcement, violence, police	violence, guns, enforcement, illegal, gun, law, immigrants	climb, criminals, guns, firearms, marijuana, sanctuary, violence	climb, law, enforcement, crime, police, immigration, violence
3	**football**	first, get, back, right, going, one, got	ball, yards, tennessee, quarter, touchdown, videos, play	videos, touchdown, yards, snap, tennessee, ball, bristol	videos, touchdown, ball, game, yards, coach, quarterback
4	**video**	see, right, just, video, can, back, one	video, phone, saw, sir, correct, pictures, yes	video, footage, phone, recording, phones, images, cameras	video, phone, sir, yes, okay, correct, see
5	**city/ mayor**	new, city, will, mayor, york, building, nbc	mayor, city, bridge, nbc, project, providence, council	champion, mayor, bridge, providence, construction, citys, city	champion, city, mayor, providence, montana, nbc, york
6	**court**	court, case, judge, trump, attorney, justice, will	jury, judge, attorney, trial, hunter, court, indictment	testifying, indictment, jury, mar-alago, prosecution, lawyers, merrick	testifying, trump, court, hunter, jury, attorney, documents
7	**president**	president, going, house, white, said, foreign, just	foreign, president, presidents, white, administration, congress, secretary	foreign, presidents, cabinet, president-elect, broadly, bipartisan, summit	foreign, president, congress, obama, presidents, administration, secretary
8	**fbi**	information, department, government, fbi, security, committee, report	fbi, information, committee, chairman, agencies, agency, data	trusted, inspector, cyber, agency, breach, agencies, privacy	trusted, fbi, investigation, government, information, committee, intelligence
9	**-**	gtgt, reporter, said, people, gtgtgt, dont, say	gtgt, reporter, gtgtgt, cnn, jake, plane, ten	gtgt, gtgtgt, liar, reporter, brooke, jake, cnn	gtgt, liar, gtgtgt, reporter, cnn, e-mail, e-mails

Topic	Label	Highest Prob	FREX	Lift	Score
10	**cooking**	just, little, going, can, okay, like, right	cheese, chicken, okay, cream, recipe, sauce, little	garlic, oven, recipes, sauce, chocolate, flavors, toss	toss, cheese, recipes, sauce, okay, garlic, gonna
11	**israel/ hamas**	israel, ukraine, russia, war, military, iran, will	hamas, gaza, israel, israeli, putin, iran, russia	casualties, gaza, hamas, israelis, missiles, israeli, palestinians	ukraine, hamas, israel, casualties, russia, gaza, iran
12	**country**	country, people, america, will, american, states, united	america, freedom, nation, rights, veterans, country, abortion	thus, values, liberty, freedom, religious, dignity, marriage	thus, america, country, americans, democracy, american, abortion
13	**clinton**	clinton, hillary, shes, campaign, debate, said, obama	clinton, hillary, clintons, sanders, bernie, emails, campaign	wore, clintons, clinton, hillary, bernie, sanders, server	clinton, hillary, wore, clintons, sanders, obama, bernie
14	**joy**	like, show, love, guy, greg, one, shes	greg, jesse, movie, laughter, guy, funny, film	joy, movie, movies, jeanine, jesse, greg, song	joy, greg, jesse, laughter, movie, love, jeanine
15	**-**	know, think, going, dont, thats, like, just	know, think, mean, dont, youre, thats, theres	gosh, mean, neil, nobodys, know, think, youre	gosh, know, think, mean, going, dont, youre
16	**biden**	biden, joe, border, house, democrats, president, republicans	biden, joe, border, mccarthy, bidens, democrats, harris	maga, ainsley, pelosi, kamala, kayleigh, newsom, mccarthy	biden, maga, border, democrats, republicans, bidens, joe
17	**money**	money, million, dollars, pay, tax, bill, state	dollars, tax, pay, money, budget, taxes, million	friendship, taxpayers, income, revenue, tax, dollars, budget	friendship, tax, money, dollars, budget, taxes, million
18	**isis**	isis, attack, iraq, syria, now, attacks, war	isis, iraq, syria, terror, terrorist, terrorism, muslim	islam, isis, brutal, qaeda, syrian, refugees, iraqi	isis, brutal, syria, iraq, terrorist, terror, islamic
19	**station**	news, now, says, county, live, tonight, today	abc, maryland, metro, wsbt, county, bend, news	heal, suzanne, fairfax, allison, arlington, patrice, georges	county, heal, wsbt, abc, news, metro, fairfax
20	**health**	health, can, care, get, medical, now, hospital	cancer, doctor, doctors, disease, patients, health, medical	disease, ear, diagnosed, symptoms, virus, doctors, vaccine	ear, health, patients, cancer, disease, hospital, vaccine

Topic	Label	Highest Prob	FREX	Lift	Score
21	**trump**	trump, donald, hes, republican, think, election, going	voters, candidates, polls, trump, donald, iowa, republican	odd, rubio, marco, electorate, rnc, mitt, romney	trump, donald, odd, republican, voters, election, republicans
22	**china**	china, now, energy, world, company, new, chinese	energy, china, market, climate, prices, chinese, industry	aim, pipeline, industry, prices, consumers, electric, markets	aim, china, chinese, prices, economy, inflation, climate
23	**weather**	going, see, will, now, right, weather, well	snow, storm, weather, temperatures, rain, storms, winds	thunderstorms, temperatures, cloudy, crew, storms, showers, snow	crew, snow, temperatures, storms, rain, weather, storm
24	**station 2**	morning, year, fox, green, people, well, can	bay, christmas, green, holiday, wisconsin, rachel, event	tap, zoo, museum, birds, peterson, parade, fishing	tap, bay, appleton, oshkosh, fox, wisconsin, christmas
25	**school**	school, kids, students, children, parents, schools, high	students, campus, school, schools, teachers, student, parents	alan, campus, campuses, teacher, superintendent, teachers, students	school, students, alan, kids, schools, parents, student
26	**sports**	game, team, year, play, one, win, season	sports, game, games, fans, players, football, team	exclusive, notre, dame, tournament, nfl, baseball, irish	exclusive, game, notre, dame, football, coach, sports
27	**crime**	found, say, case, man, two, -year-old, death	murder, -year-old, arrested, prison, victim, sexual, charged	cop, sexually, sentenced, dna, sexual, allegedly, murder	cop, murder, investigators, police, charges, arrested, -year-old
28	**community**	can, need, work, people, will, community, make	community, help, work, rhode, working, resources, need	medicine, rhode, resources, partnership, community, nonprofit, resource	medicine, community, rhode, thank, need, communities, families
29	**media**	media, people, news, fox, said, saying, story	media, twitter, racist, social, post, youtube, tucker	usa, musk, racist, elon, carlson, twitter, tucker	usa, media, twitter, racist, musk, fox, elon

Topic	Label	Highest Prob	FREX	Lift	Score
30	**family/ church**	family, life, just, years, day, know, time	family, life, father, church, loved, mom, friends	hats, pastor, sisters, grandfather, pray, remembered, jesus	hats, family, father, life, son, mom, mother

Table 7: Topics for 2014 STM. See appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**isis**	isis, iraq, military, syria, united, israel, will	isis, syria, israel, iraq, hamas, gaza, israeli	hamas, iraqi, islamic, joy, qaeda, syria, syrian	isis, joy, iraq, syria, hamas, gaza, israel
2	**movie**	north, movie, film, korea, show, kim, theater	korea, movie, movies, film, theater, kim, hollywood	brutal, theaters, korea, movies, cyber, comedy, movie	brutal, movie, korea, film, hollywood, north, theaters
3	**ukraine/ russia**	ukraine, russia, russian, putin, bridge, key, president	ukraine, russia, russian, putin, calm, bridge, ukrainian	calm, putin, russia, russian, russians, ukraine, ukrainian	calm, ukraine, russia, russian, putin, ukrainian, sanctions
4	**video**	video, phone, see, security, camera, call, cell	video, phone, cell, camera, cameras, tape, ray	video, cell, ray, roger, phones, phone, cameras	video, phone, nfl, cell, cameras, camera, surveillance
5	**court**	court, case, says, said, today, attorney, now	judge, court, attorney, documents, prosecutors, trial, guilty	los, courtroom, judge, allegations, pleaded, sentenced, lawsuit	los, court, attorney, judge, prosecutors, charges, investigation
6	**nbc**	last, two, one, three, night, ago, years	nbc, last, night, island, three, ago, providence	unlikely, patrice, nbc, providence, susie, tony, frank	unlikely, nbc, providence, rhode, island, last, night
7	**health/ ebola**	health, ebola, care, hospital, medical, now, patients	ebola, patients, disease, health, doctors, patient, virus	patients, professionals, cdc, ebola, outbreak, virus, infected	ebola, professionals, health, hospital, patients, virus, disease
8	**city/ mayor**	city, will, new, says, mayor, building, now	mayor, city, council, project, property, marijuana, construction	authority, mayor, mayors, council, marijuana, city, construction	authority, city, mayor, council, project, property, marijuana

Topic	Label	Highest Prob	FREX	Lift	Score
9	**shopping**	can, like, one, use, get, just, new	shop, buy, store, stores, sell, items, products	spicy, app, products, stores, shop, shopping, plastic	spicy, store, can, products, shop, stores, food
10	**police**	police, say, man, now, county, live, tonight	suspect, prince, police, victim, investigators, -year-old, georges	basement, stabbed, detectives, plater, roz, year-old, gunman	police, basement, county, investigators, suspect, victim, abc
11	**-**	think, people, dont, thats, going, want, can	think, question, welcome, dont, understand, sort, talk	welcome, frankly, perspective, shouldnt, agree, necessarily, legitimate	welcome, think, people, dont, question, want, youre
12	**money**	money, state, million, dollars, pay, company, business	jobs, million, tax, money, dollars, pay, budget	chinese, manufacturing, taxpayers, taxes, taxpayer, tax, jobs	chinese, dollars, tax, money, million, jobs, taxes
13	**weather**	water, ice, says, winter, river, power, lake	ice, warning, water, river, trees, fish, boat	warning, dnr, boats, fishing, swimming, flooding, trees	warning, water, ice, winter, lake, fish, trees
14	**cooking**	going, just, little, okay, can, like, really	recipe, sauce, cheese, flavor, chicken, cream, butter	teaspoon, butter, flavor, flour, onion, onions, oven	mustard, sauce, recipe, cheese, recipes, flavor, cream
15	**-**	know, like, just, really, think, got, dont	know, mean, yeah, like, ive, really, hes	wow, weird, nervous, mean, know, guess, magic	wow, know, yeah, like, think, hes, mean
16	**family**	family, life, just, children, son, shes, child	son, family, father, child, mother, life, daughter	sleep, father, son, brother, sister, child, daughter	sleep, family, child, son, mother, children, father
17	**wsbt**	county, says, south, wsbt, channel, bend, kelly	wsbt, bend, joseph, dog, elkhart, desk, channel	wsbts, registered, fillmore, wsbs, denise, elkhart, joseph	registered, county, wsbt, bend, elkhart, kelly, joseph
18	**football**	game, team, play, one, season, win, football	football, game, sports, players, games, win, team	congratulations, playoffs, touchdown, redskins, quarterback, playoff, championship	congratulations, game, football, notre, dame, players, coach

Topic	Label	Highest Prob	FREX	Lift	Score
19	**events**	fox, green, people, bay, year, event, wisconsin	tickets, annual, fox, event, packers, museum, music	champion, ronaldo, donation, tickets, organizers, annual, parade	champion, fox, bay, green, packers, wisconsin, appleton
20	**media**	new, story, women, media, york, news, show	york, media, book, wrote, twitter, women, read	dice, diana, tweeted, magazine, twitter, tweet, york	dice, media, women, york, book, gtgtgt, twitter
21	**weather 2**	snow, weather, tomorrow, will, morning, going, now	snow, wind, rain, degrees, temperatures, weather, winds	chills, cloudy, snow, wind, showers, forecast, meteorologist	chills, snow, temperatures, rain, weather, degrees, storm
22	**morning**	well, good, right, morning, yeah, can, just	yeah, fun, pauline, cool, good, morning, yes	chili, pauline, zoo, garden, emily, awesome, deem	chili, pauline, fun, yeah, morning, cool, awesome
23	**school**	school, students, high, kids, schools, college, program	students, school, schools, student, college, campus, teachers	students, classroom, materials, teachers, elementary, teacher, superintendent	school, students, materials, schools, student, kids, teachers
24	**police 2**	police, officer, officers, gun, brown, shot, michael	officer, gun, ferguson, officers, enforcement, wilson, michael	convenience, ferguson, officer, guns, cop, missouri, louis	police, convenience, officer, officers, ferguson, jury, gun
25	**veterans**	world, country, will, today, american, years, people	veterans, america, nation, honor, world, american, church	americas, veterans, nation, veteran, vietnam, honor, sergeant	americas, veterans, war, world, american, america, afghanistan
26	**reporter**	gtgt, reporter, said, say, dont, people, yes	gtgt, reporter, gtgtgt, cnn, yes, listen, didnt	reporter, gtgt, cnns, gtgtgt, cnn, brooke, anderson	reporter, gtgt, gtgtgt, cnn, cnns, happened, said
27	**president**	president, house, republicans, obama, republican, will, democrats	republicans, republican, democrats, clinton, senate, hillary, election	agenda, republicans, democrats, republican, democrat, hillary, immigration	agenda, republicans, president, democrats, republican, hillary, clinton

Topic	Label	Highest Prob	FREX	Lift	Score
28	**fire**	fire, car, road, morning, just, live, now	fire, cars, driver, road, car, truck, drivers	flames, firefighters, driver, intersection, lanes, brianne, firefighter	flames, fire, car, firefighters, driver, cars, crash
29	**plane**	plane, flight, air, search, airport, information, area	plane, flight, aircraft, ocean, airport, pilot, ship	malaysian, aviation, islands, malaysia, plane, flight, wreckage	plane, islands, flight, aircraft, malaysian, search, airlines
30	**states**	will, going, now, right, line, see, get	line, california, zone, virginia, train, space, station	zone, trains, california, ring, mexico, train, line	zone, virginia, line, space, going, california, station

Table 8: Topics for 2015 STM. See appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**planned parenthood**	phone, planned, officer, body, parenthood, cell, fired	planned, phone, parenthood, cell, videos, fired, camera	submit, planned, parenthood, footage, images, phone, videos	parenthood, officer, submit, planned, phone, cell, videos
2	**video**	video, car, shot, stop, saw, man, pulled	video, van, pulled, car, shot, bus, leg	video, van, recording, yelling, screaming, belt, leg	video, car, van, shot, pulled, -year-old, neck
3	**police**	police, officers, black, gun, city, officer, community	gun, baltimore, black, officers, gray, guns, violence	ferguson, policing, recover, freddie, protests, baltimore, gray	police, officers, recover, officer, baltimore, gun, black
4	**congress**	house, white, congress, bill, will, republicans, senate	congress, senate, speaker, house, vote, bill, legislation	speaker, con, boehner, bipartisan, lawmakers, veto, shutdown	senate, con, congress, republicans, democrats, republican, vote
5	**cancer**	women, children, can, health, kids, child, cancer	cancer, disease, doctor, brain, study, doctors, health	cancer, disease, grace, symptoms, medication, diagnosed, brain	grace, cancer, health, disease, patients, children, child

Topic	Label	Highest Prob	FREX	Lift	Score
6	**animals**	year, home, people, come, day, dog, event	dog, dogs, christmas, owner, animal, store, animals	furniture, pets, pet, donate, dogs, animals, animal	furniture, dog, dogs, store, animal, dollars, year
7	**obama**	president, obama, policy, foreign, speech, administration, world	foreign, policy, obama, president, obamas, presidents, speech	web, foreign, obamas, policy, obama, president, oval	president, obama, web, foreign, policy, barack, speech
8	**station**	now, live, news, city, will, new, abc	metro, maryland, rhode, island, sam, providence, abc	sale, bowser, rhode, trains, providence, sweeney, buses	sale, metro, providence, city, rhode, county, abc
9	**fun**	fun, anybody, join, dance, joy, laugh, cheering	fun, join, joy, anybody, dance, laugh, cheering	joy, join, fun, cheering, laugh, dance, guide	fun, joy, join, anybody, dance, laugh, cheering
10	**trump**	trump, donald, hes, republican, carson, bush, new	trump, donald, carson, polls, iowa, stage, poll	stage, carsons, trump, donald, trumps, romney, carson	stage, trump, donald, carson, jeb, republican, bush
11	**-**	gtgt, reporter, said, gtgtgt, get, dont, like	gtgt, reporter, gtgtgt, cnn, yes, ten, jake	reporter, gtgt, gtgtgt, translator, cnns, jake, don	reporter, gtgt, gtgtgt, cnn, e-mail, translator, jake
12	**weather**	will, water, morning, snow, tomorrow, see, day	snow, weather, water, rain, storm, temperatures, degrees	futurecast, rain, inches, showers, snow, thunderstorms, clouds	snow, temperatures, rain, weather, showers, water, degrees
13	**technology**	new, government, information, federal, use, company, can	technology, data, company, employees, cyber, systems, companies	lowest, consumers, technology, cyber, users, systems, data	lowest, government, data, cyber, federal, technology, company
14	**football**	game, team, one, play, year, season, football	game, football, sports, games, notre, dame, team	cubs, nfl, playoff, tournament, coaches, irish, playoffs	cubs, notre, game, dame, sports, football, irish
15	**money**	money, million, tax, dollars, pay, cut, jobs	tax, cut, money, taxes, million, jobs, pay	cut, tax, taxes, growth, taxpayers, revenue, income	cut, tax, money, taxes, dollars, economy, budget
16	**-**	people, can, will, think, want, need, make	important, things, process, need, sure, talk, forward	inner, newstalk, challenges, bruce, decisions, discussions, conversations	inner, people, think, need, important, will, can

Topic	Label	Highest Prob	FREX	Lift	Score
17	**cooking**	can, just, okay, right, like, yeah, little	okay, yeah, gonna, cheese, nice, cool, eat	cheese, oven, recipes, salad, butter, chocolate, cooking	recipes, yeah, okay, gonna, cheese, chocolate, great
18	**-**	know, going, right, well, thats, get, just	know, theyre, mean, going, really, lot, yeah	wave, know, sort, theyre, mean, whats, probably	know, wave, mean, going, right, theyre, think
19	**family**	like, just, years, family, life, one, love	movie, friends, amazing, book, loved, love, dad	justin, song, diana, movie, instagram, movies, sing	justin, movie, film, love, book, family, music
20	**school**	school, wsbt, students, south, says, county, bend	students, wsbt, bend, school, schools, indiana, joseph	students, copeland, com, crenshaw, elkart, fillmore, teachers	wsbt, com, school, students, bend, elkhart, county
21	**debate**	debate, candidates, think, night, last, rubio, governor	debate, walker, carly, candidates, debates, rubio, marco	cnbc, moderators, winners, ferina, debates, karly, carly	debate, rubio, candidates, winners, carly, marco, debates
22	**presidential candidates**	clinton, hillary, shes, sanders, campaign, democratic, state	sanders, clinton, biden, hillary, bernie, clintons, server	biden, sanders, server, clintons, hearings, bernie, emails	hillary, clinton, biden, sanders, clintons, bernie, hearings
23	**immigration**	people, country, governor, law, states, immigration, going	immigration, illegal, governor, border, law, country, laws	citizenship, creation, undocumented, latino, mexico, amnesty, illegal	immigration, governor, creation, law, border, immigrants, citizenship
24	**police**	police, fire, now, say, just, one, scene	scene, fire, driver, injuries, hospital, injured, firefighters	flames, firefighters, gunshots, suv, scene, transported, driver	police, flames, scene, hospital, county, investigators, injuries
25	**isis**	isis, military, iran, will, iraq, deal, syria	nuclear, iran, troops, russia, assad, military, forces	sanctions, iranians, irans, assad, kurds, ukraine, troops	irans, isis, iran, syria, iraq, nuclear, assad
26	**plane**	officials, information, one, security, plane, now, may	plane, flight, pilot, airport, officials, search, sources	bodies, airlines, flight, pilots, plane, pilot, drone	bodies, flight, plane, investigation, passengers, aircraft, airport

Topic	Label	Highest Prob	FREX	Lift	Score
27	**court**	case, court, judge, attorney, said, today, charges	charges, judge, prison, jury, attorney, charged, trial	hernandez, courtroom, sentenced, jurors, prosecutors, jury, sentencing	hernandez, court, charges, attorney, jury, murder, prosecutors
28	**media**	think, dont, said, say, know, people, hes	media, dont, guy, mean, think, doesnt, agree	curious, journalists, media, stupid, offended, apologize, ridiculous	curious, think, media, dont, know, mean, hes
29	**terrorism**	isis, attack, people, paris, attacks, terror, terrorism	paris, refugees, terrorists, islam, terror, terrorist, muslims	massacre, jihad, paris, radicalized, refugees, islam, bernardino	isis, refugees, muslims, paris, muslim, massacre, islam
30	**church**	people, church, pope, faith, religious, today, rights	pope, marriage, faith, church, flag, gay, religious	introduced, pope, -sex, gay, marriage, cuba, bible	introduced, pope, church, religious, marriage, faith, gay

Table 9: Topics for 2016 STM. See appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**police**	police, black, officers, gun, officer, shot, community	gun, officers, officer, police, charlotte, shooting, shot	submit, shootings, cops, gun, charlotte, officers, tulsa	police, officers, officer, gun, submit, charlotte, shooting
2	**government**	government, security, will, federal, department, new, information	agencies, data, cyber, federal, agency, services, government	rent, recommendations, management, veto, agencies, lawmakers, software	rent, federal, government, cyber, security, agencies, department
3	**plane**	air, plane, train, one, space, force, new	plane, unbelievable, train, flight, flying, space, air	unbelievable, pilot, airplane, jet, plane, landing, profile	unbelievable, plane, flight, train, aircraft, air, passengers
4	**family**	just, like, family, life, one, kids, years	mother, son, father, kids, family, daughter, mom	music, elementary, mom, joy, mother, daughters, girl	music, kids, mother, family, parents, father, children
5	**-**	way, different, make, line, get, trying, put	way, different, line, sometimes, ways, gets, rules	way, bottom, line, different, impression, useful, sometimes	way, different, rules, line, sometimes, ways, gets

Topic	Label	Highest Prob	FREX	Lift	Score
6	**court**	case, video, court, judge, will, evidence, justice	video, judge, attorney, case, charges, court, charged	video, jury, guilty, sentenced, lawsuit, trial, charged	video, attorney, judge, court, justice, charges, jury
7	**congress**	president, house, party, republican, republicans, obama, democrats	senate, ryan, house, democrats, republicans, paul, party	pelosi, richard, mcconnell, senate, ryans, reid, schumer	richard, democrats, senate, republicans, republican, president, obama
8	**presidential candidates**	trump, clinton, hillary, donald, debate, shes, think	debate, shes, debates, playing, hillary, candidates, clinton	moderator, moderators, playing, holt, debates, debate, lester	playing, clinton, trump, hillary, donald, debate, debates
9	**football**	game, year, team, one, back, tonight, will	season, game, football, larry, sports, games, notre	championship, dame, limited, notre, larry, irish, sports	limited, notre, dame, football, game, players, sports
10	**-**	gtgt, reporter, cnn, gtgtgt, campaign, one, says	reporter, present, gtgtgt, cnn, gtgt, cnns, tapper	present, reporter, tapper, cnns, sara, gtgtgt, aides	gtgt, reporter, present, gtgtgt, cnn, trumps, cnns
11	**weather**	now, will, just, right, county, south, city	county, snow, storm, wsbt, bend, weather, road	hurricane, meteorologist, slow, snow, danielle, suzanne, inches	slow, wsbt, county, snow, elkhart, bend, storm
12	**isis**	isis, war, syria, iraq, military, now, forces	isis, iraq, syria, troops, islamic, terrorists, syrian	baghdad, pentagon, caliphate, fighters, isil, iraqi, mosul	pentagon, isis, syria, iraq, iraqi, mosul, syrian
13	**family/ church**	women, men, woman, born, said, bill, say	women, born, birth, sexual, church, christmas, christian	delivered, jewish, birth, marriage, abortion, certificate, pope	delivered, women, christmas, israel, born, church, abortion
14	**-**	going, know, people, well, want, get, great	great, thank, going, know, want, ive, youve	appointment, fantastic, hopefully, appreciate, vets, luck, thank	appointment, going, know, people, great, thank, well
15	**foreign defense**	russia, iran, north, nuclear, defense, putin, korea	defense, putin, korea, iran, russia, nuclear, sanctions	koreas, putin, sailors, defense, irans, jong-un, korea	defense, russia, korea, iran, nuclear, putin, russian

Topic	Label	Highest Prob	FREX	Lift	Score
16	**trump**	trump, donald, hes, trumps, campaign, president-elect, president	president-elect, pence, mike, romney, transition, mitt, trump	fortune, gingrich, bannon, swamp, pence, mattis, newt	trump, donald, president-elect, romney, fortune, trumps, pence
17	**-**	gtgt, said, dont, people, think, say, want	gtgt, dont, jake, said, didnt, yes, listen	buddy, jake, gtgt, anderson, wolf, inappropriate, corey	gtgt, buddy, jake, think, gtgtgt, people, dont
18	**america**	will, country, people, america, american, americans, make	applause, america, cheers, education, class, nation, inner	allen, applause, cheers, inequality, poverty, education, wage	applause, allen, cheers, america, jobs, hillary, country
19	**-**	know, dont, think, like, right, thats, just	mean, yeah, dont, know, okay, youre, like	blacks, tucker, cuz, neil, weird, yeah, kimberly	blacks, think, mean, yeah, know, dont, hes
20	**us/race**	people, country, speech, american, say, want, america	racist, flag, amendment, constitution, hate, muslim, freedom	web, anthem, racist, bigot, flag, liberty, amendment	web, racist, flag, speech, muslim, constitution, amendment
21	**media**	media, news, new, york, press, fox, now	media, press, university, campus, coverage, fox, students	balanced, journalism, journalists, howard, campus, media, buzz	media, balanced, students, campus, fox, press, news
22	**-**	think, really, lot, well, theres, thats, see	sort, really, think, kind, terms, certainly, obviously	luxury, sort, brexit, perspective, terms, reflection, broader	think, luxury, sort, really, lot, things, kind
23	**delegates**	trump, sanders, cruz, donald, bernie, ted, republican	cruz, ted, sanders, bernie, delegates, rubio, convention	con, cruz, ted, delegates, kasich, caucuses, rubio	cruz, sanders, delegates, con, bernie, ted, trump
24	**terrorism**	now, attack, new, two, people, one, police	authorities, brussels, injured, bomb, paris, suspect, attack	com, rahami, belgian, attacker, bomber, suspects, plot	com, police, brussels, fbi, investigators, terror, injured

Topic	Label	Highest Prob	FREX	Lift	Score
25	**immigration**	immigration, country, wall, illegal, going, border, will	immigration, illegal, border, immigrants, sanctuary, mexico, mexican	undocumented, deport, deportation, factory, sanctuary, aliens, immigration	immigration, factory, immigrants, illegal, border, sanctuary, mexico
26	**election**	trump, vote, clinton, election, voters, states, hillary	voting, polls, electoral, votes, pennsylvania, ohio, poll	stops, electoral, stein, jill, ballots, battleground, recount	stops, trump, clinton, polls, vote, voters, electoral
27	**health**	health, can, water, care, medical, now, get	doctor, medical, health, pneumonia, doctors, insurance, cancer	improvement, patients, symptoms, doctor, disease, diagnosed, doctors	health, improvement, pneumonia, patients, medical, doctor, doctors
28	**clinton emails**	clinton, hillary, fbi, foundation, emails, information, state	emails, foundation, classified, server, fbi, email, e-mails	deliberately, server, classified, podesta, emails, wikileaks, comey	clinton, fbi, emails, classified, server, e-mails, hillary
29	**money**	money, tax, million, jobs, business, going, pay	tax, money, trade, taxes, dollars, companies, market	gate, tax, prices, trillion, audit, market, rates	tax, gate, money, taxes, jobs, dollars, trade
30	**cuba**	president, united, states, obama, world, will, american	united, cuba, prime, states, president, minister, castro	cuba, permanent, castro, prime, british, britain, communist	permanent, president, united, cuba, obama, castro, states

Table 10: Topics for the topical content STM for Sinclair purchased station data.

Topic	Words
1	cameras, photos, captured, photo, wjla, footage, images
2	climb, time
3	dancing, parade, dance, candy, cheer, band, golf
4	videos, video, posted, media, facebook, twitter, shows
5	honored, behalf, introduce, celebrate, colleagues, supportive, achieve
6	commissioners, proposal, properties, council, citys, apartments, mayor
7	emergency, drug, staffing, deaths, drugs, procedures, prevention
8	bristol, friendship, flag, ford, williams, courage, flags
9	touchdowns, touchdown, undefeated, quarterback, halftime, scored, coach
10	spicy, flour, yummy, chocolate, garlic, recipes, flavors
11	violence, americans, weapons, global, domestic, religious, border
12	tea, super, ideas, magic, style, interesting, traditional
13	flames, firefighters, firefighter, crash, smoke, crashed, blaze
14	flooding, snowfall, flooded, sidewalks, snow, pipe, water
15	joy, excited, winners, join, sleeping, wave, scary
16	artists, donations, museum, concert, donate, festival, organizers
17	animals, animal, humane, dogs, dog, adoption, pet
18	prosecutors, sentenced, prosecutor, courtroom, pleaded, convicted, lawsuit
19	bridge, drivers, lanes, bridges, engineers, gas, crashes
20	thunderstorms, showers, inland, gusts, clouds, sunshine, winds
21	redskins, players, baseball, nfl, playoff, football, player
22	greg, farmers, tropical, sunny, corn, fishing, clay
23	teachers, teacher, students, superintendent, academic, elementary, colleges
24	book, guy, dad, god, married, cuz, mom
25	gunshot, gunman, police, suspects, suspect, custody, homicide
26	workforce, governments, companies, contractors, infrastructure, consumers, taxpayers
27	lawmakers, republicans, senate, congress, democrats, bipartisan, legislature
28	winnebago, sturgeon, marinette, dnr, deer, lakes, oshkosh
29	roz, plater, year-old, rockville, cheetah, -year-old, grandmother
30	sort, obviously, folks, interesting, mentioned, necessarily, maybe

Table 11: Topic-Covariate Interactions for the topical content STM for Sinclair purchased station data.

Topic	Group	Words
1	Before	girl, ray, hurt, hadnt, pictures, capture, submit
	After	cell, body, surveillance, recorded, light, footage, speed
2	Before	climb
	After	-
3	Before	katie, carrie, congratulations, entertainment, audience, anniversary, stage

Topic	Group	Words
	After	shoppers, gifts, gift, toys, sale, shopping, eve
4	Before	post, page, youtube, facebook, twitter, recognize, shows
	After	social, moment, pictures, released, heres, online, moments
5	Before	brad, wedding, bell, lisa, birthday, hair, awards
	After	rhode, island, providence, lord, assembly, pray, governor
6	Before	annie, copeland, roth, rick, bend, wsbts, spencer
	After	taxpayers, revenue, funding, taxpayer, taxes, bern, funds
7	Before	security, cyber, consumer, digital, cell, sheriffs, sheriff
	After	symptoms, hospitals, virus, patients, disease, cancer, goshen
8	Before	veterans, afghanistan, cemetery, military, war, soldiers, warwick
	After	virginia, heather, marion, bank, loan, johnny, richmond
9	Before	irish, tournament, penn, hockey, carl, benton, kimberly
	After	greenville, friendship, vegetable, henry, kingsport, science, suv
10	Before	festival, easter, holiday, book, babies, amanda, reduce
	After	pauline, classroom, flowers, meteorologist, birds, awesome, fitness
11	Before	cancer, patients, disease, therapy, symptoms, diagnosed, shutdown
	After	afghanistan, russia, terrorist, iraq, terror, troops, soldiers
12	Before	emily, pauline, deem, angela, colors, rachel, cute
	After	gop, donald, trump, cruz, hillary, democratic, republican
13	Before	tornado, suv, homeowner, korff, lightning, suspicious, insurance
	After	jeanette, reyes, trains, rail, metro, flights, riders
14	Before	cherry, heat, ski, storms, supply, residential, restrictions
	After	boats, beaches, roadway, bern, coastal, intersection, atlantic
15	Before	dinner, excited, winners, scary, sleeping, joy, wave
	After	joining, wave, joy, excited, winners, sleeping, join
16	Before	shopping, shoppers, christmas, sales, gift, gifts, volunteer
	After	ceremony, tribute, parade, exhibit, veterans, breast, memorial
17	Before	egg, eggs, breakfast, forecast, population, deer, sleeping
	After	racing, farm, lisa, race, girl, races, phil
18	Before	arrests, suspected, recorded, ford, steven, dcs, cranston
	After	murder, homicide, fairfax, murdered, custody, baltimore, death
19	Before	trains, rail, riders, passengers, metro, intersection, airport
	After	consumer, consumers, fees, cents, revenue, dollars, costs
20	Before	snow, slippery, marinette, ecu, oshkosh, appleton, pete
	After	macon, debris, warner, interstate, wgxa, trees, houston
21	Before	marathon, runners, race, receiver, races, racing, kickoff
	After	hockey, tournament, jordan, championship, basketball, coaching, teammates
22	Before	garden, flowers, boat, soil, farm, boats, lawn
	After	amanda, elkhart, deputies, sheriffs, wsbt, bend, wgxa
23	Before	toys, technology, santa, talented, pilot, projects, connecticut

Topic	Group	Words
	After	athlete, athletes, craven, buses, trauma, pandemic, ecu
24	Before	theater, movie, inspired, baby, mothers, wsb, library
	After	guys, robert, egg, sport, skill, congratulations, retirement
25	Before	surveillance, rousey, metro, accused, morgan, beaten, cell
	After	detective, protesters, autopsy, murder, protest, trauma, departments
26	Before	taxes, prices, tax, bills, sales, fees, cents
	After	cyber, intelligence, pentagon, privacy, homeland, matters, providers
27	Before	mayor, mayors, vincent, candidates, sector, clinton, brianne
	After	immigration, tax, shutdown, proposal, marijuana, funding, proposed
28	Before	doran, beth, schlicht, follow-, alex, ship, chad
	After	montana, kalispell, bozeman, montanas, missoula, butte, snow
29	Before	homicide, detective, body, cruz, eve, lord, reyes
	After	rousey, brianne, patrice, graham, warwick, cemetery, girl
30	Before	realize, waited, worried, biggest, opened, days, deals
	After	bruce, yeah, conversations, convention, buildings, appreciate, helpful

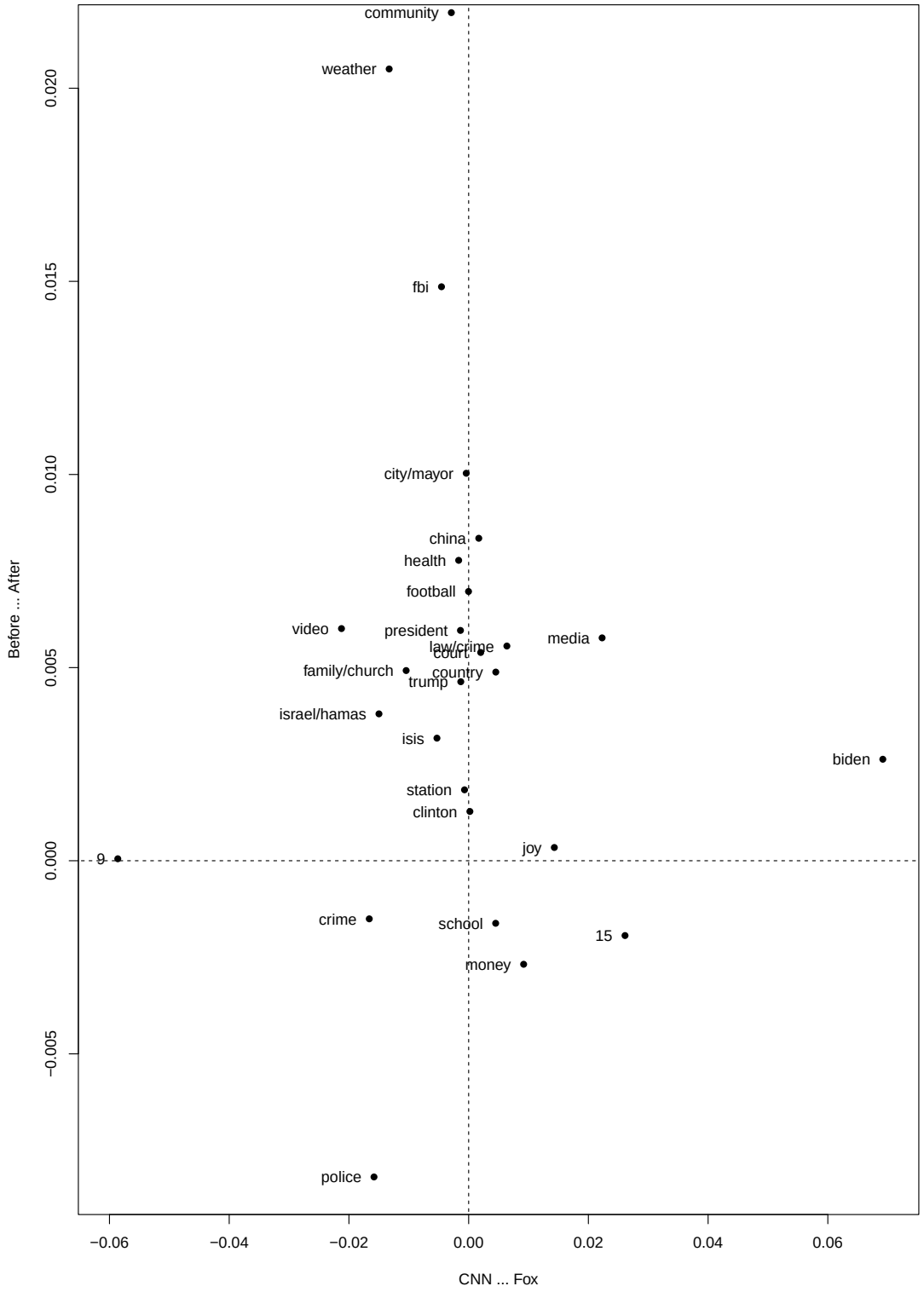


Figure 5: Results for STM on **all data**. Change in topic proportion shifting from CNN to Fox on the x-axis and before Sinclair purchase to after Sinclair purchase on the y-axis. Topics 10 (cooking), 26 (family), and 24 (station 2) are omitted, but can be found in Figure 2a.

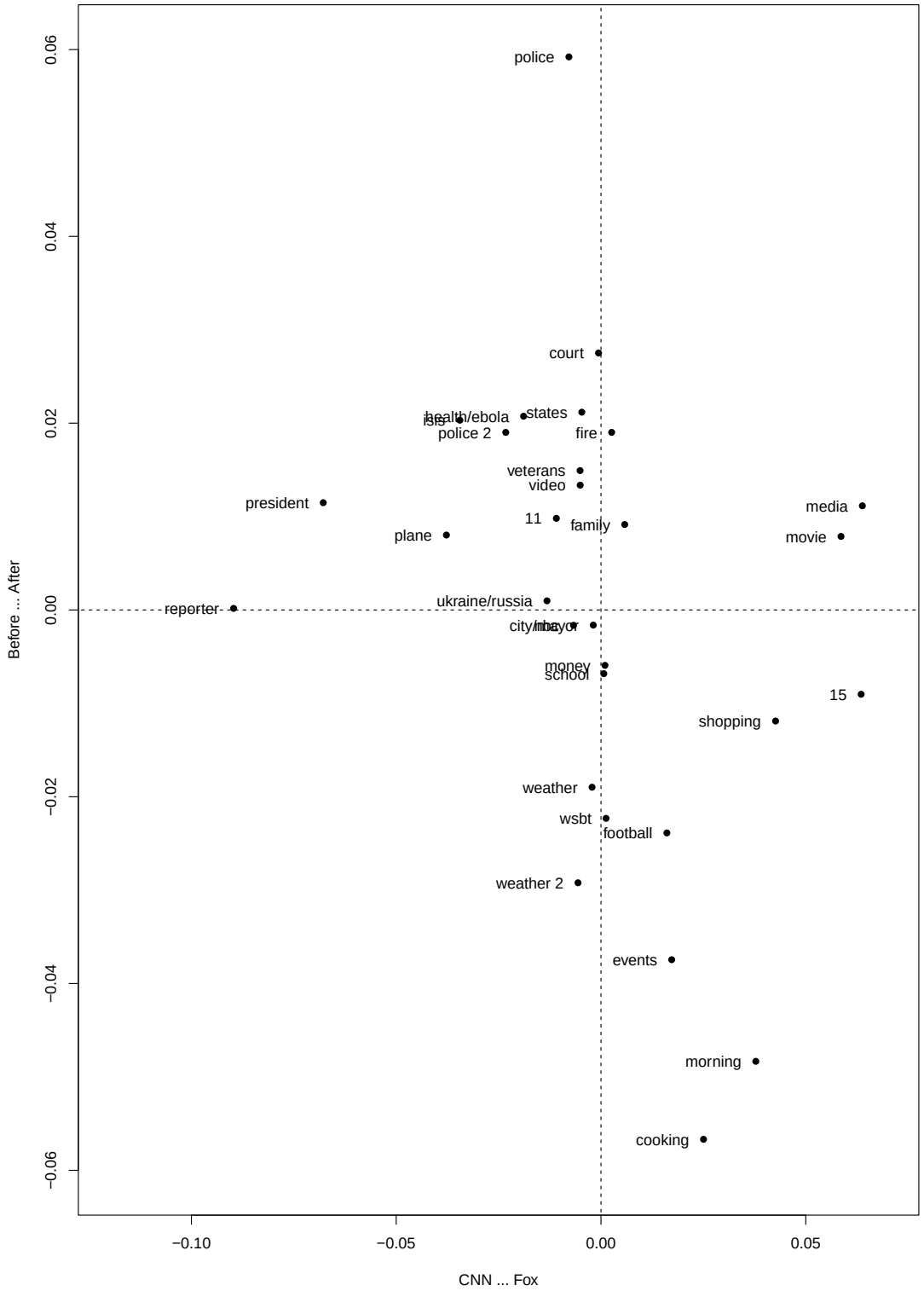


Figure 6: Results for STM on **2014 data**. Change in topic proportion shifting from CNN to Fox on the x-axis and before Sinclair purchase to after Sinclair purchase on the y-axis.

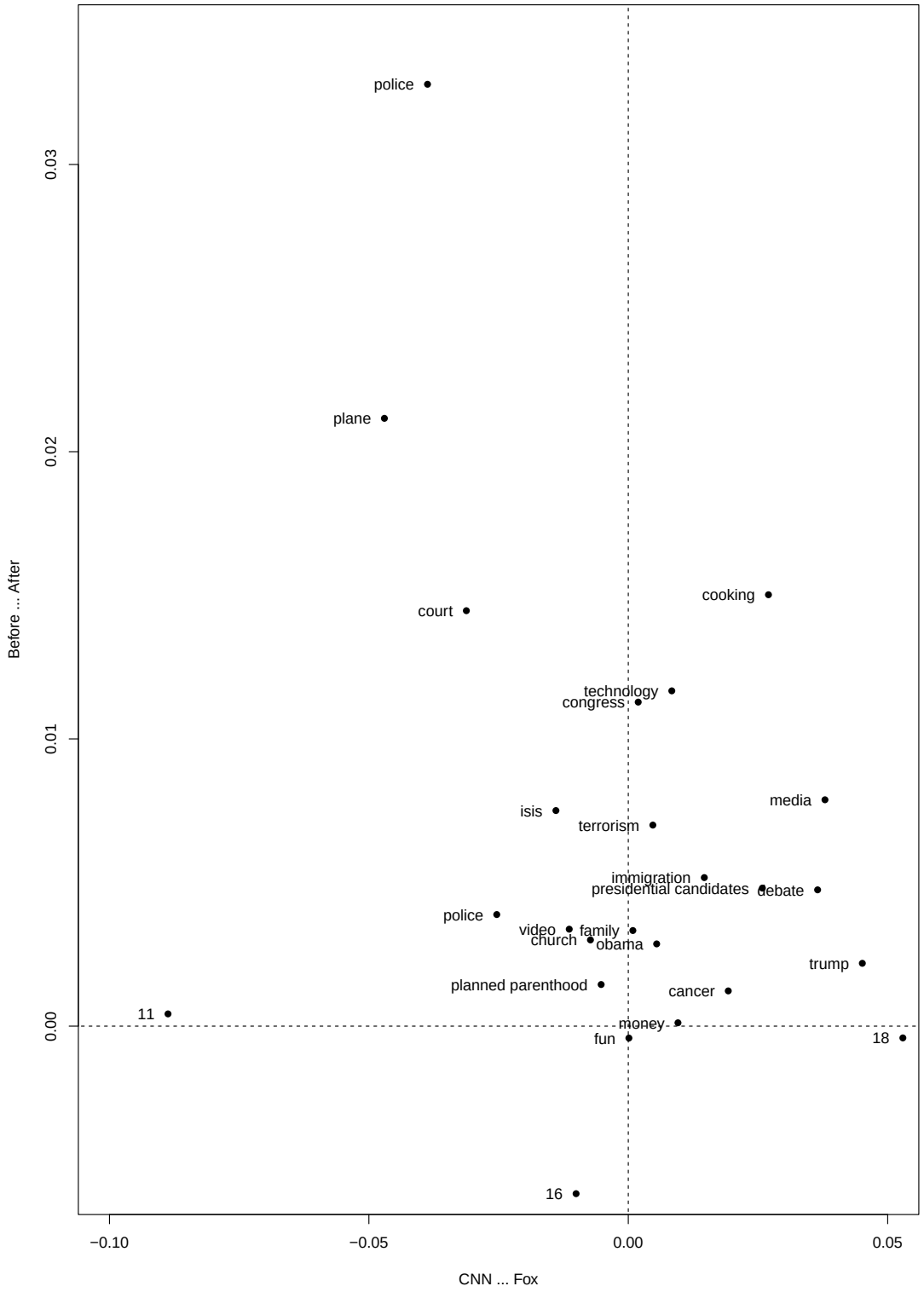


Figure 7: Results for STM on **2015 data**. Change in topic proportion shifting from CNN to Fox on the x-axis and before Sinclair purchase to after Sinclair purchase on the y-axis. Topics 6 (animals), 8 (station), 12 (weather), 14 (football), and 20 (school) are omitted, but can be found in Figure 2c.

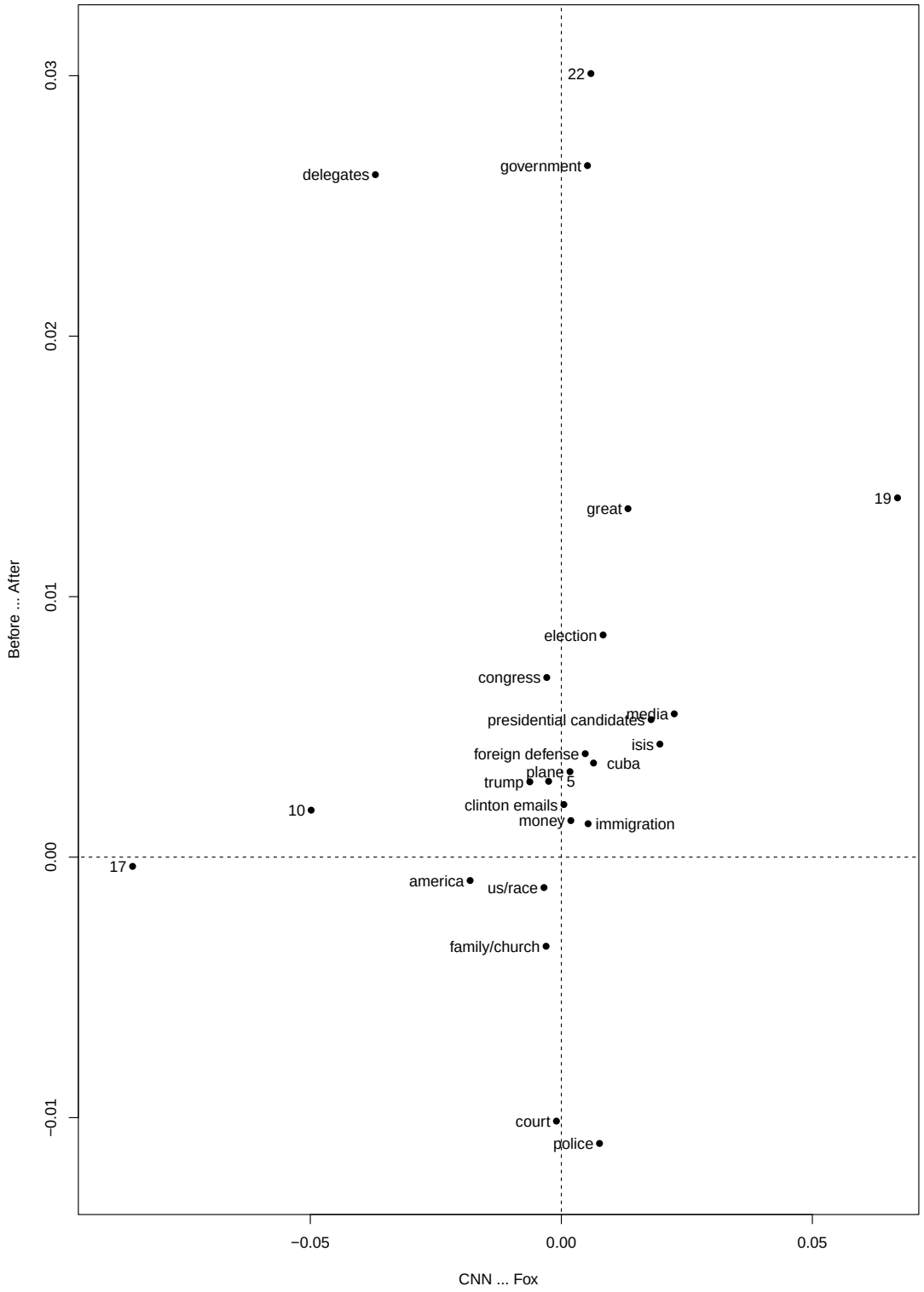


Figure 8: Results for STM on **2016 data**. Change in topic proportion shifting from CNN to Fox on the x-axis and before Sinclair purchase to after Sinclair purchase on the y-axis. Topics 4 (family), 9 (football), 11 (weather), 24 (terrorism), and 27 (health) are omitted, but can be found in Figure 2d.

Figure 9: STM with before/after purchase as a topical content covariate. These plots show words within a topic which are strongly associated with coverage before Sinclair takeover as opposed to after Sinclair purchase.

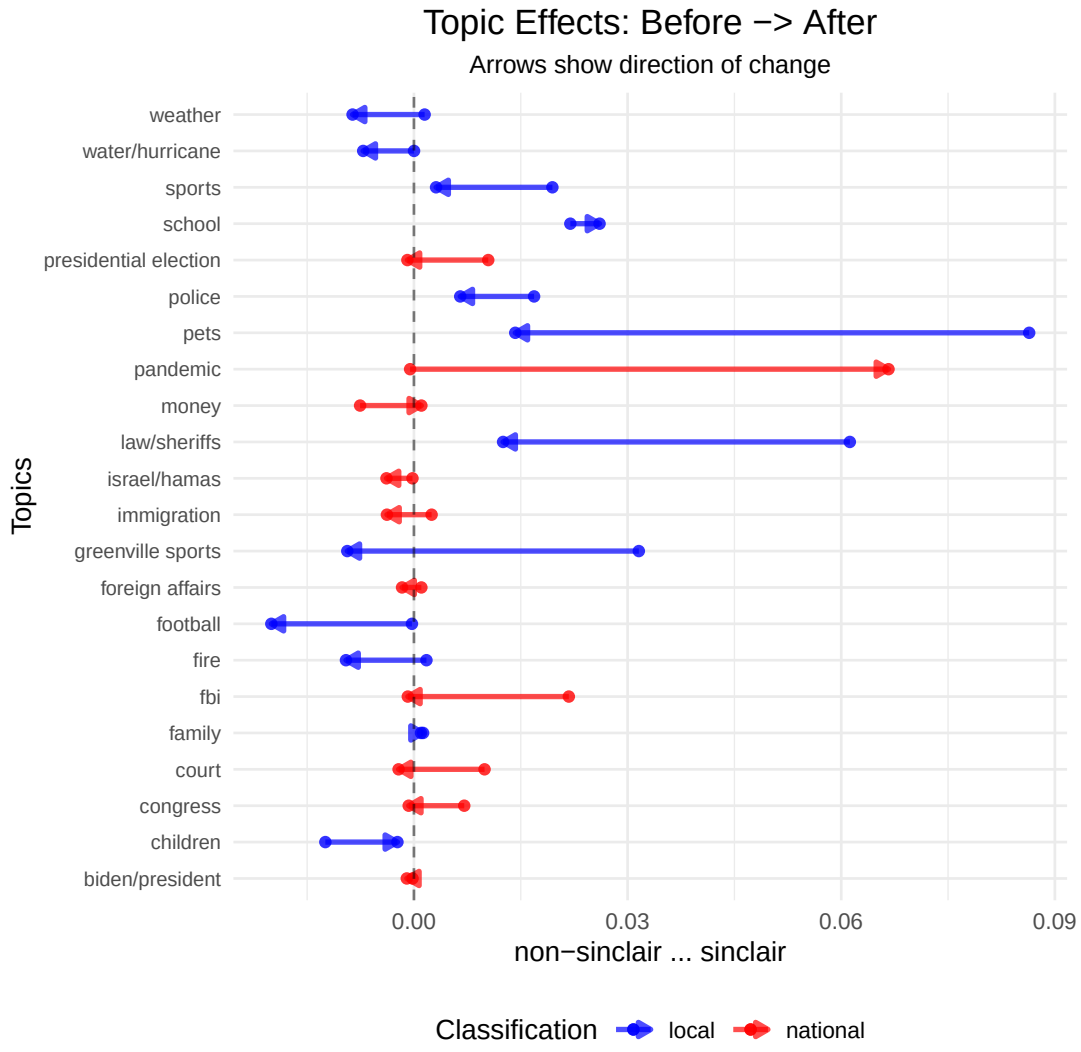


Figure 10: Results for STM on **all paired data**. Change in topic proportion shifting from non-Sinclair to Sinclair affiliate on the x-axis, and shift before and after purchase date is shown with arrows. Red denotes national topics and blue denotes local topics. Topic list is shown on the y-axis. Topics with unclear national/local interpretation are omitted here, and included in Figure 11.

Table 12: Topics for Paired Analysis STM with all data. See appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**immigration**	border, people, country, new, law, governor, city	border, illegal, texas, immigration, migrants, immigrants, cities	asylum, climb, cartels, migrant, migrants, sanctuary, aliens	border, climb, immigration, migrants, immigrants, illegal, sanctuary
2	**police**	police, car, morning, say, officers, happened, just	scene, police, officers, crash, officer, shooting, vehicle	submit, crash, shooter, scene, injuries, accident, fatal	submit, police, officers, scene, crash, car, injuries

Topic	Label	Highest Prob	FREX	Lift	Score
3	**court**	case, court, trump, judge, will, president, former	supreme, judge, trial, indictment, court, legal, lawyers	testifying, fani, indictment, willis, wade, lawyers, supreme	trump, testifying, court, judge, trial, supreme, donald
4	**media**	media, news, new, video, fox, york, show	video, media, twitter, fox, tucker, post, movie	video, instagram, twitter, outlets, tucker, platforms, magazine	video, media, fox, twitter, york, social, facebook
5	**football**	first, going, get, got, game, right, hes	ball, yards, coach, game, gonna, rivals, play	videos, rivals, clock, bennett, pals, snap, powered	videos, touchdown, yards, game, coach, ball, quarterback
6	**israel/ hamas**	ukraine, israel, war, now, military, will, hamas	hamas, gaza, israel, israeli, forces, hostages, ukrainian	counteroffensive, hamas, hezbollah, hostages, idf, palestinian, casualties	hamas, ukraine, gaza, israel, casualties, russia, putin
7	**foreign affairs**	president, united, will, states, administration, secretary, foreign	foreign, china, secretary, nuclear, countries, administration, president-elect	foreign, sanctions, korea, nuclear, cuba, taiwan, ambassador	foreign, putin, president, russia, china, nuclear, president-elect
8	**law/ sheriffs**	county, found, office, law, charges, case, sheriffs	sheriffs, murder, sheriff, arrested, charges, charged, prison	colonel, sheriffs, deputies, homicide, sheriff, arrested, murder	colonel, sheriffs, county, charges, murder, police, investigators
9	**TN**	city, county, will, says, johnson, tennessee, news	channel, johnson, tennessee, josh, city, kingsport, sarah	champion, newschannel, commissioners, eleven, tennessees, channel, improvements	champion, county, tennessee, kingsport, city, bristol, johnson
10	**money**	money, dollars, million, tax, pay, jobs, will	tax, jobs, taxes, dollars, companies, money, billion	unbelievable, tax, wages, income, paycheck, investments, medicare	unbelievable, tax, dollars, inflation, taxes, economy, money
11	**gender/ race**	people, country, women, america, american, will, black	rights, america, black, freedom, hate, women, racist	pledge, religious, racism, protest, freedoms, gender, religion	pledge, america, women, americans, thank, democracy, american

Topic	Label	Highest Prob	FREX	Lift	Score
12	**water/ hurri-cane**	water, now, area, just, can, right, will	water, bridge, plane, storm, flight, airport, coast	yep, ocean, flight, hurri-cane, bridge, debris, rail	yep, water, storm, hur-ricane, river, airport, aircraft
13	**advertis-ements**	new, christmas, store, now, get, car, bristol	christmas, store, holiday, sale, restaurant, shop, sales	beats, christ-mas, santa, sale, holiday, stores, restaurant	beats, bristol, christmas, kingsport, furniture, abington, cus-tomers
14	**congress**	house, republi-cans, bill, re-publican, sen-ate, democrats, party	senate, speaker, republicans, mccarthy, congressman, capitol, republi-can	maga, speaker, mccarthy, sen-ate, schumer, senators, mc-connell	maga, re-publicans, democrats, republican, senate, speaker, congress
15	**filler words 1**	way, make, can, get, see, back, sure	way, make, done, long, sure, ways, making	way, shannon, ways, connect-ing, shape, fig-ure, apart	way, make, can, see, get, done, ways
16	**greenville sports**	one, tonight, game, green, win, greenville, first	greenville, green, score, devils, daniel, win, blue	chevrolet, dev-ils, warriors, greenville, boone, daniel, finds	chevrolet, greenville, touchdown, devils, game, boone, score
17	**filler words 2**	know, think, go-ing, dont, thats, people, like	think, know, mean, dont, youre, thing, kind	flash, mean, sort, know, think, honestly, dont	think, know, flash, mean, going, people, dont
18	**pandemi**	health, will, cases, state, can, people, county	testing, virus, health, cases, distancing, masks, tests	trusted, virus, vaccinated, dis-tancing, quar-antine, testing, outbreak	trusted, health, missoula, virus, kovat, testing, county
19	**weather**	morning, snow, see, will, weather, going, day	snow, tempera-tures, showers, weather, rain, forecast, de-grees	snow, toss, showers, sun-shine, cooler, cloudy, temper-atures	toss, snow, temperatures, montana, show-ers, missoula, kalispell
20	**fbi**	information, in-vestigation, fbi, department, re-port, evidence, questions	fbi, letter, classified, chairman, investigation, committee, evidence	heal, fbi, over-sight, server, classified, docu-ment, letter	fbi, heal, in-vestigation, classified, documents, evidence, com-mittee

Topic	Label	Highest Prob	FREX	Lift	Score
21	**filler words 3**	gtgt, reporter, said, say, dont, women, gtgtgt	gtgt, reporter, gtgtgt, e-mails, usa, cnn, e-mail	gtgt, gtgtgt, usa, reporter, e-mail, e-mails, aides	gtgt, usa, reporter, gtgtgt, e-mails, e-mail, cnn
22	**children**	can, children, kids, care, help, parents, child	cancer, children, mental, parents, doctor, child, doctors	tent, cancer, parent, doctors, diagnosed, doctor, pregnant	tent, children, kids, parents, patients, child, hospital
23	**family**	just, family, years, like, life, time, know	book, friends, family, life, mom, loved, father	jay, lord, funeral, grandfather, queen, larry, mom	jay, family, book, father, life, mom, thank
24	**fire**	fire, smoke, firefighters, burning, fires, tha, burn	tha, thi, tth, firefighters, ths, whe, fire	ant, ere, tha, tht, tit, ahe, aim	fire, aim, tth, tha, thi, firefighters, ath
25	**presidential election**	trump, donald, clinton, campaign, hillary, election, hes	hillary, clinton, donald, debate, trump, campaign, voters	gentlemen, hampshire, romney, rnc, hillary, mitt, battleground	trump, donald, clinton, hillary, gentlemen, election, voters
26	**pets**	right, just, little, got, can, like, yeah	dog, little, yeah, beautiful, fun, love, okay	tail, adoption, dogs, dog, chicken, delicious, sugar	tail, gonna, dog, adoption, yeah, thank, fun
27	**MT**	montana, says, missoula, people, help, community, will	mtn, montanas, park, helena, medicine, missoula, montana	medicine, nbcmontanacom, wildlife, mtns, helena, mtn, recreation	medicine, montana, missoula, mtn, flathead, bozeman, montanas
28	**school**	school, students, community, year, just, schools, kids	students, music, school, campus, schools, student, university	music, arts, festival, campus, teachers, students, classes	music, students, school, campus, schools, kids, community
29	**biden/ president**	biden, joe, president, dont, will, hunter, people	biden, joe, hunter, bidens, greg, brian, jesse	anti, jeanine, kamala, ainsley, bidens, carley, newsom	biden, bidens, joe, anti, hunter, president, democrats
30	**sports**	game, team, play, just, playing, know, got	playing, sports, football, team, play, game, field	playing, nfl, soccer, baseball, basketball, sport, league	playing, game, football, montana, coach, sports, players

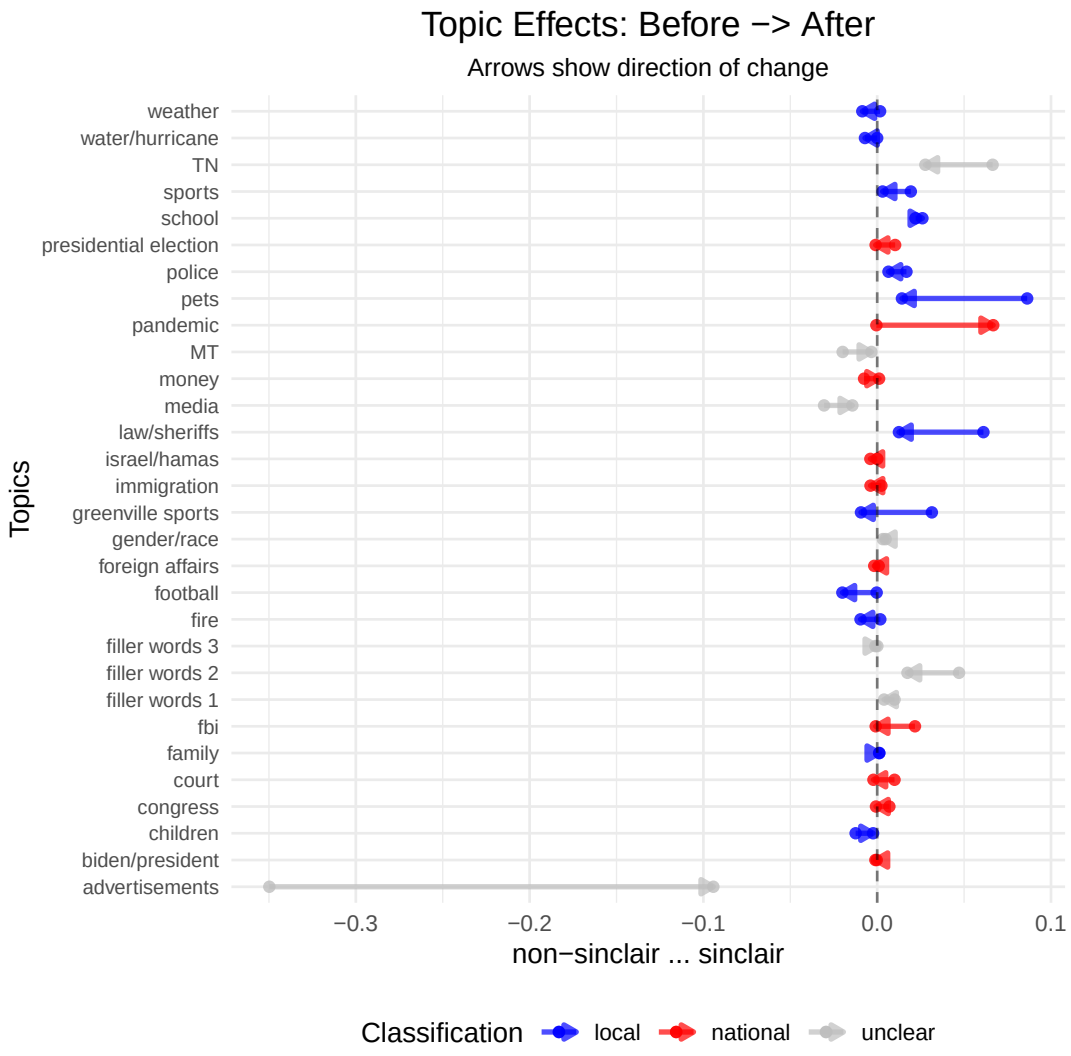


Figure 11: Results for STM on **all paired data**. Change in topic proportion shifting from non-Sinclair to Sinclair affiliate on the x-axis, and shift before and after purchase date is shown with arrows. Red denotes national topics, blue denotes local topics, and gray topics are unclear. Topic list is shown on the y-axis. The graph with only national/local topics can be found in Figure 10.

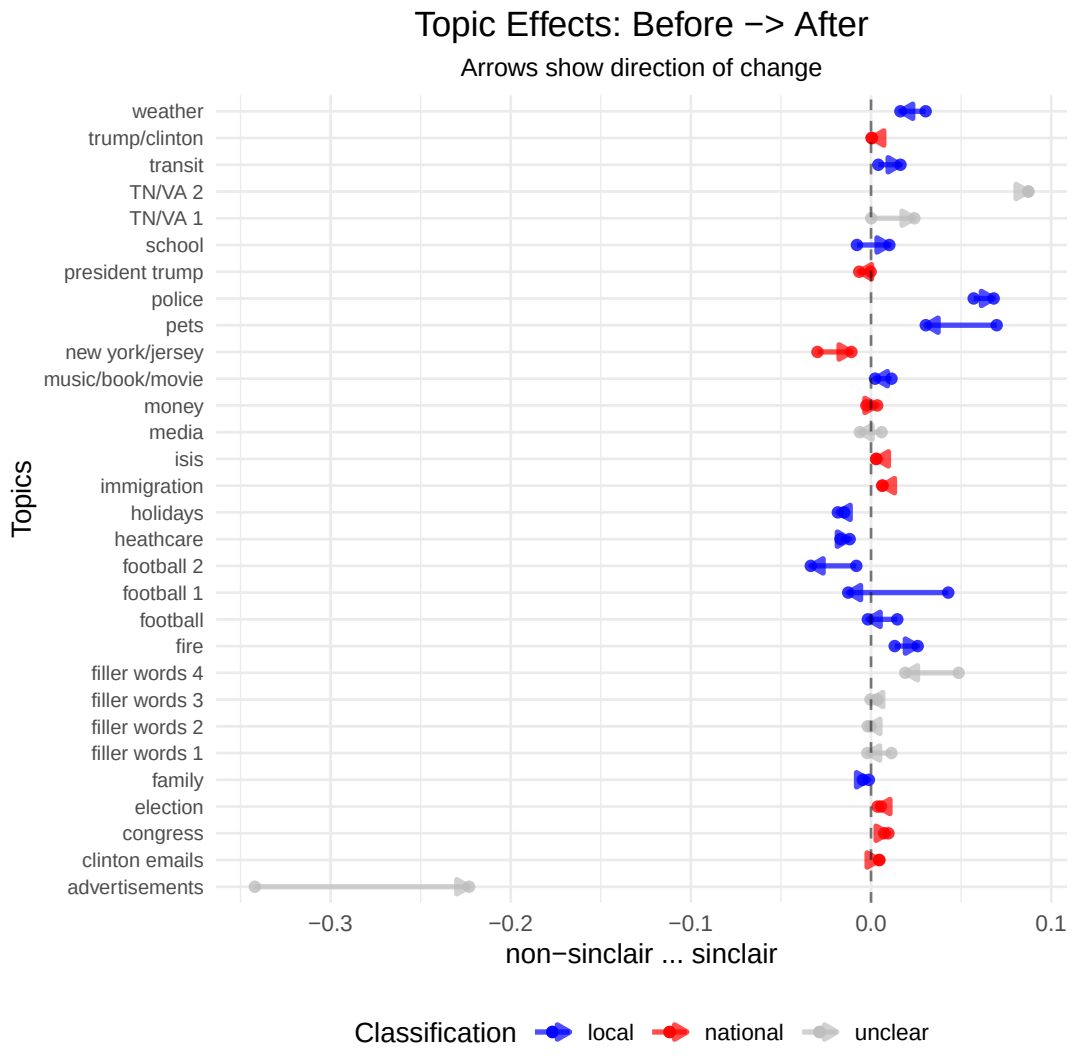


Figure 12: Results for STM on **paired data before 2020**. Change in topic proportion shifting from non-Sinclair to Sinclair affiliate on the x-axis, and shift before and after purchase date is shown with arrows. Red denotes national topics, blue denotes local topics, and gray topics are unclear. Topic list is shown on the y-axis. The graph with only national/local topics can be found in Figure 4.

Table 13: Topics for Paired Analysis STM with data before 2020. See appendix for column information.

Topic	Label	Highest Prob	FREX	Lift	Score
1	**new york/ jersey**	new, york, national, times, record, jersey, join	york, new, jersey, national, record, join, miller	jooy, york, jersey, new, miller, join, record	new, york, jooy, national, jersey, join, record
2	**media**	women, media, news, said, press, fox, saying	women, media, twitter, press, sexual, comments, magazine	journalists, sexually, ypcom, fzone, twitter, women, magazine	women, media, sexual, ypcom, fox, twitter, journalists
3	**football 1**	one, tonight, first, game, win, nothing, back	score, quarter, final, nothing, devils, lady, touchdown	aint, warriors, daniel, gate, scores, boone, hits	aint, touchdown, devils, score, greenville, game, crockett
4	**filler words 1**	can, way, make, get, want, thats, sure	way, make, can, sure, lets, put, done	way, ways, make, sure, can, try, glad	way, can, lets, make, want, talk, sure
5	**president trump**	trump, hes, donald, president, president-elect, secretary, trumps	president-elect, transition, cabinet, romney, mitt, trumps, secretary	upload, cabinet, mattis, flynn, giuliani, sessions, bolton	trump, president-elect, donald, romney, secretary, trumps, mitt
6	**money**	jobs, money, going, business, tax, will, million	jobs, tax, companies, billion, economy, money, obamacare	unbelievable, trillion, jobs, billion, companies, carrier, regulations	unbelievable, tax, jobs, obamacare, taxes, economy, companies
7	**filler words 2**	gtgt, reporter, say, dont, gtgtgt, said, cnn	gtgt, reporter, gtgtgt, cnn, jake, videos, wolf	videos, gtgt, gtgtgt, reporter, cnns, cnn, jake	gtgt, videos, reporter, gtgtgt, cnn, cnns, jake
8	**holidays**	great, got, right, well, just, christmas, come	christmas, fun, parade, event, excited, tickets, folks	palswebcom, merry, christmas, parade, festival, santa, celebration	christmas, fun, parade, palswebcom, festival, daytime, merry
9	**clinton emails**	clinton, fbi, information, investigation, emails, election, hillary	fbi, emails, e-mails, comey, email, classified, cyber	classified, heal, hacked, hackers, e-mails, hack, hacking	clinton, fbi, e-mails, heal, hillary, comey, investigation

Topic	Label	Highest Prob	FREX	Lift	Score
10	**immigration**	people, country, law, will, president, americans, america	immigration, rights, immigrants, flag, illegal, americans, law	immigrants, pledge, sanctuary, religion, protests, religious, constitution	pledge, immigration, immigrants, sanctuary, law, americans, federal
11	**TN/VA 1**	city, johnson, will, bristol, kingsport, street, now	johnson, city, bristol, kingsport, project, downtown, champion	champion, construction, citys, johnson, city, project, bristol	champion, johnson, city, kingsport, bristol, downtown, tri-cities
12	**filler words 3**	john, space, don, bob, hero, beer, wine	ray, hero, tth, don, tin, ship, bob	ath, aan, ahe, tha, whe, aon, ihe	ray, tth, ihe, tnd, ahe, tin, tng
13	**music/ book/ movie**	like, know, one, really, yeah, show, just	book, music, movie, film, song, show, love	music, movie, movies, book, film, songs, sing	music, movie, book, film, song, yeah, love
14	**weather**	morning, today, now, day, will, see, well	morning, tomorrow, weekend, afternoon, sunday, hours, live	webcom, temperatures, morning, forecast, tomorrow, wednesday, rain	morning, webcom, tomorrow, weather, temperatures, rain, forecast
15	**healthcare**	health, care, can, hospital, medical, help, also	health, patients, cancer, medical, disease, doctors, treatment	doctors, trusted, medication, symptoms, disease, patients, cancer	health, patients, trusted, hospital, medical, disease, doctors
16	**isis**	isis, war, president, military, will, united, now	isis, syria, iran, nuclear, military, iraq, forces	mosul, pit, sanctions, assad, iranian, iraqi, nato	isis, syria, russia, iran, pit, iraq, putin
17	**football**	team, game, playing, play, season, year, football	playing, players, games, sports, basketball, football, team	playing, nfl, league, basketball, players, baseball, athletes	playing, game, football, players, coach, games, etsu
18	**TN/VA 2**	county, says, tennessee, state, will, news, virginia	county, josh, tennessee, sarah, carter, board, sullivan	sponsored, defuse, jackie, burnie, commissioner, nate, tennessees	county, tennessee, sponsored, sullivan, unicoi, channel, defuse

Topic	Label	Highest Prob	FREX	Lift	Score
19	**fire**	fire, now, people, one, officials, just, attack	fire, scene, alert, attack, firefighters, authorities, montana	update, firefighters, fire, shooter, fires, alert, flames	update, fire, firefighters, police, officials, montana, authorities
20	**police**	police, said, case, say, officers, officer, man	officer, sheriffs, officers, murder, police, charges, charged	year-old, submit, deputies, sheriffs, murder, jury, aggravated	police, submit, sheriffs, officers, investigators, investigation, charges
21	**football 2**	gonna, first, now, back, get, game, hes	ball, gonna, science, thomas, yards, yard, touchdown	chevrolet, thomas, clock, bennett, rivals, crockett, ball	chevrolet, crockett, touchdown, yards, gonna, game, ball
22	**transit**	car, water, just, road, get, plane, train	water, miles, crash, plane, train, bus, car	slow, flight, miles, crashed, drivers, engine, driver	slow, crash, water, car, highway, driver, plane
23	**filler words 4**	know, think, going, people, dont, well, like	think, know, mean, dont, going, youre, thing	aim, mean, tucker, sort, know, think, neil	think, know, mean, going, people, aim, dont
24	**congress**	president, house, republican, party, obama, democrats, republicans	senate, senator, republicans, democrats, republican, cruz, governor	pelosi, ron, rubio, senate, jeb, christie, cruz	republican, democrats, obama, republicans, ron, senate, president
25	**pets**	little, right, just, like, look, got, yeah	dog, adoption, animals, shelter, dogs, animal, okay	animals, appalachian, adoption, adorable, chocolate, kitchen, cream	appalachian, adorable, adoption, shelter, gonna, dog, animal
26	**election**	trump, clinton, election, vote, donald, hillary, states	voting, vote, electoral, votes, voters, michigan, polls	mexican, electoral, battleground, recount, electorate, stein, rigged	trump, clinton, hillary, election, donald, voters, electoral
27	**advertisements**	now, home, store, free, one, get, buy	sale, sales, furniture, store, shop, shopping, buy	wallace, accessories, furniture, sales, app, sale, brands	wallace, furniture, sale, sales, store, kingsport, abington

Topic	Label	Highest Prob	FREX	Lift	Score
28	**family**	family, just, people, life, know, help, years	family, children, families, father, mother, life, mom	properties, funeral, mom, journey, honor, mothers, moms	properties, family, children, veterans, kids, church, mother
29	**trump/ clinton**	trump, donald, clinton, hillary, debate, think, said	debate, hillary, clinton, donald, trump, shes, candidates	absolute, moderator, debate, debates, lester, temperament, universe	trump, clinton, hillary, donald, debate, absolute, clintons
30	**school**	school, students, video, schools, university, college, kids	video, students, campus, school, schools, student, elementary	video, campus, elementary, teachers, teacher, classes, students	video, school, students, campus, student, schools, gun

Learning Moral Diversity: Modelling Individual Perspectives in Moral Classification of Texts

Yi Ren, Lewis Mitchell, Matthew Roughan
School of Mathematical Sciences, Adelaide University
{yi.ren, lewis.mitchell, matthew.roughan}@adelaide.edu.au

Abstract

Understanding moral values in social media text offers insight into moral judgement formation, and supervised NLP models trained on crowdsourced data have achieved strong classification performance. However, most approaches simplify the problem by aggregating multiple annotators' labels into a single "ground truth", overlooking the inherent subjectivity of the task. In practice, there are disagreements between annotators caused by personal viewpoint or inherent ambiguities, particularly for short tweets. Here, we extend a pretrained language model with a layer that learns annotator-specific features. Our model improves predictions of individual annotations and yields representations that reveal meaningful insights into annotators' moral perspectives. We show that models trained on aggregated labels may hide variation and give a misleading impression of performance. Overall, we demonstrate that disagreement reflects the inherent subjectivity of the task and that modelling individual perspectives creates benefits for moral classification of texts.

1 Introduction

Morality plays a vital role in shaping people's opinions and forming judgement towards social events. Accordingly, analysing morality allows for better understanding of what people believe and how people interact and form communities. We can explore these beliefs through opinions and stances expressed via language, in particular online content. Such analysis has inspired diverse research directions—from analysing political ideology and polarisation (Haidt and Graham, 2007) to understanding how people engage in public-health discourse (Jiang et al., 2025; Zhou et al., 2024). Thus, it is extremely valuable to be able to extract moral values from human-created content.

Many Natural Language Processing (NLP) techniques have been applied and integrated with an empirically validated psychological framework called **Moral Foundations Theory** (MFT) (Graham et al., 2009, 2013; Haidt, 2012). MFT decomposes human beliefs into five moral foundations, each with its own virtue and vice axis: *Authority/Subversion*, *Care/Harm*, *Fairness/Cheating*, *Loyalty/Betrayal* and *Purity/Degradation*. Approaches including lexicons (Araque et al., 2020; Frimer et al., 2019; Graham and Haidt, 2012; Hopp et al., 2021) and supervised machine learning models (Beiró et al., 2023; Huang et al., 2022; Lin et al., 2018) have been employed to classify text according to MFT.

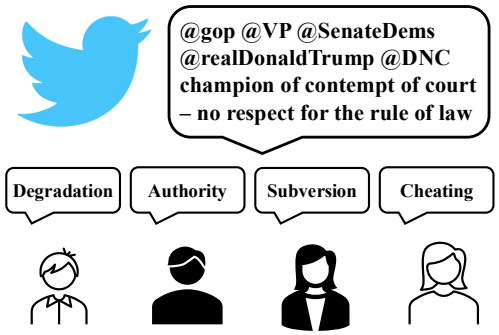


Figure 1: A tweet example with disagreeing labels given by four annotators in the Moral Foundations Twitter Corpus (Hoover et al., 2020), demonstrating how individuals interpret moral expressions differently.

Recently, transformer-based large language models have shown promising results in classification tasks. BERT in particular have been widely applied due to its ability to generate rich contextual and semantic embeddings (Devlin et al., 2019). Fine-tuning BERT with crowdsourced data has become a standard and effective approach for achieving state-of-the-art performance in moral value classifications (Guo et al., 2023; Nguyen et al., 2024; Preniqi et al., 2024).

However, moral judgement is inherently subjective—individuals hold different beliefs and inhabit different contexts that lead to them inter-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 83–94

preting content differently and prioritising different foundations. Additionally, classifying texts from social discourse further extends the subjectivity due to the ambiguity of text data, particularly in highly abbreviated contexts such as tweets. Hence, crowdsourced training datasets in this field often exhibit substantial disagreement among annotators (Figure 1), yet existing work typically disregards this information, treating disagreement as unstructured noise. They typically train a universal classifier treating moral classification as if there exists a "ground truth" that can be derived from aggregated annotators' responses.

In reality, the disagreements are an important part of the content of the analysis of morality. A model trained to derive one "smoothed out" response will miss the inherent conflicts and subtleties that are so much of the human experience. A more sophisticated approach is to build models that learn how individuals differ in their judgement.

In this work, we take a step towards this goal by training classifiers that model annotators in crowdsourced data specifically. We do this by adding a neural network layer on top of finetuning of BERT. This additional layer learns how a particular annotator interprets moral content differently from the shared text embeddings. Furthermore, the added layer can be seen as an interpretable representation that capture meaningful differences between individuals, revealing biases and tendencies towards different moral values. Our results show substantial improvement in predicting individual annotations and highlight concerns that training classifiers on aggregated labels may appear highly accurate but mask inconsistencies across annotators.

This work makes three key contributions:

1. A modelling approach that captures individual perspectives in moral value classification, accounting for task subjectivity;

2. We demonstrate that annotator biases and tendencies in crowdsourced datasets are learnable features rather than noise, where our model yields on average, a 10.2 % improvement in classification accuracy when compared with finetuned BERT over five foundations; and

3. We raise concerns about aggregating annotations into a single ground truth label in such task, urging future modelling approaches to incorporate individual-level variation.

2 Related Work

2.1 NLP for MFT

Many studies focus on supervised learning approaches where classifiers are trained to map texts to moral values. Classical machine learning models such as logistic regression and support vector machines, as well as deep learning models such as long short-term memory networks, have been widely adopted (Araque et al., 2020; Beiró et al., 2023; Hoover et al., 2020; Lin et al., 2018; Trager et al., 2022). More recent work uses transformer-based pretrained language models, particularly BERT and its variants. Trager et al. (2022) report baseline performance from finetuning BERT models on a labelled dataset of Reddit posts. Nguyen et al. (2024) and Preniqi et al. (2024) provide in-depth analyses of BERT fine-tuning procedures and evaluations of performance in practice. Supervised learning approaches have shown promising results in classification tasks. However, training of supervised models typically rely on large-scale crowdsourced datasets that often exhibit annotation disagreement (as noted earlier). Existing works often handle this disagreement through label aggregation, implicitly assuming the existence of a ground-truth label that can be derived through summarisation or averaging of the various inputs. Yet studies in moral psychology demonstrate that moral judgement is inherently subjective and varies across individuals (Haidt, 2012), indicating that aggregation may hide meaningful differences in how people interpret the same content.

2.2 Subjectivity in Moral Judgement

Human moral judgement is widely recognised as subjective and shaped by individual differences. Haidt's social intuitionist model highlights how moral judgements arise from intuitive, socially and culturally shaped processes (Haidt, 2001). This model later informed MFT, which underpins most NLP research on moral value classification and posits that moral judgement varies across individuals (Haidt, 2012). Cultural differences were among the most influential factors shaping variability in moral judgement. Early work shows that people from different countries make different moral evaluations, even when presented with the same scenarios (Haidt et al., 1993). Subsequent studies using the Moral Foundations Questionnaires further validated the impact of demographic and cultural differences in moral intuitions (Atari et al., 2023;

Graham et al., 2011). Furthermore, studies also show that interpretation of morality shifts depends on social identities (Ellemers and Van der Toorn, 2015; Koleva et al., 2012). One must then consider individual perspectives when forming moral judgement. Liscio et al. (2022) explicitly analyse the relationship between model performance and annotator agreement, and suggest modelling approaches to incorporate annotator (dis-)agreement.

2.3 Data Perspectivism

Recent work on subjective NLP tasks has increasingly challenged the assumption of having a single“ground truth" and the use of aggregated labels, advocating instead for modelling individual annotator perspectives. Under the emerging paradigm of *data perspectivism*, disagreement in annotation is treated as a meaningful signal rather than noise (Cabitza et al., 2023). Prior studies have explored various multi-annotator learning strategies, including incorporating annotator statistics and learning personal latent vectors (Kanclerz et al., 2022), learning multi-task and multi-label frameworks (Davani et al., 2022), and annotator-aware representations (Mokhberian et al., 2024). These approaches consistently show that modelling individual annotation patterns can achieve comparable and sometimes better performance to majority-vote baselines while better capturing uncertainty and variability in human judgement. Additionally, methods that leverage annotator metadata demonstrate that improvements often arise from learning annotator-specific behaviour rather than shared demographic patterns (Orlikowski et al., 2025). Within MFT NLP, prior work has explored perspectivist approaches from different angles. Golazizian et al. (2024) explored cost-efficient approaches combine multi-task learning with few-shot annotator adaptation to incorporate new annotator perspectives while reducing annotation cost. In parallel, Alvarez Nogales and Araque (2024) take an early step by training separate classifiers on subsets of annotators and combining their outputs via prompt-based ensemble; however, annotation sparsity and limited per-annotator data lead to substantial variability in annotator-specific predictions.

Collectively, this line of work highlights the importance of preserving perspectives in subjective tasks and motivates approaches, such as ours, that directly model individual annotators in MFT NLP and reveal insights into the impact of annotators and the subjective nature of moral foundations.

3 Data

In this work we use the **Moral Foundations Twitter Corpus** (MFTC) (Hoover et al., 2020), comprising 35,108 tweets collected from Twitter (now X), across seven topics identified by hashtags. Twenty-three annotators were trained to manually assign labels for moral foundations and their polarity expressed in each separate tweets. An additional “non-moral" label is included for annotators to flag tweets with no presence of any moral foundations. Every tweet received between 3 to 8 annotations, the majority receiving 3 or 4.

Foundation	Moral	Absent	Proportion
Authority	18473	109945	14.39
Care	26141	102277	20.36
Fairness	24635	103783	19.18
Loyalty	17437	110981	13.58
Purity	11982	116436	9.33

Table 1: Annotation distributions for all five foundations, with the proportion of annotations for moral classes (*virtue* and *vice*), showing class imbalance between the moral and absent classes.

Table 1 shows the label distribution for all five foundations, where each foundation is considered for the entire dataset. There exists a substantial class imbalance between the moral classes and the absent class.

#Annotator	Mean	Median	Min	Max	s.d
23	5585	4588	560	19556	4747

Table 2: Summary statistics of the number of tweets each annotator has annotated, showing the sparsity of annotation structure.

While the label distribution provides an overview of class prevalence, it does not capture how annotations are distributed across annotators. The variability in contributing annotations adds more complexity to the problem and we first show that with summary statistics in Table 2. On average, each annotator labelled about one-seventh of the entire tweet collection. The large standard deviation highlights the variability of numbers of tweets each annotator labelled—a small number of annotators covered a majority of the dataset. This imbalance leads to sparse co-annotation, as visualised in Figure 2. Edges in the graph have weights that represent the number of tweets two annotators both labelled and edges with weight less than 500 are re-

moved to identify clusters where annotators within one cluster have been presented similar twitter content. The disconnectedness of the co-annotation network indicates that subsets of annotators never label any common items. As a result, information does not propagate across all annotators, preventing us from identifying global latent variables that helps explain annotator labelling ability and item labelling difficulty. This motivates the consideration of modelling approaches that are not restricted by disconnectedness and can make use of such locally structured annotation data, which we describe in the following section.

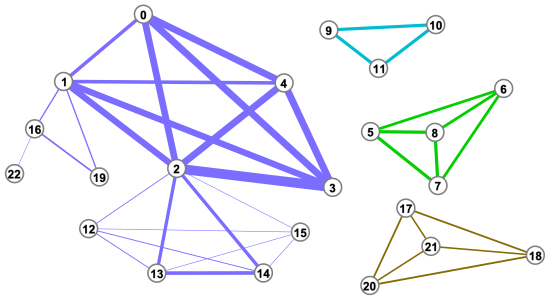


Figure 2: Co-annotation network between annotators. Edges represent the number of tweets co-annotated, and edges with weight less than 500 are filtered out. There exists several clusters of annotators which allows us to compare within groups of annotators that are labelled similar items. The disconnectedness shows that information is not shared across all annotators.

Given the large amount of missing values in the annotation structure as shown above, we report inter-annotator agreement using Krippendorff's α (Krippendorff, 2018) for each foundation in Table 3. An α value of 0 indicates agreement at the level of chance, while a value of 1 indicates perfect consensus. We observe low to moderate agreement across all foundations, with *Authority*, *Loyalty* and *Purity* exhibiting noticeably lower agreement than the others, suggesting a higher degree of subjectivity in their labelling. Overall, these highlight the need to explicitly capture diverse perspectives, as simple aggregation methods may be inadequate under such low levels of agreement.

Authority	Care	Fairness	Loyalty	Purity
0.263	0.349	0.401	0.301	0.254

Table 3: Krippendorff's α computed across 23 annotators per foundation, indicating low to moderate inter-annotator agreement and reflecting the subjective nature of moral-value labelling.

Some prior work has disregarded the polarity of a foundation by merging virtue and vice labels or treated virtue and vice dimensions as two individual foundations. These approaches allow researchers to simplify the problem and avoid additional "noise". However, we preserve the virtue and vice labels, as we aim to study the subjectivity of moral judgement not only identifying whether a foundation is expressed, but also in how annotators differ in assigning opposing moral valences.

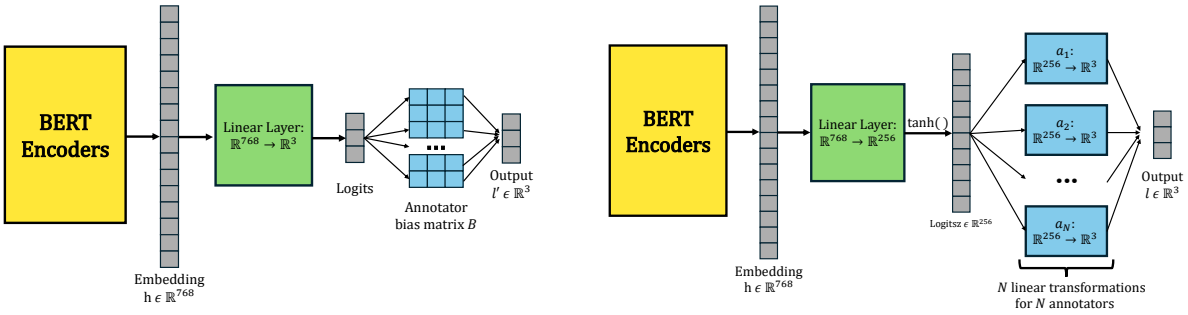
4 Methods

4.1 Problem Setup

Our goal is to predict the moral foundation expressed in text while accounting for individual perspectives of the human annotators. Each text instance x_i is annotated by annotator a_j in a group of N annotators, producing a set of labels for all present moral values. For each foundation $k \in \{1, ..., 5\}$, we extract the label $y_{ij}^{(k)}$ and consider the tuple $(x_i, a_j, y_{ij}^{(k)})$ as a single observation. This way we keep annotators' individual labels rather than aggregating annotations into a single "ground truth" label. We simplify the multi-label classification problem into single-label, multi-class classification tasks by considering each foundation separately. We train a separate classifier $f_k(x_i, a_j; \theta_k)$ for each of the five moral foundations (*Authority, Care, Fairness, Loyalty, Purity*) to predict $y_{ij}^{(k)} \in \{1, 2, 3\}$, representing labels *virtue, vice* and *absent*.

4.2 Model Overview

Our model extends a finetuned BERT classifier by incorporating neural network structures that learn annotator-specific features. This design is inspired by the *Crowd Layer* framework, which applies neural network designs that directly learn from crowd-sourced labels from multiple annotators (Rodrigues and Pereira, 2018). Adapting this framework to our setting, we introduce the **Annotator Layer** that adjusts the shared text features according to the annotators, allowing the model to capture systematic difference in annotation patterns and annotators' individual perspectives. The model has 3 parts (Figure 3). Firstly, the pretrained language model BERT takes texts as inputs and outputs contextual embeddings. We use `BERT-base-uncased` to encode text x_i, producing an embedding $h_i \in \mathbb{R}^{768}$ from the final hidden layer `[CLS]` token. The BERT model is



(a) Bias-only variant that adjusts the final output probability based on the annotators' biases towards each class, yielding an interpretable bias matrix B.

(b) Linear transformation variant that expresses annotators' labelling patterns with linear layers, yielding better representation power of the individual perspectives.

Figure 3: BERT finetuning with **Annotator Layer** that models annotators' individual labelling pattern and features.

finetuned with the following layers during the training process. The second component is a projection layer that maps the 768-dimensional BERT hidden representation to a lower-dimensional feature vector. The output dimensionality is determined by the following Annotator Layer variants.

Bias-only The objective of this variant is to provide interpretable annotator features learned from data. The projection layer has an output dimension of 3 that corresponds to the number of classes. It applies a linear transformation on the embedding h_i to get the base logits l_i: $l_i = Wh_i + b$.

For the Annotator Layer, we use an $N \times c$ matrix for N annotators where each row of the matrix corresponds to the biases of an annotator towards each class. For the base logits l_i with annotator id a_j, we adjust logits according to the annotator by computing $l'_{ij} = l_i + B_j^T$, where $B \in \mathbb{R}^{N \times c}$ is the annotator bias matrix and B_j^T denotes the transpose of the j-th row, representing annotator a_j's bias across all classes.

Linear Transformation The objective of this variant is to provide greater predicting power as we use much more parameters to represent the annotators. The projection layer has a tunable output dimension which we choose to use 256 as an intermediate value between 768 and 3. We apply a *tanh* activation function to the output. For the Annotator Layer, each annotator a_j is a linear transformation with a 3×256 weight matrix and a 3×1 bias vector. We compute the adjusted logits by $l'_{ij} = W_{a_j} z_i + b_{a_j}$.

Both variants produce a 3 dimensional vector l'_{ij}, we then apply a *softmax* function to yield the predicted probability distribution:

$$p_{ij} = \text{softmax}(l'_{ij}), \quad \hat{y}_{ij}^{(k)} = \arg\max_c p_{ij}^{(c)},$$

and the predicted class $\hat{y}_{ij}^{(k)}$ for text x_i and annotator a_j is the class with the greatest probability.

4.3 Training Objective

We train the model with the cross-entropy loss and include two regularisation terms:

1. **L2 Norm (Weight Decay):** standard L2 regularisation is applied to the model parameters to prevent overfitting.

2. **Centred Bias Penalty:** For the bias-only variant of the Annotator Layer, we add a centred penalty on the bias matrix B:

$$\mathcal{R}_{\text{centre}} = \left\| \frac{1}{N} \sum_{j=1}^{N} B_j \right\|_2^2.$$

This is to ensure that the average annotator has zero bias towards each of the classes.

In summary, the model combines a shared representation of each text with annotator-specific adjustments. BERT encodes the input into a contextual embedding, which is then mapped to base logits via a linear classification layer. The Annotator Layer modifies these logits according to the learned parameters of the annotator who supplied the label, modelling their individual perspectives. Via backpropagation, updates to the BERT parameters incorporate annotator information, enabling the encoders to refine text representations according to annotators' various moral perspectives.

5 Experiments

Here we demonstrate two main aspects of our approach: (1) its effectiveness in predicting individual annotations, and (2) the interpretability of the learned annotator features.

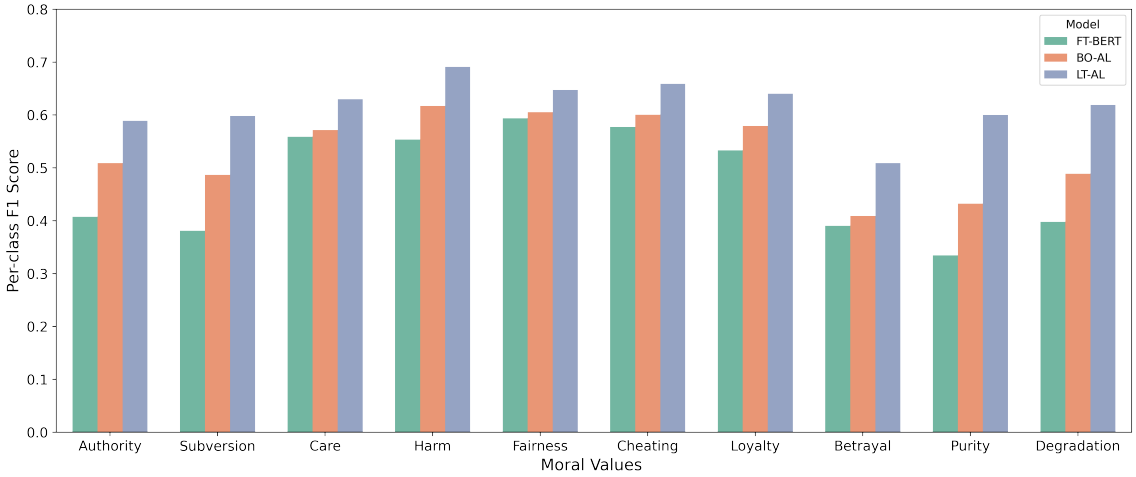


Figure 4: Average per-class F1 scores (five runs on different splits) for 10 moral values across five foundations (without class *absent*), comparing three models: FT-BERT, BO-AL and LT-AL, evaluated on raw annotations. Both BO-AL and LT-AL models outperforms the baseline FT-BERT model.

We begin by preprocessing the texts following steps outlined in Appendix A.1. We then separate the data into subsets that correspond to each of the five foundations, where we keep the cleaned texts, annotator ids and annotations. For each foundation, we create five folds using a `StratifiedGroupKFold` from scikit-learn (Pedregosa et al., 2011), which maintains the overall label distribution across folds and prevents data leakage by ensuring that all annotations belonging to the same tweet are grouped together in either the train or test set. Models are built and evaluated using the 5 non-overlapping splits and any metrics reported is an average score calculated across the five splits.

We compare the two variants of Annotator Layer (the Bias-only Annotator Layer (**BO-AL**) and Linear Transformation Annotator Layer (**LT-AL**)) with a baseline of finetuned BERT without an Annotator Layer (**FT-BERT**). The latter adds a linear classifier that maps embeddings from BERT into a 3-class probability distribution and the parameters are finetuned. Finetuned BERT is currently regarded as the state-of-the-art approach in moral value classification. We deploy the same pretrained `BERT-base-uncased` and apply an identical training process wherever possible, allowing our experiments to also function as an ablation study. Training details and values of hyperparameters are recorded in Appendix A.2. We report classification performance using F1 scores to better reflect performance (than classification accuracy) under the dataset's imbalanced label distribution.

6 Results

By modelling annotator-level features, Annotator Layers yield clear prediction improvements for individual annotations compared to the baseline.

	FT-BERT	BO-AL	LT-AL
Authority	57.3	64.6	**71.3**
Care	67.4	70.2	**75.1**
Fairness	69.7	71.1	**74.9**
Loyalty	62.1	64.3	**70.0**
Purity	56.3	62.7	**72.9**
Overall	62.6	66.8	**72.8**

Table 4: Macro F1 scores for each foundation across three models: FT-BERT, BO-AL, and LT-AL. The addition of Annotator Layer improves classification performance, where the linear transformation variant yields the most improvement of 10.2% in F1 score, averaged across all foundations.

Figure 4 shows that as we increase the complexity of neural network structures that represent the annotators (from none for FT-BERT, to linear layers for LT-AL), the performance improves across all foundations. With more model parameters, the LT-AL model has greater representational power for modelling individual annotators, leading to a largely improved classification performance over the baseline when predicting individual annotations. Even the BO-AL model yields a clear performance gain, despite adding only a small bias matrix (69 parameters). We observe particularly large improvement in F1 scores for *Authority*, *Loyalty* and *Purity*. Table 4 further validates these results, showing im-

	Virtue			Vice			Absent		
	FT-BERT	BO-AL	LT-AL	FT-BERT	BO-AL	LT-AL	FT-BERT	BO-AL	LT-AL
A	40.8	50.9	58.9	38.1	48.7	59.8	93.0	94.1	95.0
C	55.9	57.1	63.0	55.4	61.7	69.1	90.8	91.8	93.1
F	59.4	60.6	64.7	57.7	60.1	65.9	92.1	92.7	93.9
L	53.3	57.9	64.1	39.0	40.9	50.9	93.8	94.2	94.9
P	33.5	43.2	60.0	39.8	48.9	61.9	95.5	96.0	96.8

Table 5: Per-class F1 scores (Virtue, Vice, Absent) for each foundation (abbreviated by their initial letters) across three models: FT-BERT, BO-AL, and LT-AL. The greatest classification improvements occurs in the moral classes compared to the Absent class.

provements in macro F1 scores across all foundations when adding Annotator Layers.

Table 5 shows the greatest classification performance gains occur in moral classes (*virtue* and *vice*). For instance, the greatest improvement over the baseline model is the classification of *purity* (virtue aspect of foundation *Purity*) with the LT-AL model, yielding an increase of 0.265 in F1 score. All three models perform comparably when it comes to classifying the *absent* class (a tweet that is not expressing a certain moral value). However, we still see an improvement for the *absent* class with the addition of Annotator Layers, even when the baseline is already sufficiently strong.

7 Discussion

7.1 Interpretability of Annotator Layer

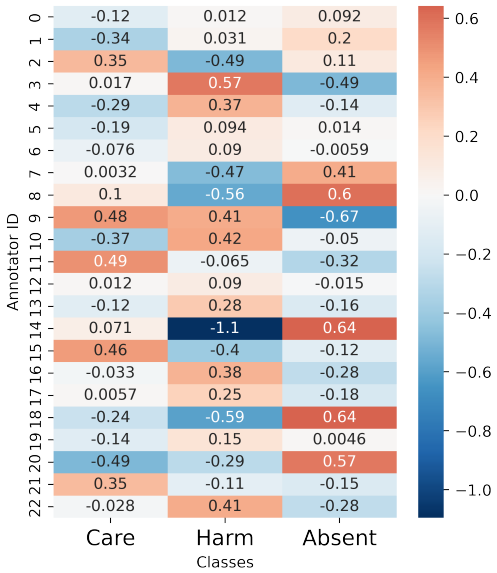


Figure 5: Bias matrix for the Care foundation extracted from the bias-only Annotator Layer. Positive values (see colourbar) indicate positive bias. This matrix gives information regarding how annotators give labels with different tendencies for the same foundation.

To illustrate the interpretability of the bias-only Annotator Layer, we extract the bias matrix from the trained models and analyse annotators' biases towards each class. Figure 5 visualises annotators' biases per foundation level. Positive values indicate biases towards a class whereas negative values indicate biases against a class. Several bias patterns are observed in the bias matrix; we use the *Care* foundation as an example. Annotator 2 shows a moderate bias toward the virtue aspect *Care* and, correspondingly, a bias against the vice aspect *Harm*. However, this complementary behaviour between the two polarities does not always hold. When an annotator possesses a tendency towards or against one moral class, the complementary class may instead be *absent*. We observe this pattern in several annotators (e.g. Annotator 3 and 7). In some cases, annotator exhibit biases towards or against both moral classes, with the *absent* class acting as the complement.

We also show bias weights for a subset of annotators for the *Care* and *Fairness* Foundations. Annotators 5, 6, 7 and 8 belong to the same co-annotation cluster (Figure 2) meaning they were presented similar tweet contents. Hence, we pick this group to show how the labelling patterns share similarity and difference across foundations. Figure 6 shows the biases over two foundations for the group of selected annotators. We observe that Annotators 7 and 8 show a consistent tendency to favour the *absent* class, with only minor differences in their biases toward the *virtue* classes across the two foundations. In contrast, Annotators 5 and 6 both have different tendencies between the two foundations. Both show similar patterns for *Care*, favouring the *vice* class and labelling against the *virtue* class. However, for *Fairness*, they exhibit opposite patterns. Annotator 5 favours the *absent* label and gives fewer *virtue* labels, whereas Annotator 6 shows the reversed behaviour. These obser-

vation suggests that, while some bias patterns are consistent, they should be analysed separately for each foundation rather than assuming a universal annotator bias.

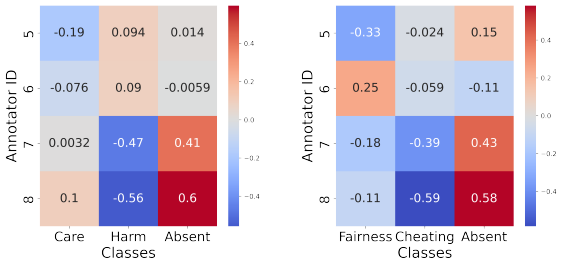


Figure 6: Bias Heatmaps of Foundations Care (Left) and Fairness (Right), for Annotators 5, 6, 7 and 8, showing annotators with similar bias patterns in one foundation can have opposite bias patterns in another.

By summarising these observations, we may identify groups of annotators who share similar perspectives when interpreting moral values for each foundation. These groups exhibit consistent patterns in moral judgement, such as virtue/vice-oriented annotators, annotators who frequently give moral labels, and those who give moral labels more cautiously, leading to dominating non-moral (absent) annotations. Such patterns suggest meaningful "annotator types", revealing insights into the diversity of moral judgement and providing potential categorisation for all individuals, not just annotators. This provides possible modelling approaches that learn group behaviours instead of modelling individual annotators, such as mixture-of-experts models where experts represents groups of people with similar perspectives. One can also study correlations between individual political ideology, cultural background and other demographic factors with the "types" that we identify, gaining insights into the development of diverse perspectives.

To examine whether these bias patterns capture information beyond simple annotator-level label preferences, we construct a non-learned empirical baseline using relative label preferences $P(c|a)/P(c)$, where each annotator's labelling distribution is normalised by the overall class distribution to account for class imbalance. Replacing the learned bias matrix with this empirical preference yields comparable performance, suggesting that both approaches shift the predicted probability distribution in similar directions for each annotator. However, we observe that while the direction of these shifts is broadly aligned, the magnitude of the learned biases differs from the empirical preference matrix. This indicates that the learned bias is not merely reproducing simple dataset statistics. The Jensen-Shannon divergence between the learned bias and empirical preference matrices averages 0.446 ± 0.070 across annotators, ranging from 0.329 to 0.562, indicating that the two representations are not closely aligned at the distribution level. Moreover, the text encoder is jointly finetuned with the bias matrix and may encode annotator related information, which is not isolated in this analysis. These results suggest that, despite similar empirical performance, the bias-only model still captures meaningful structure beyond label frequency statistics.

7.2 Raw Annotations and Aggregated Labels

Using the trained models with Annotator Layers, we apply a simple rule to obtain an aggregated label for each tweet. For each tweet, we activate all "annotators" in the Annotator Layer, regardless of their IDs, and obtain 23 predictions, yielding 23 probability distributions over the 3 classes. We then average these distributions and select the class with the highest mean probability as the aggregated prediction. For this analysis, we train the baseline model FT-BERT directly on aggregated labels, mimicking the common practice used when fine-tuning BERT for moral value classification.

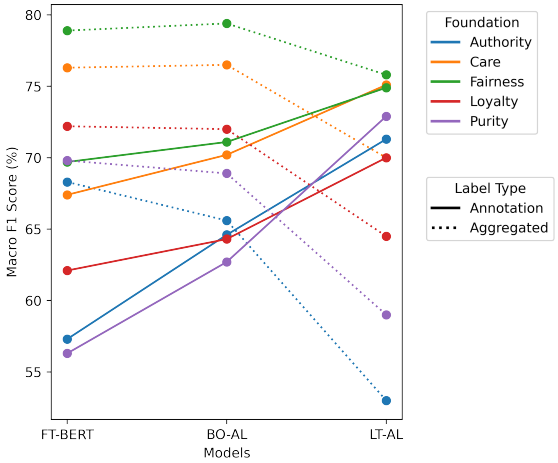


Figure 7: Trend of macro F1 scores as the modelling capacity to represent the annotators increases, compared for predicting raw annotations and aggregated labels. As the capacity to model individual perspectives increases (from FT-BERT to LT-AL), we see an increase in classification performance on raw annotations and a decrease in classification performance on aggregated labels.

We show the changes in macro F1 scores in Fig-

ure 7 as we move from standard finetuned BERT, to the two variants of the Annotator Layers. The horizontal axis can be interpreted as an increasing capacity to model annotator-specific features from left to right. We've already shown in the **Results** section that as we increase the capacity, the prediction accuracy of raw annotations also increases (solid lines in Figure 7). When evaluating on aggregated labels, the FT-BERT and BO-AL models achieve comparable performance, whereas the LT-AL model shows a substantial decrease. This decrease when evaluating on aggregated labels is more obvious where we see the largest gain when evaluating on raw annotations. Interestingly, the transition from BO-AL to LT-AL yields the strongest improvement on raw annotations, but it also produces substantial decrease under aggregated-label evaluation. This is expected, the LT-AL models are designed to capture annotator-specific features and better learn how individuals interpret texts and assign labels. When activating all annotators and collapsing their outputs into a "consensus" label, additional noise is introduced due to the sparse annotation structure. In essence, we are asking the learned annotator representations to make predictions on tweets that the corresponding annotators never see, with potentially great domain differences. These observations demonstrate that a model that is capable of representing annotators' individual perspectives does not necessarily agree with the aggregated labels. This highlights an important limitation of training models on aggregated labels: a strong performance by such a model may hide substantial underlying variations in individual perspectives and may not reflect the true effectiveness of the model.

7.3 Ambiguity of Foundations

Foundation	Type	Virtue	Vice	Mean
Care	I	7.1	13.7	10.4
Fairness	I	5.3	8.2	6.8
Authority	B	18.1	21.7	**19.9**
Loyalty	B	10.8	11.9	**11.4**
Purity	B	26.5	12.1	**19.3**

Table 6: Performance improvement in F1 scores (%) between LT-AL and FT-BERT for the five moral foundations. The binding (B) foundations benefits more from modelling annotators' individual perspectives when compared to the individualising (I) foundations.

While the addition of Annotator Layers improves overall performance, the gains vary across foundations, suggesting that some moral foundations exhibit greater ambiguity and therefore benefit more from annotator-specific modelling. Greater improvements occur in *Authority*, *Loyalty*, and *Purity*, compared to *Care* and *Fairness* (Table 6). This pattern mirrors the distinction between individualising foundations (*Care, Fairness*) and binding foundations (*Authority, Loyalty, Purity*). Individualising foundations are generally considered more morally relevant and are endorsed across the political spectrum, whereas binding foundations tend to receive endorsement from a smaller portion of the population (Graham et al., 2009). We've shown that human annotators exhibit lower agreement on the binding foundations (Table 3), as measured by Krippendorff's α, indicating greater ambiguity. This pattern align with both the observed performance gains and the psychological distinctions between binding and individualising foundations.

8 Conclusion

In this work we introduced the Annotator Layer for moral classification of texts that captures annotator-specific moral perspectives and annotation patterns, extending on finetuning BERT models. Our experiments demonstrate improved classification performance of individual annotations in crowdsourced dataset, along with interpretable representations of annotators' bias patterns. It is shown that disagreement between annotators in such subjective tasks is a learnable feature instead of annotation noise. The results suggest that relying solely on aggregated labels can hide important information. We hope this work encourages future research to move beyond training a universal classifier that predicts a "ground truth" and develop models that better reflect diversity of moral judgement and understand the subjectivity of moral classification of texts.

9 Limitations

Our work has two primary limitations.

First, although the dataset publication notes that annotator metadata (e.g., demographic information, political ideology and moral values measured by MFQ) exists, this information was not available to us and is therefore not incorporated into the analysis. As a result, while the Annotator Layer learns annotator-specific features and identified potential differences of annotation patterns between groups of annotators, we cannot directly examine

how these patterns relate to known characteristics. Studies in moral psychology have validated that these characteristics have a direct impact to human moral judgement. Hence, access to such data can help validate the learned representations and provide explanations to some of the observed patterns.

Second, our approach does not provide a strong mechanism for producing high-quality aggregated predictions to moral values. We've demonstrated the bias-only variant's comparable classification performance to finetuned BERT on aggregated labels, but the linear transformation variant has shown a substantial decrease in performance. Many downstream applications ultimately requires a single, aggregated label for a text observation, yet our annotator-specific models requires annotator (human) information to provide accurate predictions which typically lacks in these tasks. The model learns fine-grained human-specific behaviours and does not generalise well for aggregated labels. Our naive approach to obtain consensual labels by activating all annotator corresponding neural network structures and averaging the prediction distributions introduces noise, especially given the sparse and uneven co-annotation structure. Developing principled aggregation methods that leverage annotator features is a vital future direction.

Acknowledgements This work was supported with supercomputing resources provided by the Phoenix HPC service at Adelaide University.

References

Anny D. Alvarez Nogales and Oscar Araque. 2024. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 67–77, Torino, Italia. ELRA and ICCL.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowl. Based Syst.*, 191(105184):105184.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *J. Pers. Soc. Psychol.*, 125(5):1157–1188.

Mariano Gastón Beiró, Jacopo D'Ignazi, Victoria Perez Bustos, María Florencia Prado, and Kyriaki Kalimeri. 2023. Moral narratives around the vaccination debate on facebook. In *Proceedings of the ACM Web Conference 2023*, page 4134–4141, New York, NY, USA. Association for Computing Machinery.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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. ACM.

Naomi Ellemers and Jojanneke Van der Toorn. 2015. Groups as moral anchors. *Curr. Opin. Psychol.*, 6:189–194.

J. A. Frimer, R. Boghrati, J. Haidt, J. Graham, and M. Dehgani. 2019. The moral foundations dictionary for linguistic analyses 2.0. https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/moral-foundations-dictionary/. Accessed: 25 November 2025.

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, Miami, Florida, USA. Association for Computational Linguistics.

Jesse Graham and Jonathan Haidt. 2012. The moral foundations dictionary. https://moralfoundations.org/other-materials/. Accessed: 25 November 2025.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, Advances in experimental social psychology, pages 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.*, 96(5):1029–1046.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *J. Pers. Soc. Psychol.*, 101(2):366–385.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:281–291.

J Haidt, S H Koller, and M G Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *J. Pers. Soc. Psychol.*, 65(4):613–628.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.*, 108(4):814–834.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Penguin UK.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Soc. Justice Res.*, 20(1):98–116.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Personal. Sci.*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behav. Res. Methods*, 53(1):232–246.

Xiaolei Huang, Alexandra Wormley, and Adam Cohen. 2022. Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, New York, NY, USA. ACM.

Julie Jiang, Luca Luceri, and Emilio Ferrara. 2025. Moral values underpinning COVID-19 online communication patterns. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2642–2650, New York, NY, USA. ACM.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, 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.

Spassena P Koleva, Jesse Graham, Ravi Iyer, Peter H Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *J. Res. Pers.*, 46(2):184–194.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. Sage Publications, Thousand Oaks, CA.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 552–559.

Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. 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 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. 2024. Measuring moral dimensions in social media with mformer. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1134–1147.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn:

Machine learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: A fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 433–442, New York, NY, USA. ACM.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. 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.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Alvin Zhou, Wenlin Liu, Hye Min Kim, Eugene Lee, Jieun Shin, Yafei Zhang, Ke M. Huang-Isherwood, Chuqing Dong, and Aimei Yang. 2024. Moral foundations, ideological divide, and public engagement with u.s. government agencies' COVID-19 vaccine communication on social media. *Mass Communication and Society*, 27(4):739–764.

A Training Details

A.1 Text Preprocessing

We clean the twitter texts by first removing URLs, non-alphanumeric characters, punctuations and retweet markers. All text is then lowercased, and user mentions are replaced with the token "@user". Stopwords may optionally be removed, though we found this to have negligible effect on model performance.

A.2 Training Process and Hyperparameters

We implement and train all models using Pytorch (Paszke et al., 2019) v2.7 and optimise the parameters using the AdamW optimiser (Loshchilov and Hutter, 2017). The initial learning rate is 2e-5 for the BERT parameters and is 1e-4 for the parameters in the linear layer and Annotator Layer, with linear decay and no warm-up. The lower learning rate is used to update the parameters of BERT moderately, avoiding deterioration of BERT's ability of capturing semantic and contextual meaning with the embeddings. In training, we use a batch size of 8, and the maximum input text length is set to be 64 tokens as all texts in the dataset are short in length. We set the L2 regularisation coefficient to 0.01 and the centred bias penalty to 0.05. We train the models for 5 epochs and freeze the BERT parameters during the first epoch to allow the newly added layers to stabilise before full fine-tuning. All experiments are run using one Nvidia A100-SXM4-40GB GPU.

Launch and Aftermath: Contrasting Social Media Responses to Chatbot Releases. The Cases of Meta's Galactica and OpenAI's ChatGPT

Maximilian Weber[1], **Johannes B. Gruber**[2]

[1]University of Mainz, Germany, [2]GESIS, Germany

Correspondence: maximilian.weber@uni-mainz.de

Abstract

In November 2022, Meta's Galactica and OpenAI's ChatGPT were released within fifteen days of each other, two transformer-based language models that were architecturally similar and built on comparable underlying technology, yet experienced starkly different outcomes. Where they diverged was not in technical kind but in domain positioning and epistemic framing: Galactica was explicitly marketed as a reliable scientific assistant, while ChatGPT was presented as a general-purpose conversational tool. Using Twitter data collected via the Twitter Research API, we conduct a comparative analysis of early social media discourse surrounding both models. Through sentiment classification, zero-shot harm and risk annotation, and LLM-based topic modeling, we find that negative sentiment escalated rapidly for Galactica while remaining comparatively stable for ChatGPT in the release period. Galactica experienced a marked escalation in criticism during its first week, eventually structuring much of the conversation. In contrast, ChatGPT's early discourse remained more evenly distributed across hype, experimentation, practical engagement, and criticism. We argue that domain positioning and epistemic expectations, rather than any meaningful technological difference, played a central role in shaping public perception, with Galactica's scientific presentation making its well-documented hallucinations appear far more damaging in public opinion.

1 Introduction

In November 2022, two large language models (LLMs) were released to the public with strikingly different trajectories. Meta introduced Galactica on November 15, 2022, a specialized LLM designed explicitly for scientific applications. Introduced as capable of summarizing academic literature, solving mathematical problems, generating Wikipedia articles, writing scientific code, and annotating molecules and proteins, the model was positioned as a tool for the research community (Taylor et al., 2022). However, within just two days of its public demo launch, Meta withdrew the service following criticism. According to Meta's Chief AI Scientist Yann LeCun, the model was effectively driven offline by public backlash: "Galactica, the LLM for scientists from Meta [...]. It was murdered by a ravenous Twitter mob. The mob claimed that what we now call LLM hallucinations was going to destroy the scientific publication system" (Yann LeCun [@ylecun], 2023).

In contrast, OpenAI's ChatGPT, released just fifteen days later on November 30, 2022, experienced rapid and widespread adoption, becoming a mainstream phenomenon that reportedly reached 700 million weekly active users within three years of its launch.[1]

This contrast provides a quasi-experimental opportunity to examine the two model launches: Why did one model face shutdown within days, while the other achieved remarkable success? Were the concerns raised about Galactica equally applicable to ChatGPT, but less central to the overall public discourse?

This study addresses three research questions through comparative analysis of social media discourse. First, how did public discourse differ between the two launches? Second, did harm- and risk-related discourse escalate within Galactica tweets over time? Third, did the target domain (scientific knowledge versus general-purpose use) affect public reception?

2 Previous research

While a growing body of research has examined public discourse around LLM releases on social media, particularly around ChatGPT (Koonchanok et al., 2024; Weber, 2024; Rauchfleisch et al.,

[1] https://openai.com/index/how-people-are-using-chatgpt/ (OpenAI usage report, as of mid-2025).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 95–102

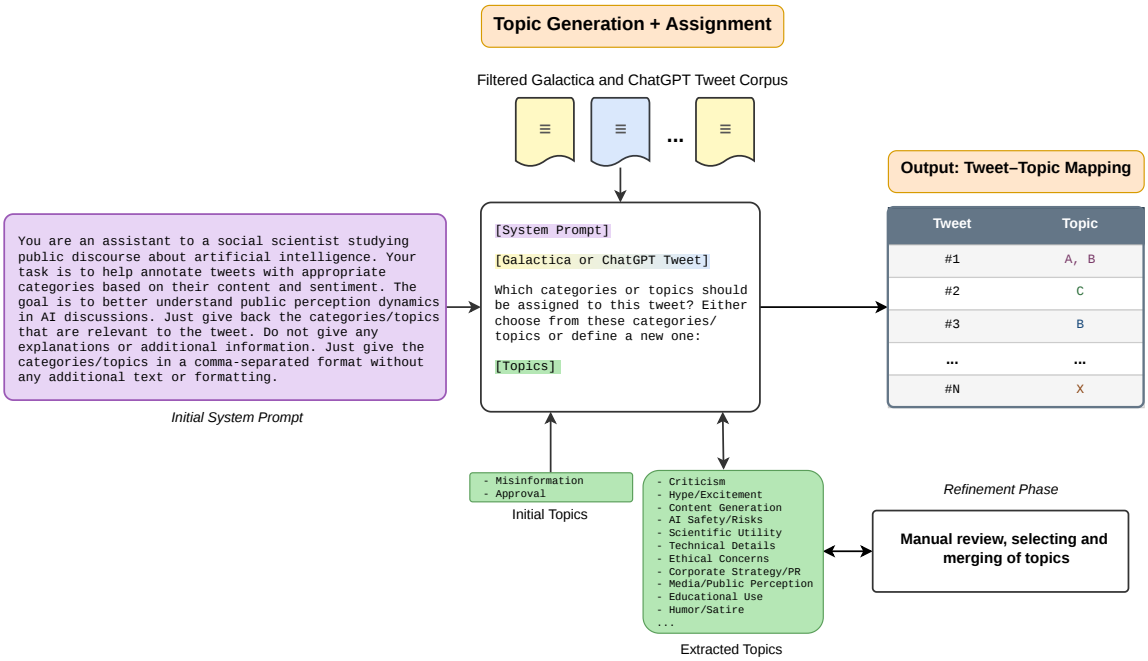


Figure 1: Overview of the Topic Generation and Assignment Pipeline. Given a filtered corpus of Galactica and ChatGPT tweets and a set of manually-curated initial topics, an LLM iteratively assigns topics to each tweet. After every 100 documents, the framework enters a refinement phase in which topics are manually reviewed, merged, and selected, yielding a final curated topic list and a topic assignment for all sampled documents.

2025), comparative analyses of social media reactions to Galactica versus ChatGPT remain absent from the literature. Chartier-Edwards et al. (2024) provide a critical account of the Galactica release, arguing that the controversy reflects tensions between AI for science promissory marketing claims, and epistemic expectations in scientific domains. They further highlight concerns about scientific misinformation and the framing of LLMs as epistemic oracles. However, they do not conduct a large-scale analysis of social media discourse, which we address in this study.

3 Data and Methods

3.1 Data

We conduct an analysis of English-language Twitter discourse during the initial release periods of both models. Using the Twitter Research API, we collected 109,694 initial tweets: 2,077 tweets mentioning "Galactica" from November 15 through December 15, 2022, and 107,726 tweets mentioning "ChatGPT" from November 30 through December 7, 2022, with 283 tweets mentioning both systems. To ensure comparability, our main analyses focus on the first eight days following each release.

Since the term *Galactica* is also associated with unrelated content on social media, such as the television series *Battlestar Galactica*, we first filtered out tweets that did not refer to Meta's Galactica model. To do so, we randomly sampled 250 tweets mentioning Galactica and manually annotated them as relevant or irrelevant. We then embedded all tweet texts using the snowflake-arctic-embed2 model (via rollama: Gruber and Weber, 2024) and trained a logistic regression classifier with LASSO regularization. The classifier achieved an accuracy of 0.908 and a macro F1 score of 0.873, after which it was applied to classify all remaining tweets mentioning Galactica.

3.2 Sentiment Annotation

To assess the emotional tone of tweets, we employed a RoBERTa-based classifier fine-tuned on Twitter data (Loureiro et al., 2022) and trained to annotate tweets for negative, neutral, and positive sentiment[2]. Each tweet was preprocessed following the conventions (replacing @mentions with `@user` and URLs with `http`) before classification. The model assigns a probability score to each of the three sentiment classes; the class with the highest

[2] `cardiffnlp/twitter-roberta-base-sentiment-latest`

score is taken as the predicted sentiment.

3.3 Harms and Risks Annotation

To identify Galactica tweets discussing potential harms or risks, we employed zero-shot annotation using `meta-llama/Llama-3.3-70B-Instruct` (Grattafiori et al., 2024), loaded in 4-bit NF4 quantization from Hugging Face. Each tweet was passed to the model with the following prompt: *Is this tweet about potential harm and risk of the AI model (LLM) Galactica? Answer with just yes or no.* Classification was based on the normalized probabilities of yes and no tokens. The system prompt instructed the model to act as an assistant to a social scientist studying public discourse about AI, providing context that tweets mentioning Galactica refer to Meta's large language model designed to assist scientific research. To evaluate annotation quality, two human annotators independently labeled a random sample of 100 tweets, achieving an inter-annotator agreement of $\kappa = 0.67$ (84% agreement). Evaluated against each annotator separately, the model achieved a κ of 0.688 and a macro-average F1 of 0.843 against annotator 1, and a κ of 0.766 and a macro-average F1 of 0.883 against annotator 2, indicating substantial agreement between automated annotations and human judgment.

3.4 Topic Classification

Recent work has explored the use of generative large language models for topic modeling and theme extraction through prompt-based frameworks (Pham et al., 2024; Sharma, 2025; Liu et al., 2025; van Wanrooij and Manhar, 2024). These approaches typically leverage generative LLMs to generate candidate topics and refine them through iterative prompting. We adopt a similar pipeline to Pham et al. (2024), as can be seen in Figure 1, but omit their second annotation phase in which a final fixed topic list is used to label a held-out set of documents. The resulting LLM-based zero-shot topic modeling and annotation pipeline incorporates a human-in-the-loop refinement stage, in which topics are reviewed and consolidated by the researchers. This design gives us direct control over the granularity and coherence of the final topic set, ensuring that the extracted themes are both meaningful and well-suited to the domain of AI discourse on social media.

Topic generation and assignment were performed using the open-weight Llama 3.3 70B (`llama3.3:70b-instruct-q4_K_M`), an instruction-tuned variant with 4-bit quantization, run via Ollama. The system prompt framed the task as assisting a social scientist in annotating public discourse about artificial intelligence, and restricted output to a comma-separated list of topic labels without additional explanation. Topic classification was applied to all tweets in both corpora: a random subset of 1,640 tweets for ChatGPT due to computational constraints, and the full set of Galactica tweets.

Day	Galactica (%)	ChatGPT (%)	Sig.
1	5.1 [2.0, 12.5]	12.0 [9.5, 15.2]	
2	12.1 [8.9, 16.1]	17.5 [16.6, 18.5]	*
3	35.6 [30.3, 41.3]	18.0 [17.4, 18.7]	***
4	33.6 [26.4, 41.6]	18.8 [18.1, 19.6]	***
5	49.1 [39.9, 58.3]	18.0 [17.4, 18.6]	***
6	42.4 [32.8, 52.6]	20.3 [19.8, 20.8]	***
7	31.2 [24.5, 38.8]	20.2 [19.7, 20.8]	**
8	35.8 [28.2, 44.1]	21.2 [20.6, 21.8]	***

Table 1: Negative sentiment rate (%) for the first 8 days after each model launch. Wilson 95% CIs in brackets. Significance based on Fisher's exact test (BH-adjusted): * $p < .05$, ** $p < .01$, *** $p < .001$.

4 Results

Table 1 and Figure 2 illustrate the temporal dynamics of sentiment during the first eight days following each model's release. For Galactica, negative sentiment accounts for only 5.1% of tweets on Day 1. However, this share rises rapidly over the subsequent days, reaching 35.6% by Day 3 and peaking at 49.1% on Day 5. Following the announcement that the demo was taken offline (November 17), overall tweet volume declines substantially.

ChatGPT exhibits a markedly different pattern. Although tweet volume increases steadily after release, reaching substantially higher levels than Galactica, negative sentiment remains comparatively stable throughout. On Day 1, 12.0% of ChatGPT tweets are classified as negative, and this share fluctuates between 17% and 21% across the first week. Unlike Galactica, ChatGPT does not experience a comparable escalation in negative sentiment.

As shown in Table 1, differences between the two models are statistically significant on all days except on the respective release date (Fisher's exact test), particularly from Day 3 onward. Overall, negative sentiment intensified markedly for Galactica while remaining stable for ChatGPT.

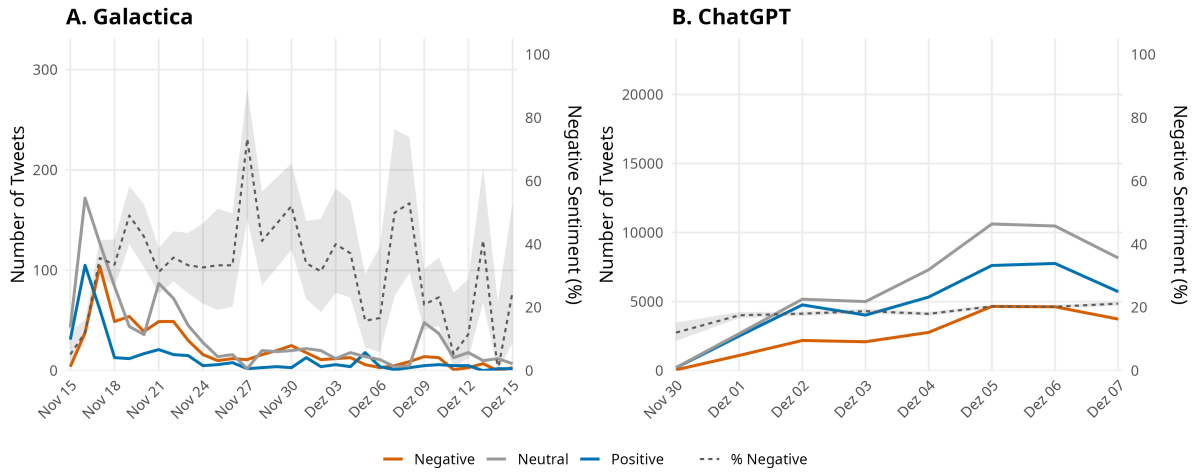


Figure 2: Temporal dynamics of Galactica (Figure A) and ChatGPT (Figure B) tweets. Lines show daily sentiment counts; the dashed line indicates the percentage of negative-sentiment tweets.

Focusing on Galactica alone, potential harm and risk annotation follows a similar pattern (Table 3). Harm-related tweets constitute only 8.3% on Day 1, but this share rises rapidly, reaching 43.9% by Day 3 and 57% by Day 5. Thus, while criticism was not dominant at launch, it came to structure a majority of Galactica discourse within the first week. Notably, following the model being taken offline on November 17, discourse shifted further toward harm- and risk-related concerns, suggesting that the withdrawal itself amplified rather than resolved the debate.

Figures 3 and 4 show the distribution of topics identified via zero-shot topic classification. ChatGPT discourse is dominated by *Innovation*, followed by *Technical Details*, *Hype/Excitement*, and *Criticism*. Although *Criticism* and *Misinformation* are present, they do not structure the majority of the conversation.

In contrast, Galactica-related discourse is led by *Criticism*, followed by *Scientific Utility* and *AI Safety/Risks*. Hype-related categories are comparatively less dominant. This suggests that Galactica's reception became increasingly structured around concerns about epistemic risk and factual reliability, particularly after the first three days following its release. During the initial three days (marked in red), however, tweets more frequently focused on *Scientific Utility* and technical details, with *Criticism* becoming more prominent thereafter.

This suggests that while concerns were present in the discourse around both models, they were relatively more prominent in the case of Galactica, particularly regarding misinformation, truthfulness, credibility, and scientific legitimacy. ChatGPT discourse, while not free of criticism, was more broadly characterized by exploratory and practical engagement.

Model	Example Tweet
Galactica	"Is this really what AI has come to, automatically mixing reality with nonsense so finely we can no longer recognize the difference?"
Galactica	"Shocked that it only took a handful of questions before Meta's new Galactica model produced racist garbage when asked about linguistic prejudice."
Galactica	"A whole new level of AI-generated academic misconduct to deal with now."
ChatGPT	"It boggles my mind how the world keeps spinning like nothing happened despite #ChatGPT. People don't understand the danger."
ChatGPT	"#ChatGPT could make it easy to cheat on written tests and homework. You can no longer give take-home exams."
ChatGPT	"ChatGPT is down and I'm having an existential crisis because I can't paste my code in. Please come back."

Table 2: Illustrative tweets from the early release period of Galactica and ChatGPT. Tweets are lightly edited for clarity and anonymized.

5 Discussion

Returning to our research questions, the findings reveal differences in the public trajectories of the two model launches. First, public discourse differed not only in tone but in temporal dynamics: while both models were accompanied by early criticism, Galactica's reception shifted rapidly toward potential harm- and risk-centered discourse, which came

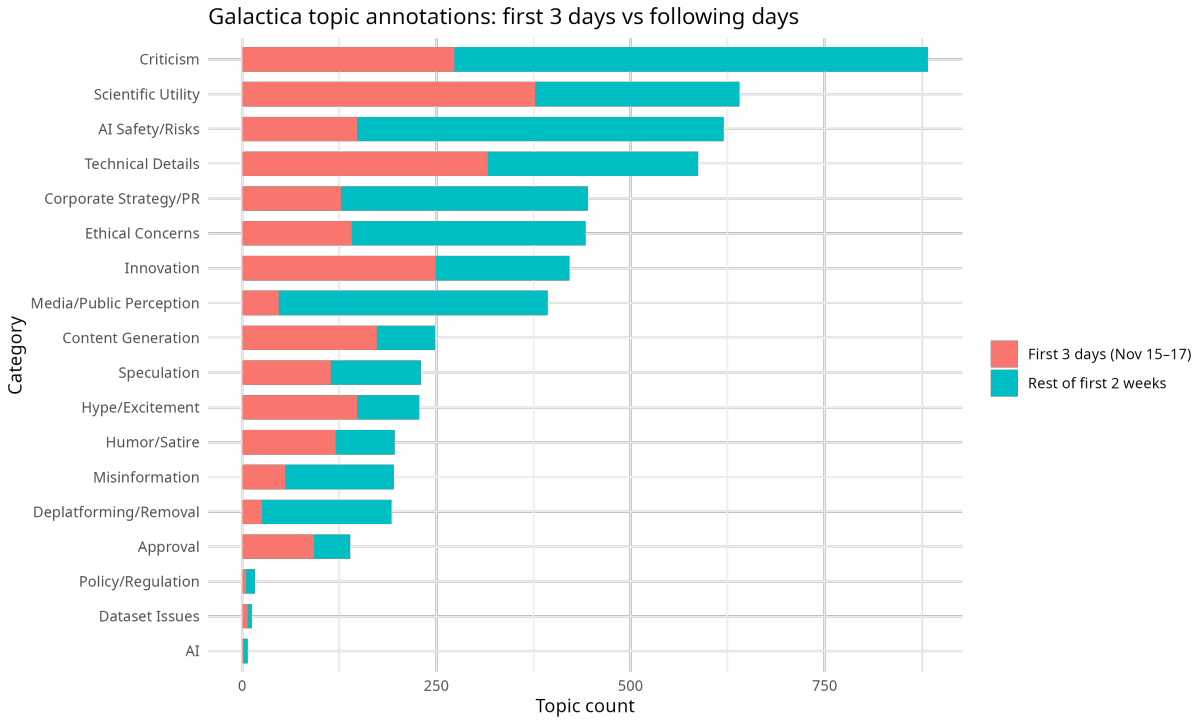


Figure 3: Topic distribution in Galactica-related tweets during the first 2 weeks following release, derived from zero-shot topic classification using Llama 3.3 (70B).

to structure a majority of tweets within days. ChatGPT discourse, by contrast, expanded in volume while maintaining a comparatively stable mix of enthusiasm, experimentation, practical engagement, and critique. Second, these patterns are consistent with the interpretation that domain positioning shaped public reception (Chartier-Edwards et al., 2024). Arguably, it is possible that ChatGPT performed better for users, leaving them with a more positive outlook on the technology. However, Taylor et al. (2022) themselves benchmarked Galactica against `text-davinci-002`, a model in the GPT-3 family (Brown et al., 2020) that underpinned the initial release of ChatGPT, and found that Galactica outperformed it on scientific knowledge probes and several bias and toxicity benchmarks. The main technical difference was the reinforcement learning from human feedback (Ouyang et al., 2022) applied in `text-davinci-003`, the model ChatGPT used at launch, which was intended in part to make the model less prone to assert falsehoods with confidence. Yet even this technical difference reflects a broader strategic divergence that led to the outcome we observed: Galactica's framing as a reliable scientific assistant likely heightened epistemic expectations, making hallucinations and factual errors normatively consequential within a domain where credibility is central. ChatGPT's general-purpose positioning, in contrast, appears to have allowed criticism to coexist with hype rather than dominate the discourse. While our design does not permit strong causal claims, the comparative evidence suggests that epistemic framing and expectation management play a critical role in shaping the early public legitimacy of AI systems.

5.1 Limitations

This study has several limitations. Our dataset comprises only original tweets, excluding retweets, and is restricted to the initial release period of both models, leaving unexamined longer-term shifts in discourse. Additionally, topic classification for ChatGPT was conducted on a random subsample of 1,640 tweets, which may not capture the full breadth of discussion around the model.

Furthermore, since the tweets analyzed predate the release of the Llama 3.3 models, it is possible that some of this content was included in the models' pretraining or fine-tuning data. This could introduce bias, as the model may have been exposed to content about Galactica and ChatGPT during training.

Following Galactica's withdrawal on November 17, discourse necessarily became retrospective, as users could no longer interact with the model.

This asymmetry with ChatGPT, which remained live throughout, should be considered when interpreting the results.

6 Conclusion

Our analysis reveals differences in public reception between Galactica and ChatGPT. Galactica faced substantially more criticism, particularly centered on concerns about misinformation and the generation of false or misleading scientific information. ChatGPT, by contrast, generated more positive discourse characterized by hype, experimentation, and demonstrations of utility, despite being subject to many of the same technical limitations. This comparative case study highlights the role of domain positioning, epistemic expectations, and expectation management in shaping public acceptance of AI technologies.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Nicolas Chartier-Edwards, Etienne Grenier, and Valentin Goujon. 2024. Galactica's dis-assemblage: Meta's beta and the omega of post-human science. *AI & SOCIETY*.

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.

Johannes B. Gruber and Maximilian Weber. 2024. rollama: An R package for using generative large language models through Ollama. *arXiv preprint*. ArXiv:2404.07654 [cs].

Ratanond Koonchanok, Yanling Pan, and Hyeju Jang. 2024. Public attitudes toward chatgpt on twitter: sentiments, topics, and occupations. *Social Network Analysis and Mining*, 14(1):106.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. 2025. LLM-guided semantic-aware clustering for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18420–18435, Vienna, Austria. Association for Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 27730–27744, Red Hook, NY, USA. Curran Associates Inc.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. 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 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Adrian Rauchfleisch, Joshua Philip Suarez, Nikka Marie Sales, and Andreas Jungherr. 2025. Winning and losing with Artificial Intelligence: What public discourse about ChatGPT tells us about how societies make sense of technological change. *Telematics and Informatics*, 103:102344.

Yash Sharma. 2025. MALTopic: Multi-Agent LLM Topic Modeling Framework. In *2025 IEEE World AI IoT Congress (AIIoT)*, pages 0707–0712, Seattle, WA, USA. IEEE.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Cascha van Wanrooij and Omendra Kumar Manhar. 2024. Topic Modeling for Small Data using Generative LLMs. In *Proceedings of the 36th Benelux Conference on Artificial Intelligence (BNAIC) and the 33rd Belgian-Dutch Conference on Machine Learning (BeNeLearn) 2024*, Utrecht, Netherlands.

Maximilian Weber. 2024. Social Group Differences in the Social Media Discussion about ChatGPT and Bing Chat. In *ACM Web Science Conference*, pages 114–118, Stuttgart Germany. ACM.

Yann LeCun [@ylecun]. 2023. Galactica, the LLM for scientists from Meta, was released a couple of weeks before ChatGPT but was taken down after... [Post].

A Appendix

Day	% Harms/Risks	95% CI
1	8.3	[3.9, 17]
2	21.5	[17.1, 26.7]
3	43.9	[38.2, 49.8]
4	45.0	[36.4, 53.9]
5	57.3	[47.6, 66.4]
6	57.6	[47, 67.6]
7	55.7	[47.2, 63.9]
8	54.5	[45.2, 63.4]

Table 3: Potential harms/risks prediction rate (%) for the first 8 days after Galactica's launch. 95% CIs based on Wilson score interval.

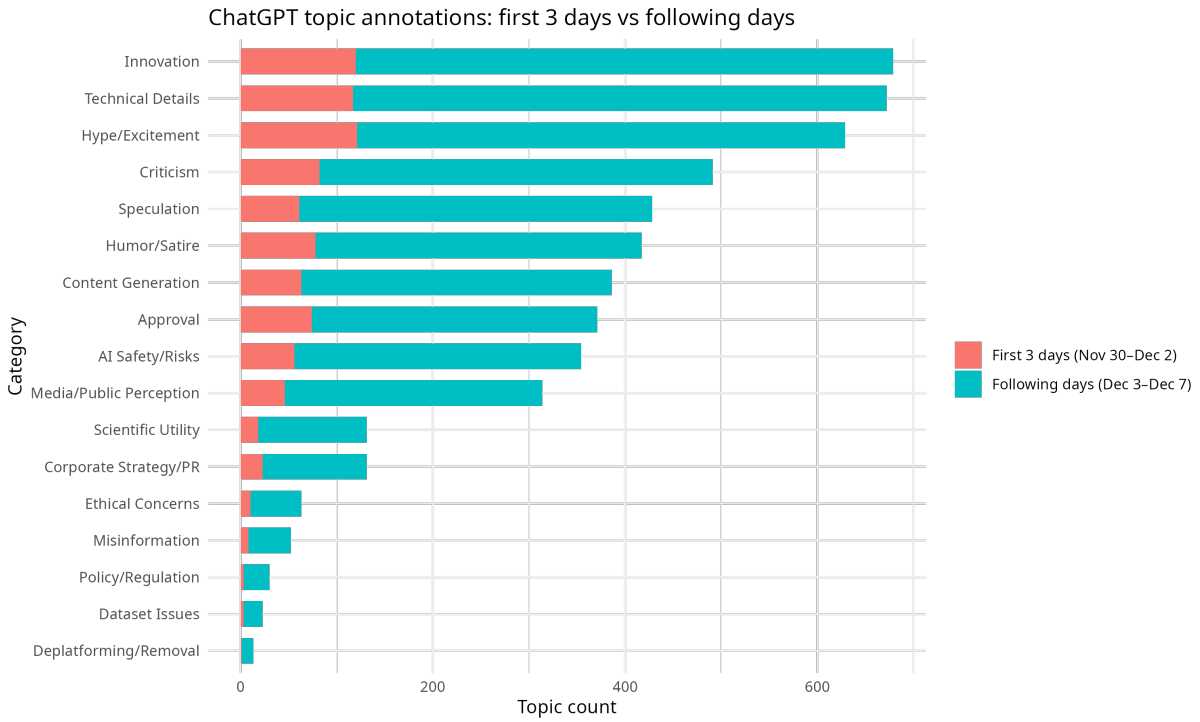


Figure 4: Topic distribution for ChatGPT tweets during the first eight days following release. Topics are derived from zero-shot topic classification using Llama 3.3 (70B) applied to a random subsample of 1,640 tweets.

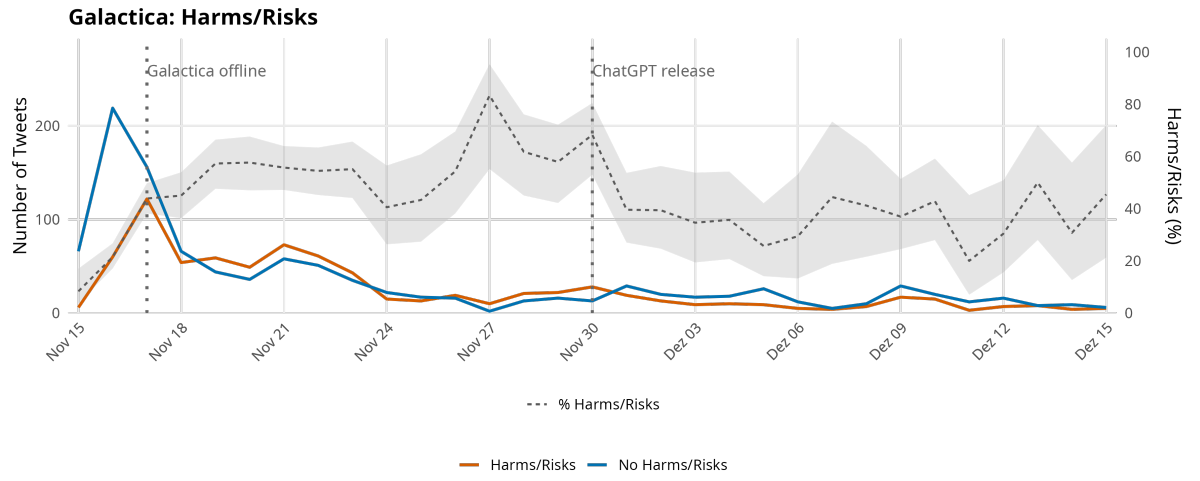


Figure 5: Temporal dynamics of potential harm/risk predictions for Galactica tweets. Lines show daily counts of predicted harms/risks and no harms/risks; the dashed line indicates the percentage of tweets classified as harms/risks, with shaded band representing the 95% Wilson confidence interval.

When Do LLMs Need Human Experts? Evidence for Social Science from Jurisprudential Classification

Caroline Cheng [1] **Edward H. Stiglitz** [2] **David Mimno** [1] **Matthew Wilkens** [1]
[1] Cornell University [2] Cornell Law School
{cyc59, js2758, mimno, wilkens}@cornell.edu

Abstract

Social scientists increasingly use large language models (LLMs) to classify text at scale, raising a key question: when can LLMs replace expert human annotation? Prior work found that earlier generative models failed on complex social science tasks while fine-tuned BERT succeeded, but whether current frontier-scale models close this gap remained untested. We investigate this question on a challenging legal reasoning task—classifying paragraphs from U.S. Supreme Court opinions as employing formal, grand, or no reasoning. Testing frontier LLMs including GPT-5.2 and leading open-weight alternatives, we find that even the most capable prompted models consistently underperform fine-tuned BERT. Only when high-parameter-count generative LLMs are fine-tuned on human-annotated training data does performance improve, and fine-tuned BERT remains a cost-effective alternative. Contrary to a common view, our results demonstrate that scaling to frontier-size LLMs does not eliminate the need for expert annotation on tasks requiring deep domain expertise—a finding with important implications for computational social science measurement.

1 Introduction

Social scientists increasingly rely on LLMs to classify text at scale, replacing expensive human coding (Ziems et al., 2024). But for tasks requiring deep domain expertise, can LLMs replace expert annotators? This question has direct implications for measurement validity: if LLM classifications diverge from expert judgments on complex constructs, downstream social science inferences may be biased (Egami et al., 2023).

Recent work finds that zero-shot LLM performance on social science coding tasks can be remarkably low, and that supervised fine-tuning on human-labeled data substantially improves smaller open-weight models (Halterman and Keith, 2025). However, these studies tested early or small models; whether frontier-scale LLMs close the gap with fine-tuned domain models is untested.

We conduct an up-to-date survey of the performance of LLMs against a challenging legal reasoning benchmark established in Thalken et al. (2023): classifying United States Supreme Court opinions as employing "formal" reasoning, "grand" reasoning, or no legal reasoning.[1] This task requires understanding of jurisprudential philosophy, rather than merely common or surface-level legal knowledge, making it a strong test for whether LLMs can match expert human annotators on domain-specific measurement tasks common to social science research. Similar to Halterman and Keith (2025), Thalken et al. (2023) found that generative LLMs perform poorly on their task without human annotation and instead observed strong performance with lightweight fine-tuned, in-domain BERT models.

Yet results as in Thalken et al. (2023) must be viewed as a snapshot, and a long-standing view is that models will outpace humans through scaling laws and greater application of compute (Kaplan et al., 2020). We now use this difficult legal benchmark to re-evaluate the question of when LLMs need human experts using more advanced frontier models, including OpenAI's most recent "reasoning" model.

We find that: (1) more recent in-domain fine-tuned models perform comparably to the original in-domain BERT models on this task; (2) prompted SOTA in-context LLMs continue to under-perform fine-tuned BERT models; (3) only fine-tuned SOTA generative LLMs—trained on human annotated samples—surpass the BERT baseline, with GPT-4.1 achieving the strongest performance.

Our findings illustrate that in a highly complex and specialized domain, scaling to frontier-size rea-

[1]The data in Thalken et al. (2023) later became the foundation of several social science papers, including Stiglitz and Thalken (2026).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 103–112

soning LLMs does not obviate the need for human annotation, and fine-tuned BERT models remain a competitive and cost-effective alternative for social scientists.[2]

2 Related Work

LLMs for Social Science Measurement. A growing body of work examines whether LLMs can serve as reliable measurement tools for social science. Ziems et al. (2024) find that LLMs generally fail to outperform fine-tuned models on CSS classification tasks. Halterman and Keith (2025) report zero-shot LLM F_1 as low as 0.21 on political science tasks, and show that supervised fine-tuning on human-labeled data substantially improves 7–12B parameter open-weight models. They do not compare against BERT-like encoder baselines or examine frontier reasoning models. Egami et al. (2023) show that modestly inaccurate LLM labels can produce biased downstream statistical inference without correction. Chae and Davidson (2025) find that fine-tuning smaller models is competitive with zero-shot large models. Pangakis and Wolken (2025) test GPT-4 on a variety of CSS tasks and conclude that human annotation is essential. However, on social media classification, Törnberg (2025) find that GPT-4 outperforms expert human coders and supervised classifiers; see also Gilardi et al. (2023). Combined, LLM performance may turn heavily on the degree to which the task requires expert domain knowledge.

Legal NLP and This Task. There has been significant recent progress in legal NLP (Siino et al., 2025), and LLMs show strong performance on both conventional human benchmarks, such as bar exams, as well as curated-task benchmarks (Guha et al., 2023). Classifying legal philosophy, however, requires deeper expertise than many legal tasks; even the human experts in Thalken et al. (2023), which established the benchmark, disagreed sometimes on the correct classification. They found that prompt-based generative models, including GPT-4, were out-performed by lightweight fine-tuned BERT models (Thalken et al., 2023).

Benchmark Validity. On published legal benchmarks such as LegalBench[3] and BigLaw Bench,[4] SOTA generative LLMs approach perfect performance. However, with the capacity for pre-training LLMs unclear, there is concern that performance is increasing due to training on the benchmarks (Ni et al., 2025; Li and Flanigan, 2024). We evaluate on a recent, specialized, expert-annotated dataset less likely to have been included in pre-training corpora.

3 LLMs on Jurisprudential Classification

Dataset. Our dataset, established by Thalken et al. (2023), consists of paragraphs from U.S. Supreme Court opinions annotated by domain experts as containing formal reasoning, grand reasoning, or neither.[5] These categories derive from jurisprudential philosophy (Llewellyn, 1960), requiring annotators to distinguish between modes of legal reasoning rather than surface legal features. Inter-annotator agreement measured by Krippendorff's α reached 0.65 in annotation sessions, reflecting the genuine difficulty of this classification task; the original annotation procedure used an iteratively-developed codebook, a decision chart that systematically improved agreement (Thalken et al., 2023, Figs. 1 and 4). This moderate agreement establishes a human performance ceiling that contextualizes model results. The annotated corpus contains 2,748 paragraphs (329 formal, 551 grand, 1,869 none), with seeds used to sample paragraphs and "none" intentionally oversampled to account for the heterogeneity of paragraphs not engaged in legal reasoning (Thalken et al., 2023).

Several pieces of external evidence support data and construct validity: predictions from a model trained on labels for these randomly selected paragraphs recover the consensus historical periodization in legal scholarship (Thalken et al., 2023; Stiglitz and Thalken, 2026, 2024); the justice-level aggregations of predictions reflect common views about justices' jurisprudence (Stiglitz and Thalken, 2026, 2024); predictions of jurisprudence correlate with the partisanship of the authoring justice in expected ways (Thalken and Stiglitz, 2026); famous historical episodes of changes in jurisprudence of justices show up in the predictions (Stiglitz and Thalken, 2024).

[2] Code is available at: https://github.com/caroline-y-cheng/llms-legal-reasoning. We do not test retrieval-augmented generation, large-scale few-shot regimes, or agentic tool-use approaches, any of which might, in the right configuration, narrow the gap. Cost, latency, and data-governance constraints common in applied social science motivate our focus on lightweight prompting and small-to-mid-scale fine-tuning.

[3] https://www.vals.ai/benchmarks/legal_bench

[4] https://www.harvey.ai/blog/gemini-3-pro-public-preview-early-access-evaluation-results

[5] See Thalken et al. (2023) for full dataset description.

We test LLMs on the task of distinguishing types of legal reasoning between formal and grand (or none) classes.[6] We compare performance between sets of in-domain fine-tuned models, a new set of prompted generative LLMs, and a set of supervised fine-tuned generative LLMs. We chose models based on performance on other legal benchmarks and accessibility (i.e., open-weight models and model size), being cognizant of resources needed for applied researchers to fine-tune and perform inference with extremely large and expensive models. In-domain fine-tuned models include LEGAL-BERT (Chalkidis et al., 2020),[7] LEGAL-RoBERTa,[8] and LEGAL-ModernBERT.[9] Generative LLMs prompted to identify legal reasoning include GPT-OSS (120B) (OpenAI, 2025a), GPT-4.1 (OpenAI, 2023), GPT-5.2 (OpenAI, 2025b), Llama-3.1-Instruct (8B and 70B) (Grattafiori et al., 2024), Qwen3-Instruct-2507 (4B), Qwen3-A3B-Instruct-2507 (30B), Qwen3-Next-A3B-Instruct (80B) (Yang et al., 2025), and DeepSeek-V3.1 (DeepSeek-AI, 2025). A subset of these LLMs was fine-tuned.

We created five stratified splits of the annotated data with 80% of the data in the training set and 20% of the data in the test segment. The in-domain fine-tuned and generative in-context models were evaluated over five splits, while the fine-tuned generative models were evaluated over one split due to model size and cost.

3.1 In-Domain BERT Models

For the task of identifying *types* of legal reasoning in text, the strongest results in Thalken et al. (2023) derived from the procedure of fine-tuned multi-class classification based on hand-labeled annotations. BERT can be fine-tuned on the GPU of a reasonably equipped recent laptop. Following this approach, we again observe similar performance among LEGAL-BERT, LEGAL-RoBERTa, and LEGAL-ModernBERT (Table 1).

[6]We also examine performance on a simpler binary task: classifying whether the passage engages in legal reasoning (of any form, either formal or grand) or not. The results from this exercise tend to support those from the more difficult multi-class problem. We report the results from this binary task in Table B.2.

[7]In Thalken et al. (2023), the strongest results derived from fine-tuning LEGAL-BERT on the annotated dataset.

[8]https://huggingface.co/Saibo-creator/legal-roberta-base

[9]https://free.law/2025/03/11/semantic-search/

3.2 Prompted Generative LLMs

To establish a baseline for the more recent generative LLMs, we test their performance when simply given instructions equal to those presented to our human annotators. In our multi-class classification task, we explore three prompting strategies: *descriptions* of each legal-reasoning class (zero-shot); *examples*, a 3-shot prompt with one canonical paragraph per class drawn from the codebook (Appendix D); and *chain-of-thought* (CoT), which walks through the annotation decision chart (Appendix C). The three in-context examples are fixed across test items. Full prompts appear in Appendix A.

3.3 Supervised Fine-tuned Generative LLMs

We determined a subset of the new set of generative LLMs to fine-tune based on their prompt-based performance and accessibility. GPT-4.1 is OpenAI's most powerful proprietary model that can be fine-tuned via the OpenAI API. Qwen3-A3B-Instruct-2507 (30B) resulted in higher macro-averaged F_1 scores than Qwen3-Next-A3B-Instruct (80B) in the multi-class task under each of the prompting strategies (Table B.1; Table 1) and was more feasible to fine-tune. DeepSeek-V3.1 was consistently outperformed by other more feasibly trained models on the baseline measurements (Table 1).

For fine-tuning, we prepared one split of the annotated dataset as prompt-completion examples:

- *Prompt*: Each of the three multi-class prompting strategies (Appendix A) followed by a paragraph to classify.
- *Completion*: Domain-expert classification of the paragraph.

We fine-tuned GPT-4.1 with the OpenAI API[10] with OpenAI autoconfigs. The Llama (8B and 70B) and Qwen (4B and 30B) models were fine-tuned using 4-bit quantized models and parameter-efficient techniques (Dettmers et al., 2023). They were fine-tuned in 200 steps, with a 10% warm-up ratio, a maximum learning rate of 3e-5, with a weight decay of 0.01.

4 LLM Performance

First, we show that more recent in-domain fine-tuned BERT models have comparable performance differences on this task to the older LEGAL-BERT model (Table 1).

[10]https://openai.com/api/

Model	Strategy	Macro F1
Fine-Tuned BERT (5-split avg.)		
LEGAL-BERT	–	0.71
LEGAL-RoBERTa	–	0.71
LEGAL-ModernBERT	–	0.70
Prompted (5-split avg., best strategy)		
GPT-5.2	Examples	0.68
GPT-4.1	Examples	0.64
GPT-OSS (120B)	Examples	0.59
Qwen3-A3B (30B)	Examples	0.53
DeepSeek-V3.1	CoT	0.52
Qwen3-Next-A3B (80B)	Examples	0.48
Llama-3.1 (8B)	CoT	0.43
Qwen3 (4B)	Examples	0.40
Llama-3.1 (70B)	Desc.	0.36
Fine-Tuned Generative (single split, best strategy)[†]		
GPT-4.1	CoT	**0.79**
Llama-3.1 (70B)	CoT	0.76
Qwen3-A3B (30B)	CoT	0.70
Qwen3 (4B)	Examples	0.69
Llama-3.1 (8B)	CoT	0.65

Table 1: Multi-class macro F_1 summary. Best prompting strategy shown per model. [†]Single-split results; see Limitations. Full per-class results in Appendix Table B.1.

We then tested the performance of the new set of LLMs with various prompting strategies on the same splits of data and find that without fine-tuning, they all perform worse than the in-domain fine-tuned BERT models.[11] On our multi-class task, the best performing prompted LLM is GPT-5.2 when given examples of the legal reasoning classes, whose macro F_1 score was 0.68 compared to LEGAL-BERT's 0.71. Llama-3.1-Instruct (8B and 70B) without task-specific fine-tuning very rarely predicted the "none" class (Table 1).[12] Our results suggest that without fine-tuning, generative LLMs continue to fall short of alignment with human annotation on highly complex and specialized classification tasks.

The fine-tuned LLMs all performed better than their non-fine-tuned versions within the same prompting strategy. However, SFT gains are scale-dependent: fine-tuned 8B and 4B models still underperform BERT, while only GPT-4.1 and 70B-class models surpass it on the evaluated split. On the multi-class classification task, fine-tuned GPT-4.1 consistently outperformed the in-domain fine-tuned models: the macro F_1 for GPT-4.1 with CoT prompt-completion examples is 0.79, 0.08 higher than the best macro F_1 for the in-domain fine-tuned models (Table 1; Table B.1). Fine-tuned Llama-3.1-Instruct (70B) and Qwen3-A3B-Instruct-2507 (30B) sometimes outperformed the in-domain fine-tuned models (Table B.1). Because notable performance gains require fine-tuned GPT-4.1 or larger open-weight LLMs trained on expert-annotated data, fine-tuned BERT remains a competitive and substantially more cost-effective alternative.

Per-class results (Appendix Table B.1) reveal that the models have the most trouble with the Formal and Grand classes; the None class is relatively easy for the models to detect. This may be because the reasoning classes reflect specialized domain-specific usage of common terms (Schauer, 1987). Even where aggregate scores converge, error patterns differ: GPT-5.2 matches BERT's Formal F_1 (0.56) but with lower precision (0.48 vs. 0.57) and higher recall (0.70 vs. 0.56), indicating it over-predicts formal reasoning, likely triggered by surface legal language rather than the underlying concept and method of reasoning. For CSS measurement, such systematic classification errors could bias downstream inference (Egami et al., 2023). Fine-tuning produces the largest gains on these difficult legal reasoning classes: fine-tuned GPT-4.1 CoT improves Formal F_1 to 0.68 (+0.16 over best GPT-4.1 without FT), Grand to 0.77 (+0.20), and None to 0.92 (+0.08), suggesting expert-annotated training data is most valuable where the classification requires the deepest domain knowledge. The models continue to make errors with fine-tuning, but at lower rates, and the errors tend to be more balanced between precision and recall, reducing systematic over- or under-prediction of any single class.

5 Conclusion

Extending Thalken et al. (2023) with frontier-scale models, we find that scaling alone does not solve this task: even GPT-5.2, a cutting edge reasoning model, underperforms fine-tuned BERT when prompted, and only fine-tuned generative LLMs surpass the BERT baseline on the evaluated split. This contrasts with tasks where frontier LLMs match or exceed human coders (Törnberg, 2025), suggesting that the need for human annotation turns

[11] Our reported results employ a user-role and zero temperature for classification.

[12] Some models often return additional text beyond the class label; we extract the class label (if it occurs) from the text and use that as the label for evaluation. Also, just prompting the GPT-OSS and DeepSeek models often generate responses that do not contain a label in our support; we calculate the classification reports based only on the predictions within the set of true labels in our codebook.

on the depth of domain expertise a task requires. Fine-tuned BERT remains a competitive and substantially more cost-effective alternative to fine-tuned frontier LLMs on this task. For CSS researchers working on tasks requiring deep domain expertise, our findings add to growing evidence (Ziems et al., 2024; Halterman and Keith, 2025; Gu et al., 2025) that investing in high-quality human annotation yields greater returns than relying on increasingly capable models alone. We reveal limitations in SOTA LLMs' capabilities to align with human reasoning, emphasizing the continued importance of domain-expert annotation.

Limitations

The prompting strategies explored for LLM baseline performance and supervised fine-tuning followed those of Thalken et al. (2023), which are limited to the scope of the instructions provided to human annotators. We did not evaluate retrieval-augmented generation, dynamic few-shot example selection, self-consistency or majority-vote schemes, or multi-agent or tool-using setups. Any of these could in principle narrow or close the gap to fine-tuned models, and exploring how much expert supervision can be substituted by stronger inference-time procedures is a natural next step.

In addition, this up-to-date survey of LLMs was limited by model size and cost. We prioritize open-weight models, 4-bit quantized models, and models that can be fine-tuned on a single H100 GPU due to resource constraints. Better performance on this task may be possible with emerging models. For the same cost-based reasons, we limit our study to a single split of the fine-tuned generative models; comparative claims between fine-tuned generative and BERT models should be interpreted with this caveat. We also do not vary the labeled-data budget for the same reason of cost. Prior work documents that prompting can substitute for hundreds of fine-tuning examples on some classification tasks (Le Scao and Rush, 2021); scaling the labeled data budget would allow us to identify the threshold at which prompted frontier LLMs become competitive with fine-tuning. This would be a productive next-step in this analysis.

The Krippendorff $\alpha = 0.65$ inter-annotator agreement is moderate and bounds the absolute level of model–human alignment achievable on this task. This ceiling reflects the genuine difficulty of distinguishing modes of jurisprudential reasoning. Though comparisons across models are evaluated against the same labels, so team-specific annotation conventions cannot drive the relative rankings reported here, absolute F_1 values should be read against the ceiling of human agreement.

Our findings are based on the study of a single domain, jurisprudential philosophy in Supreme Court opinions. Other CSS domains may show different patterns.

Ethics Statement

No human subjects were involved in this study beyond the annotation described in Thalken et al. (2023). Our data is in the public domain.

Acknowledgments

Many thanks to Rosamond Thalken and the team for establishing the foundation for the continuation of this work and to Prof. Stiglitz, Prof. Mimno, and Prof. Wilkens for their guidance and support.

References

Youngjin Chae and Thomas Davidson. 2025. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

DeepSeek-AI. 2025. Deepseek-v3 technical report.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

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*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*, pages 1–17.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.

Karl N. Llewellyn. 1960. *The Common Law Tradition: Deciding Appeals*. W.S. Hein, Buffalo, NY.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2025. Training on the benchmark is not all you need. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24948–24956.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card.

OpenAI. 2025b. Openai gpt-5 system card.

Nicholas Pangakis and Samuel Wolken. 2025. Keeping humans in the loop: Human-centered automated annotation with generative AI. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1471–1492.

Frederick Schauer. 1987. Formalism. *Yale Lj*, 97:509.

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*, 13:18253–18276.

Edward H Stiglitz and Rosamond Thalken. 2024. Historical trends in macro-jurisprudence: A language model assessment, 1870-2023. *Md. L. Rev.*, 84:46.

Edward H. Stiglitz and Rosamond Thalken. 2026. Undertanding change in jurisprudence. *Journal of Law, Economics, and Organization*.

Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: LM annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265, Singapore. Association for Computational Linguistics.

Rosamond Thalken and Edward H Stiglitz. 2026. Measuring jurisprudence. *Journal of Law and Courts*, pages 1–22.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*.

An Yang et al. 2025. Qwen3 technical report.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Appendices

A Prompts

Our approaches to prompting LLMs to identify legal reasoning and legal reasoning classes in text:

- *In Context, Descriptions*: An in-context prompt that provides the model with descriptions of the legal reasoning classes before asking for inference on a new paragraph.
- *In Context, Examples*: An in-context prompt that provides the model with one example of each of the three legal reasoning classes before asking for inference on a new paragraph.
- *Chain-of-Thought*: A CoT prompt that provides the model with steps of reasoning to follow in order to determine the class of legal reasoning before asking for inference on a new paragraph.

These prompts are included in Figure A.1. Each prompting strategy is derived from our codebook (Appendix D) or decision chart (Appendix C).

B Full Results

Table B.1 contains the full results for the multi-class task (formal reasoning versus grand reasoning versus no interpretation). Table B.2 contains the results for the simpler binary task (interpretation versus no interpretation).

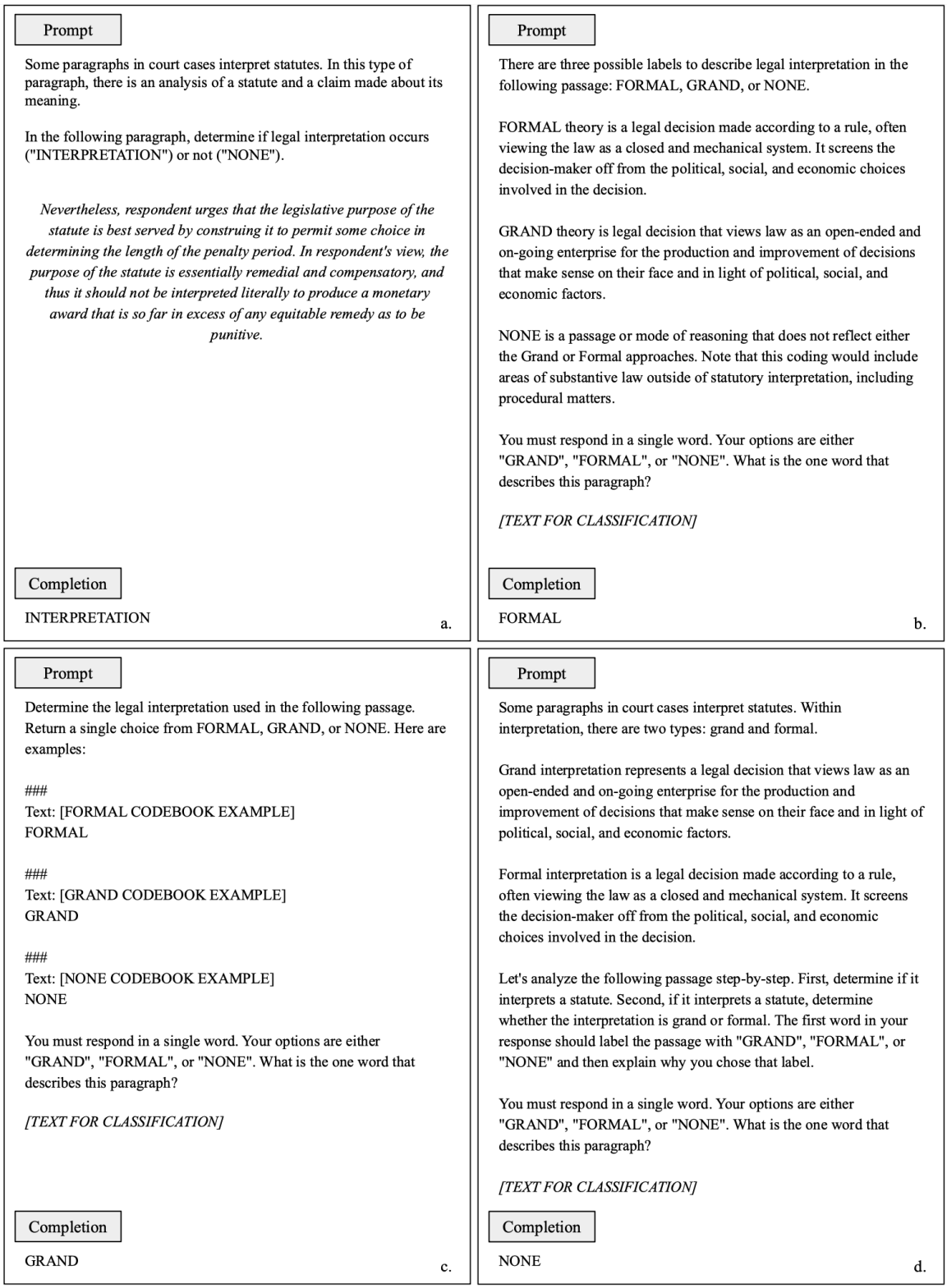


Figure A.1: Prompt a is the prompt used for identifying whether legal interpretation occurs or not. Prompt b is the prompt used for description classification of the classes of legal interpretation. Prompt c is the prompt used for few-shot classification of the classes of legal interpretation. Prompt d is the prompt used for CoT reasoning and the classes of legal interpretation.

C Decision Chart

Figure C.1 presents the decision chart provided to the annotation team. It was the basis of the CoT

Model	*Macro* F1	P	R	*Grand* F1	P	R	*Formal* F1	P	R	*None* F1	P	R
Multi-Class												
LEGAL-BERT	0.71	0.71	0.72	0.69	0.67	0.71	0.56	0.57	0.56	0.88	0.89	0.88
LEGAL-RoBERTa	0.71	0.70	0.72	0.69	0.65	0.73	0.56	0.55	0.58	0.88	0.90	0.86
LEGAL-ModernBERT	0.70	0.72	0.68	0.67	0.70	0.65	0.53	0.58	0.49	0.89	0.87	0.91
In-Context, Descriptions												
GPT-OSS (120B)	0.56	0.57	0.67	0.57	0.46	0.74	0.44	0.31	0.74	0.69	0.93	0.54
GPT-4.1	0.46	0.51	0.58	0.41	0.38	0.44	0.37	0.23	0.87	0.59	0.92	0.43
GPT-5.2	0.49	0.54	0.62	0.43	0.41	0.44	0.40	0.26	0.90	0.66	0.94	0.51
Llama-3.1-Instruct (8B)	0.38	0.43	0.48	0.21	0.28	0.16	0.32	0.20	0.82	0.60	0.82	0.47
Llama-3.1-Instruct (70B)	0.36	0.45	0.50	0.24	0.29	0.21	0.33	0.20	0.92	0.53	0.87	0.38
Qwen3-Instruct-2507 (4B)	0.38	0.46	0.51	0.39	0.27	0.72	0.36	0.27	0.56	0.39	0.85	0.26
Qwen3-A3B-Instruct-2507 (30B)	0.50	0.50	0.58	0.43	0.35	0.55	0.42	0.31	0.67	0.65	0.84	0.53
Qwen3-Next-A3B-Instruct (80B)	0.40	0.50	0.54	0.34	0.36	0.33	0.33	0.20	0.90	0.53	0.93	0.37
DeepSeek-V3.1	0.35	0.48	0.49	0.23	0.36	0.17	0.30	0.18	**0.95**	0.51	0.90	0.36
In-Context, Examples												
GPT-OSS (120B)	0.59	0.59	0.70	0.59	0.45	0.84	0.48	0.36	0.71	0.69	**0.95**	0.55
GPT-4.1	0.64	0.68	0.63	0.57	0.52	0.62	0.52	0.68	0.42	0.84	0.83	0.84
GPT-5.2	0.68	0.66	0.71	0.63	0.61	0.65	0.56	0.48	0.70	0.84	0.88	0.79
Llama-3.1-Instruct (8B)	0.17	0.40	0.30	0.23	0.14	0.58	0.24	0.20	0.30	0.05	0.86	0.02
Llama-3.1-Instruct (70B)	0.18	0.45	0.40	0.32	0.42	0.26	0.22	0.13	0.92	0.00	0.80	0.00
Qwen3-Instruct-2507 (4B)	0.40	0.51	0.50	0.42	0.27	**0.90**	0.32	0.36	0.28	0.48	0.90	0.33
Qwen3-A3B-Instruct-2507 (30B)	0.53	0.55	0.52	0.42	0.45	0.39	0.38	0.44	0.33	0.79	0.76	0.82
Qwen3-Next-A3B-Instruct (80B)	0.48	0.53	0.60	0.51	0.46	0.56	0.35	0.23	0.80	0.59	0.91	0.44
DeepSeek-V3.1	0.50	0.62	0.48	0.24	0.53	0.15	0.44	0.58	0.36	0.82	0.73	0.93
Chain-of-Thought												
GPT-OSS (120B)	0.56	0.56	0.66	0.56	0.48	0.68	0.42	0.30	0.71	0.70	0.91	0.58
GPT-4.1	0.51	0.52	0.59	0.43	0.39	0.47	0.41	0.28	0.75	0.69	0.88	0.56
GPT-5.2	0.49	0.55	0.60	0.31	0.47	0.23	0.40	0.26	0.91	0.77	0.93	0.65
Llama-3.1-Instruct (8B)	0.43	0.44	0.46	0.36	0.27	0.51	0.25	0.22	0.30	0.67	0.84	0.56
Llama-3.1-Instruct (70B)	0.33	0.47	0.48	0.24	0.31	0.20	0.30	0.18	0.94	0.46	0.91	0.31
Qwen3-Instruct-2507 (4B)	0.34	0.45	0.48	0.36	0.26	0.61	0.31	0.21	0.62	0.36	0.90	0.22
Qwen3-A3B-Instruct-2507 (30B)	0.42	0.49	0.54	0.44	0.32	0.69	0.33	0.23	0.62	0.48	0.92	0.33
Qwen3-Next-A3B-Instruct (80B)	0.41	0.49	0.53	0.40	0.33	0.52	0.32	0.21	0.74	0.50	0.92	0.34
DeepSeek-V3.1	0.52	0.56	0.59	0.28	0.50	0.19	0.49	0.36	0.79	0.80	0.82	0.78
Fine-Tuned, Descriptions												
GPT-4.1	0.73	0.74	0.74	0.75	0.71	0.80	0.55	0.59	0.51	0.90	0.91	0.90
Llama-3.1-Instruct (8B)	0.61	0.71	0.59	0.61	0.58	0.64	0.35	0.71	0.23	0.87	0.84	0.91
Llama-3.1-Instruct (70B)	0.66	0.81	0.60	0.60	0.79	0.49	0.48	**0.85**	0.34	0.88	0.80	**0.97**
Qwen3-Instruct-2507 (4B)	0.64	0.66	0.62	0.59	0.61	0.56	0.49	0.54	0.45	0.85	0.82	0.87
Qwen3-A3B-Instruct-2507 (30B)	0.68	0.72	0.66	0.70	0.71	0.69	0.48	0.59	0.40	0.88	0.85	0.91
Fine-Tuned, Examples												
GPT-4.1	0.76	0.78	0.75	0.76	0.75	0.77	0.62	0.69	0.55	0.91	0.90	0.93
Llama-3.1-Instruct (8B)	0.59	0.74	0.55	0.55	0.68	0.46	0.37	0.76	0.25	0.86	0.78	0.95
Llama-3.1-Instruct (70B)	0.74	0.76	0.72	0.75	0.75	0.74	0.57	0.65	0.51	0.90	0.88	0.92
Qwen3-Instruct-2507 (4B)	0.69	0.71	0.67	0.67	0.67	0.67	0.54	0.63	0.48	0.86	0.84	0.87
Qwen3-A3B-Instruct-2507 (30B)	0.73	0.77	0.70	0.70	0.74	0.68	0.60	0.72	0.51	0.88	0.85	0.92
Fine-Tuned, Chain-of-Thought												
GPT-4.1	**0.79**	**0.81**	**0.77**	**0.77**	**0.79**	0.75	**0.68**	0.73	0.63	**0.92**	0.90	0.94
Llama-3.1-Instruct (8B)	0.65	0.69	0.62	0.59	0.71	0.51	0.47	0.54	0.42	0.88	0.83	0.93
Llama-3.1-Instruct (70B)	0.76	0.77	0.76	0.75	0.76	0.74	0.64	0.65	0.63	0.90	0.90	0.91
Qwen3-Instruct-2507 (4B)	0.67	0.70	0.65	0.66	0.68	0.64	0.49	0.60	0.42	0.87	0.84	0.90
Qwen3-A3B-Instruct-2507 (30B)	0.70	0.73	0.68	0.68	0.71	0.65	0.56	0.64	0.49	0.88	0.85	0.91

Table B.1: Model performance averaged over five train test splits for fine-tuned and generative in-context models. Model performance on 1 train test split for fine-tuned generative models. Macro averages represent averages unweighted by class.

prompt for generative LLMs.

D Codebook

Table D.1 presents the codebook with definitions and core examples of each class, which guided annotators and was the basis of the in-context descriptions and in-context examples prompts for generative LLMs.

Model	*Macro* F1	P	R	*Interpretation* F1	P	R	*None* F1	P	R
Fine-Tuned									
LEGAL-BERT	0.82	0.81	0.83	0.76	0.72	0.80	0.88	0.90	0.86
LEGAL-RoBERTa	0.82	0.82	0.83	0.77	0.73	0.80	0.88	0.90	0.86
LEGAL-ModernBERT	0.82	0.82	0.81	0.75	0.77	0.73	0.89	0.88	0.90
Generative In-Context									
GPT-OSS (120B)	0.66	0.69	0.71	0.62	0.49	0.85	0.70	0.89	0.57
GPT-4.1	0.68	0.69	0.72	0.63	0.51	0.80	0.74	0.87	0.64
GPT-5.2	0.71	0.71	0.74	0.65	0.54	0.80	0.77	0.88	0.68
Llama-3.1-Instruct (8B)	0.60	0.64	0.65	0.56	0.44	0.78	0.64	0.84	0.53
Llama-3.1-Instruct (70B)	0.69	0.70	0.73	0.63	0.53	0.79	0.75	0.87	0.66
Qwen3-Instruct-2507 (4B)	0.58	0.65	0.65	0.57	0.42	0.86	0.59	0.87	0.45
Qwen3-A3B-Instruct-2507 (30B)	0.54	0.67	0.65	0.57	0.41	**0.94**	0.52	**0.93**	0.36
Qwen3-Next-A3B-Instruct (80B)	0.62	0.67	0.69	0.59	0.46	0.85	0.66	0.88	0.52
DeepSeek-V3.1	0.64	0.67	0.69	0.59	0.47	0.81	0.68	0.86	0.56
Generative Fine-Tuned									
GPT-4.1	**0.86**	**0.86**	**0.86**	**0.81**	0.82	0.80	**0.91**	0.91	0.92
Llama-3.1-Instruct (8B)	0.78	0.80	0.76	0.68	0.75	0.63	0.87	0.84	0.90
Llama-3.1-Instruct (70B)	0.78	0.83	0.76	0.68	**0.85**	0.56	0.88	0.82	**0.95**
Qwen3-Instruct-2507 (4B)	0.79	0.79	0.79	0.72	0.70	0.73	0.86	0.87	0.85
Qwen3-A3B-Instruct-2507 (30B)	0.81	0.81	0.81	0.75	0.75	0.74	0.88	0.88	0.88

Table B.2: Model performance for binary interpretation averaged over 5 train test splits for fine-tuned and generative in-context models. Model performance for binary interpretation on 1 train test split for fine-tuned generative models. Macro averages represent averages unweighted by class.

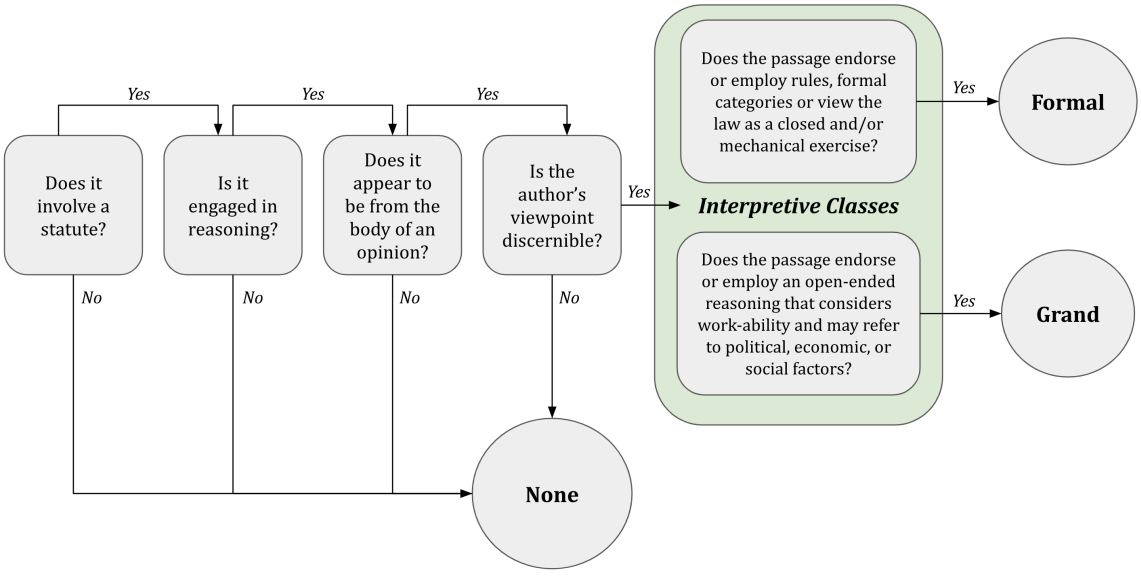


Figure C.1: The decision chart provided to annotators.

Class	Definition	Example
Formal	A legal decision made according to a rule, often viewing the law as a closed and mechanical system. It screens the decision-maker off from the political, social, and economic choices involved in the decision.	Accepting this point, too, for argument's sake, the question becomes: What did "discriminate" mean in 1964? As it turns out, it meant then roughly what it means today: "To make a difference in treatment or favor (of one as compared with others)." Webster's New International Dictionary 745 (2d ed. 1954). To "discriminate against" a person, then, would seem to mean treating that individual worse than others who are similarly situated. [CITE]. In so-called "disparate treatment" cases like today's, this Court has also held that the difference in treatment based on sex must be intentional. See, e.g., [CITE]. So, taken together, an employer who intentionally treats a person worse because of sex—such as by firing the person for actions or attributes it would tolerate in an individual of another sex—discriminates against that person in violation of Title VII. Bostock v. Clayton County
Grand	A legal decision that views the law as an open-ended and ongoing enterprise for the production and improvement of decisions that make sense on their face and in light of political, social, and economic factors.	Respondent's argument is not without force. But it overlooks the significance of the fact that the Kaiser-USWA plan is an affirmative action plan voluntarily adopted by private parties to eliminate traditional patterns of racial segregation. In this context respondent's reliance upon a literal construction of §§703 (a) and (d) and upon McDonald is misplaced. See [CITE]. It is a "familiar rule, that a thing may be within the letter of the statute and yet not within the statute, because not within its spirit, nor within the intention of its makers." [CITE]. The prohibition against racial discrimination in §§703 (a) and (d) of Title VII must therefore be read against the background of the legislative history of Title VII and the historical context from which the Act arose. See [CITE]. Examination of those sources makes clear that an interpretation of the sections that forbade all race-conscious affirmative action would "bring about an end completely at variance with the purpose of the statute" and must be rejected. [CITE]. See [CITE]. Steelworkers v. Weber
None	A passage or mode of reasoning that does not reflect either the Grand or Formal approaches. Note that this coding would include areas of substantive law outside of statutory interpretation, including procedural matters.	The questions are, What is the form of an assignment, and how must it be evidenced? There is no precise form. It may be. by delivery. Briggs v. Dorr, CITE, citing numerous cases; Onion v. Paul, 1 Har. & Johns. 114; Dunn v. Snell, CITE; Titcomb v. Thomas, 5 Greenl. 282. True, it is said it must be on a valuable consideration, with intent to transfer it. But these last are requisites in all assignments, or transfers of securities, negotiable or not. It may be by writing under seal, by writing without seal, by oral declarations, accompanied in all cases by delivery, and on a just consideration. The evidence may be by proof of handwriting and proof of. possession. It may be proved by proving the signature of the payee or obligee on the back, and possession by a third person. 3 Gill & Johns. 218.

Table D.1: Codebook definition and examples of each of the interpretive classes.

An NLP Framework for Analyzing Corporate Strategic Behavior in the Opioid Industry Documents Archive

Duy Dang Phu and Dang Van Thin
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
24520010@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

The Opioid Industry Documents Archive (OIDA) provides extensive internal corporate records that offer valuable insight into the drivers of the opioid crisis, yet its use in systematic analysis of corporate strategy remains limited. In this study, we propose an NLP-based framework to analyze strategic behavior in large-scale litigation archives, combining relevance filtering and topic modeling with large language model (LLM)-assisted interpretation. Applied to documents from Insys Therapeutics and Mallinckrodt Pharmaceuticals, our approach uncovers systematic differences in corporate strategies and organizational priorities. These results highlight the potential of integrating representation learning and LLMs for large-scale analysis in public health and corporate accountability research.

1 Introduction

The opioid crisis constitutes a severe and ongoing public health emergency, characterized by widespread opioid dependence, high rates of overdose mortality, and profound social and economic consequences. In the wake of extensive litigation against pharmaceutical manufacturers, distributors, pharmacy chains, and consulting firms, millions of internal corporate documents have been made publicly available. These disclosures provide an unprecedented window into corporate decision-making processes and strategic conduct within the opioid industry.

The Opioid Industry Documents Archive (University of California, San Francisco and Johns Hopkins University, n.d.) is a digital repository developed by the University of California, San Francisco and Johns Hopkins University. It contains internal corporate documents, emails, and presentations produced by opioid manufacturers and related organizations. Given its scope and depth, OIDA represents a uniquely valuable data source for research in social science, public health, policy, and law—particularly for studies aimed at analyzing corporate strategic behavior. However, despite its substantial research potential, systematic large-scale computational analyses of corporate strategic behavior within OIDA remain limited. The unstructured nature of the textual data, combined with relatively sparse metadata, complicates comprehensive large-scale analysis.

To address these challenges, this study develops a multi-stage computational framework for analyzing corporate strategic behavior in OIDA. The framework integrates keyword-based retrieval, transformer-based text embeddings, and a K-Nearest Neighbors classifier to refine the selection of strategy-relevant texts. We then apply topic modeling, augmented by large language model –assisted interpretation, to identify and synthesize recurring forms of corporate strategic behavior. We focus on two organizations—Insys Therapeutics and Mallinckrodt Pharmaceuticals—by systematically characterizing and comparing their corporate strategic behavior during the same historical period. As a result, we present a computational framework that integrates large-scale text analysis with LLM-assisted interpretation to enable scalable and reproducible analysis of extensive litigation corpora. Substantively, our findings provide systematic empirical evidence of recurring corporate strategic behavior patterns in the opioid industry.

2 Related Work

Computational text analysis has been widely applied to large-scale corpus exploration, including topic modeling, semantic clustering, sentiment analysis, and named entity recognition. These approaches enable structured analysis of unstructured textual data across scientific, legal, and corporate domains.

Prior work has demonstrated the utility of multi-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 113–122

stage NLP pipelines for knowledge discovery. For example, (Polpinij et al., 2026) integrates transformer-based embeddings, deep clustering, and relation extraction to construct domain-specific knowledge representations from agricultural research literature. Similarly, (Azher et al., 2025) combines BERTopic and large language models to structure and summarize limitation sections in scientific articles. Beyond scientific corpora, NLP methods have been applied to corporate disclosures and financial documents. Studies such as (Kang and Kim, 2022) and (Faccia et al., 2024) employ sentiment analysis and semantic similarity measures to examine thematic emphasis, disclosure patterns, and linguistic risk signals in corporate reports. These works illustrate how computational methods can surface structured patterns within corporate communication.

Such approaches are particularly relevant for archives like the *Opioid Industry Documents Archive*, which contains internal corporate communications documenting strategic planning, marketing activities, and regulatory positioning. While OIDA has been introduced as a valuable research resource, computational engagement with the archive remains limited. Existing work, such as OIDA-QA (Shen et al., 2025), primarily focuses on benchmark construction. However, an integrated computational framework explicitly designed to identify and characterize corporate strategic behavior in large-scale litigation archives remains absent. This gap motivates the development of the structured NLP framework proposed in this study.

3 Data Description

The dataset is collected from the Opioid Industry Documents Archive, a public repository of internal documents related to the opioid industry. The archive contains various types of documents, including emails, reports, memoranda, and presentation slides. Each document is accompanied by metadata such as document ID, document type, date, author, and other descriptive attributes.

The primary source for textual analysis in this study is the *ocr_text* field, which contains text extracted from scanned original documents using Optical Character Recognition (OCR). This field serves as the main input for NLP analysis. However, due to the nature of OCR-based extraction, the text may contain structural inconsistencies, formatting irregularities, and noise (e.g., broken words, misrecognized characters, or misplaced line breaks). These issues introduce additional preprocessing challenges before conducting downstream NLP tasks.

4 Methodology

All detailed prompt templates, model configurations, and hyperparameters are provided in the Appendix B.

4.1 Data Selection and Preprocessing

This study draws on documents from the Opioid Industry Documents Archive. We selected presentation materials associated with two major opioid manufacturers, Insys Therapeutics and Mallinckrodt Pharmaceuticals, both of which have been centrally implicated in litigation and public investigations related to the U.S. opioid crisis. Focusing on these firms enables a concentrated examination of corporate strategic practices during periods of heightened commercial activity and regulatory scrutiny.

To reduce topical noise and focus on high-level strategic content, we restricted our analysis to presentation documents. Compared to emails or general reports, presentation slides are more likely to summarize strategic planning, business positioning, and key corporate initiatives, making them more suitable for downstream semantic analysis. After collecting the presentation files, we performed text cleaning on the *ocr_text* field to remove OCR-related artifacts and malformed characters. To facilitate embedding generation, each presentation was segmented into text chunks of 300 words with a 50-word overlap between consecutive segments. This chunking strategy was determined through preliminary experiments to balance contextual completeness and embedding quality.

4.2 Data Filtering

To enable topic modeling to focus on strategic-level content, we implemented a two-stage filtering procedure.

Stage 1: Keyword-based Pre-filtering. We first constructed a domain-informed keyword list capturing common corporate strategies. Only text segments containing at least one of these keywords were retained. This step serves two purposes: (1) reducing topical noise and (2) lowering computational costs for subsequent modeling stages.

Stage 2: LLM-assisted Labeling and Similarity-based Propagation. From each company, we randomly sampled 150 text segments and used a large language model with chain-of-thought (Wei et al., 2022) prompting to infer whether each segment was strategically relevant. For text representation, we employed the embedding model *avsolatorio/GIST-Embedding-v0* (Solatorio, 2024), selected to balance computational efficiency and semantic performance.

Using the labeled subset as supervision, we trained a k-nearest neighbors classifier ($k = 5$). The value of k was determined through empirical evaluation, with F1-score used as the primary selection metric. The trained KNN model was subsequently applied to the full dataset, where strategic relevance for each segment was determined via majority voting among its nearest neighbors in the embedding space.

4.3 Topic Modeling

We employed BERTopic (Grootendorst, 2022) to cluster semantically similar text segments into coherent and interpretable topics. To ensure methodological consistency across stages, the same embedding model used during the filtering phase (*avsolatorio/GIST-Embedding-v0*) was retained for topic modeling.

For topic representation, we applied the *KeyBERTInspired* approach to extract representative keywords for each topic. Prior to generating topic representations, additional preprocessing steps were conducted, including stopword removal. This procedure was intended to reduce lexical noise, improve semantic coherence, and enhance the interpretability of the resulting topics. Recent work by Yang et al. (2025) demonstrates that large language models (LLMs) can generate high-quality topical descriptors that align closely with human judgment in topic model evaluation. This finding suggests that LLMs possess substantial potential for extracting structured and meaningful information from document clusters. Motivated by this insight, we operationalized a two-stage LLM-assisted analytical framework by designing the following prompts:

Stage 1: Prompt for Topic-Level Strategy Classification This prompt was developed to determine whether a given topic explicitly reflects one of the predefined corporate strategic categories: *Sales Expansion & Promotion*, *Influence & Narrative Management*, *Regulatory Risk Management & Evasion*, or neither. The model was instructed to rely strictly on explicit textual evidence extracted from representative document excerpts and to provide a justification for its classification decision. This step serves as a filtering mechanism to exclude topics that do not substantively reflect corporate strategies or strategies, thereby improving the validity of subsequent interpretation.

Stage 2: Prompt for Topic Interpretation This prompt was applied exclusively to topics that had been classified as strategy-related in the preceding stage. Its primary objective was to generate a structured and analytically coherent interpretation of each topic, thereby facilitating clearer substantive analysis and supporting subsequent qualitative examination.

During the prompt design process, we observed that the inclusion of certain explicitly strategy-laden terms tended to induce speculative reasoning, leading the model to produce overextended or inferential claims not firmly grounded in the source material. To mitigate this tendency, the prompt was carefully calibrated to constrain interpretive latitude. Specifically, the model was instructed to frame the described actions as observable industry strategies without inferring hidden intentions unless such intentions were directly evidenced in the text. Furthermore, the prompt required that every substantive claim in the generated explanation be traceable to the provided textual excerpts. This evidentiary constraint was intended to reinforce analytical rigor, minimize unwarranted extrapolation, and ensure that interpretations remained firmly anchored in the documentary record.

5 Analysis

After applying the LLM to strategy-classified topics, we obtained the results summarized in Table 1. To better understand the substantive characteristics of the identified topics, we now turn to a company-level qualitative analysis. This allows us to examine how strategic patterns manifest differently across organizational contexts rather than relying solely on aggregate topic counts.

5.1 Insys

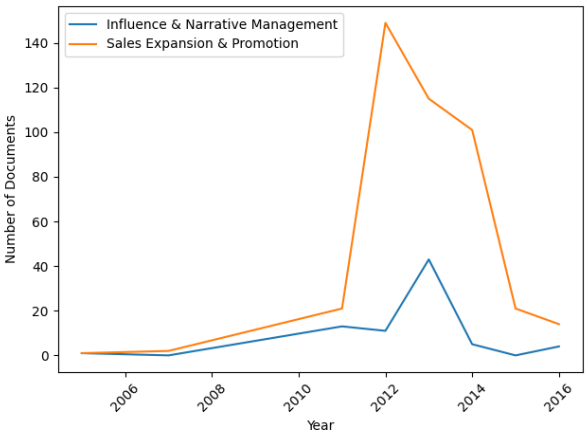
Figure 1 presents the yearly distribution of strategy-related text chunks identified for Insys. The results show that the majority of documents categorized under *Sales Expansion & Promotion* and *Influence & Narrative Management* are concentrated in the

Table 1: Strategy-related topics identified by the LLM. Sales Exp. = Sales Expansion & Promotion; Influence = Influence & Narrative Management; Reg. Risk = Regulatory Risk Management & Evasion; Total = total number of extracted topic groups.

Strategy	Insys	Mallinckrodt
Sales Exp.	24	9
Influence	5	1
Reg. Risk	0	5
Total	51	25

Note: A single topic may be assigned to more than one strategy category.

Figure 1: Number of Document per Strategy Type Over Years



2011–2015 period. Because most strategy-related materials fall within this time frame, our subsequent analysis focuses primarily on these years.

This temporal concentration overlaps with key milestones in the commercial trajectory of *Subsys*, which received approval from the U.S. Food and Drug Administration (FDA) in January 2012 for the treatment of breakthrough cancer pain in opioid-tolerant patients. Publicly available records report substantial revenue growth between 2012 and 2015, with annual revenues reaching approximately $330 million by 2015 (Opioid Industry Documents Archive, n.d.). However, it is important to note that the prominence of strategy-related content during 2011–2015 may also partly reflect the larger overall volume of available documents from this period. In other words, the higher frequency of identified strategies is not necessarily indicative of a proportional increase in strategic activity, but may be influenced by the greater density of archived materials.

To further examine the internal composition of these activities, we identified the five strategy-related topics with the highest number of text documents within the 2011–2015 period and conducted qualitative summaries using LLM-assisted abstraction of the corresponding documents. Although automated summarization may introduce minor semantic imprecision, manual inspection of a subset of documents suggests that the extracted summaries preserve their primary thematic content. Figure 2 displays the quarterly distribution of these five topics.

Substantively, the five topics capture complementary dimensions of commercialization strategy. Topic 3 reflects an integrated, multi-channel marketing framework combining awareness campaigns, targeted outreach, speaker engagement, and medical-education initiatives supported by internal staff and IT infrastructure, oriented toward expanding prescriber adoption and utilization intensity. Topic 10 centers on a structured speaker-program system in which promotional events are financially supported and monitored through prescription-linked performance metrics. Topic 11 describes organizational efforts to enhance sales-force efficiency by reallocating logistical responsibilities to specialized liaisons, thereby increasing representatives' focus on prescriber engagement and market expansion. Topics 15 and 30 both concern performance-based compensation structures, including region-specific sales targets, national adjustment factors, and layered bonus mechanisms designed to align managerial incentives with corporate revenue objectives.

The quarterly patterns suggest differentiated temporal dynamics across topics. Topic 3 appears consistently from early 2011 through late 2014, indicating that the activities captured under this topic were sustained over multiple years rather than confined to a short-term initiative. In contrast, the remaining topics emerge primarily after January 2012. Topics 10 and 11 become visible relatively early in the post-approval phase, particularly between 2012 and early 2013, while Topics 15 and 30 appear more prominently from late 2013 to mid-2014. These patterns may reflect shifts in strategic emphasis during Insys's peak commercialization period.

Taken together, the visualization suggests that most identifiable strategy-related documentation

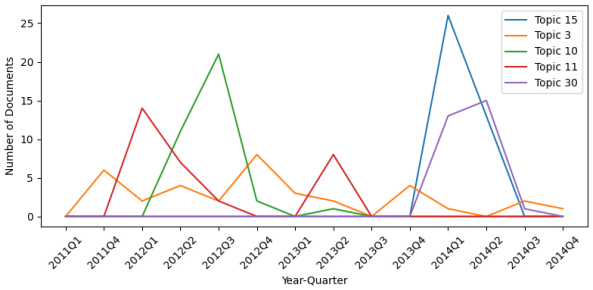


Figure 2: Number of Document per Topic Over Quarter-Years

clusters within 2011–2015, a period that also contains a relatively high volume of overall archival materials. While these observations remain descriptive and do not establish causality, they provide distributional evidence that may reflect evolving organizational priorities during a document-dense phase of product commercialization.

5.2 Mallinckrodt

Similar to the pattern observed for Insys, the majority of strategy-related documents identified in the Mallinckrodt corpus are concentrated in the 2010–2014 period. This temporal clustering spans the years immediately before and after a significant regulatory event: in 2011, Mallinckrodt came under investigation by the Drug Enforcement Administration (DEA) for failing to meet its obligations to monitor and report suspicious orders of controlled substances (Opioid Industry Documents Archive, n.d.). As with Insys, the concentration of identified strategies during this period partly reflects the relatively high volume of available documents from these years. Accordingly, our analysis focuses on 2010–2014, while acknowledging that distributional density may be influenced by archival coverage.

In contrast to the case of Insys, our analysis of Mallinckrodt's corpus reveals the presence of text chunks categorized under *Regulatory Risk Management & Evasion*. Although these documents do not constitute the largest share of strategy-related content, their emergence during a period of heightened regulatory scrutiny is analytically noteworthy. However, a closer qualitative inspection of these topics reveals that they primarily pertain to standard corporate legal and regulatory procedures, rather than explicit strategies for navigating or countering the concurrent DEA investigation. This pattern suggests that while the organization was actively managing its baseline regulatory compliance alongside commercial expansion, the automated extraction did not capture direct strategic responses to the DEA probe within this topic subset. Within the 2010–2014 window, we identified the five strategy-related topics with the highest document counts and conducted qualitative abstraction of their representative texts.

Substantively, the five topics capture complementary dimensions of Mallinckrodt's commercialization strategy during the focal period. Topic 1 reflects sustained efforts to expand prescription volume through sales-force growth, payer segmentation and rebate arrangements to secure formulary access, and the mobilization of key opinion leaders, collectively oriented toward increasing market penetration and revenue attainment. Topic 2 centers on coordinated launch campaigns that integrate sales training, defined performance metrics, cross-product promotion, expanded coverage, and patient-support mechanisms within a structured commercialization framework. Topic 11 describes a comprehensive pre-launch strategy that aligns marketing, medical affairs, and managed-markets functions around pricing, regulatory positioning, payer analytics, and unbranded market-development initiatives. Topic 10 emphasizes managed-care engagement and contract optimization, supported by systematic payer segmentation and financial modeling to balance coverage, profitability, and market share. Topic 15 concerns organizational and channel design decisions, including territory structuring, sales-force deployment, and evaluation of internal versus outsourced commercial models, thereby shaping the operational infrastructure of market entry and expansion.

The temporal distribution of these topics suggests differentiated strategic horizons. Topics 1 and 2 appear over relatively extended intervals, indicating sustained commercial initiatives. Topic 1 spans nearly the entire 2010–2014 period, while Topic 2 is prominent from early 2011 through approximately mid-period, suggesting an extended launch and expansion phase. In contrast, Topics 10, 11, and 15 exhibit more temporally concentrated patterns, consistent with time-bound planning or optimization initiatives.

Taken together, the quarterly distributions indicate a combination of long-term commercialization strategies and shorter-term strategic adjustments during a period characterized by both market expansion and increased regulatory scrutiny. While

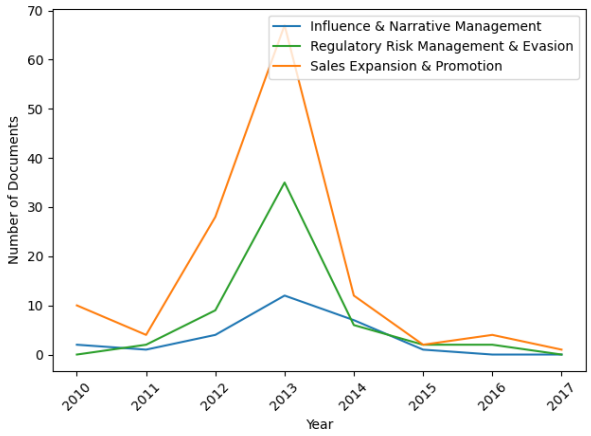


Figure 3: Number of Document per Strategy Type Over Years

these observations remain descriptive and do not establish causal relationships, they provide distributional evidence of evolving strategic emphases within Mallinckrodt's documented activities during 2010–2014.

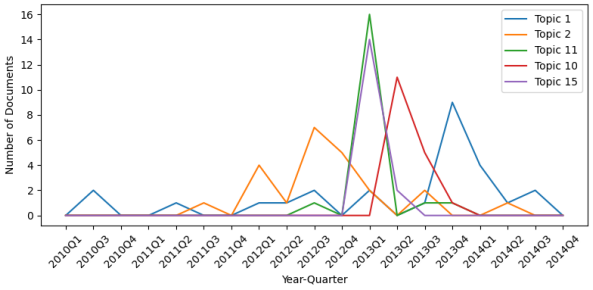


Figure 4: Number of Document per Topic Over Quarter-Years

5.3 Comparative Strategic Emphases

Synthesizing the topic distributions reveals a distinct divergence in the documented strategic priorities of the two organizations during their respective periods of peak commercialization. While both companies focused on market expansion, their operational emphases differed significantly along the commercialization value chain.

Insys's corpus reflects a localized, micro-level strategy heavily oriented toward direct prescriber engagement. Their strategic documentation indicates a highly proactive direct-to-prescriber outreach model, predominantly focusing on prescriber-level interventions. This is evidenced by the prominence of topics detailing speaker programs linked to sales volume and performance-based compensation structures designed to monitor prescription metrics. In essence, Insys's documented strategy during this period prioritized prescriber acquisition, engagement, and utilization intensity. Notably, these computationally derived patterns align closely with the historical record; subsequent federal investigations documented that Insys heavily relied on "speaker programs" to drive and reward prescription volumes (U.S. Department of Justice, 2019).

Conversely, Mallinckrodt demonstrated a macro-level strategy focused on institutional market access and commercial infrastructure. Rather than concentrating primarily on individual physicians, Mallinckrodt's internal discourse prioritized payer segmentation, pricing strategies, and contract optimization. Their emphasis on pre-launch coordination among marketing, medical affairs, and managed markets, alongside careful considerations of organizational channel design, indicates a structural approach. The documents suggest that Mallinckrodt aimed to navigate the broader distribution environment and secure formulary access to facilitate subsequent sales efforts. Notably, external historical records from (Opioid Industry Documents Archive, n.d.) indicate that Mallinckrodt became one of the largest suppliers of generic oxycodone in the United States during the 2008–2016 period. This expansion temporally overlaps with the internally documented emphasis on institutional market access and commercial infrastructure development, suggesting compatibility between strategic orientation and observed distribution scale rather than implying direct causality.

Ultimately, the computational analysis moves beyond merely quantifying strategy occurrences; it illustrates how commercialization strategies can manifest either as targeted behavioral interventions at the prescriber level (Insys) or as the structural optimization of market access and payer dynamics (Mallinckrodt).

6 Conclusion

In this paper, we proposed a structured NLP framework for analyzing corporate strategic behavior from large-scale corporate presentation documents. The framework integrates (i) keyword-based prefiltering, (ii) supervised refinement using a KNN classifier trained on embedding representations of LLM-labeled text chunks, and (iii) topic mod-

eling with BERTopic to uncover latent thematic structures. LLMs were further employed to identify strategy-relevant topics and generate human-readable explanations, thereby enhancing interpretability.

Applying this framework to documents from Insys Therapeutics and Mallinckrodt Pharmaceuticals, we systematically characterized both shared and divergent forms of corporate strategic behavior. The results demonstrate that combining embedding-based classification, topic modeling, and LLM-assisted interpretation enables a structured, scalable, and reproducible analysis of corporate strategic behavior in large archival corpora. Overall, the proposed framework reduces manual coding effort, improves analytical consistency, and accelerates the extraction of meaningful insights into corporate strategic behavior from unstructured litigation documents.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

7 Limitations and Future Work

Several stages of the analysis rely on Large Language Models (LLMs) for strategy extraction and interpretation. While this approach substantially reduces manual effort and enables scalable processing, the accuracy and interpretive validity of LLM-generated outputs have not been systematically validated by domain experts. As a result, potential issues such as misclassification, oversimplification, or latent bias may persist. Future research should incorporate structured expert evaluation and inter-rater validation frameworks to assess the reliability, consistency, and robustness of extracted strategic themes.

This study focuses exclusively on presentation documents within the OIDA dataset. Although presentations provide structured and strategy-oriented insights, OIDA contains additional document types—such as internal communications, reports, and correspondence—that may offer complementary or contrasting perspectives. Restricting the analysis to a single document category may therefore limit the comprehensiveness of the findings. Extending the scope to include multiple document types would enable a more holistic reconstruction of organizational behavior.

The approach relies mainly on qualitative interpretation rather than a shared quantitative framework, which limits the rigor of cross-corpus comparisons. Future work may explore the design of standardized quantitative evaluation protocols for more reliable cross-corpus benchmarking.

Another limitation concerns the uneven temporal distribution and inherent gaps within the dataset, which together constrain the longitudinal analysis. Documentation from certain periods—particularly prior to 2010—is relatively limited, introducing potential blind spots in the reconstruction of strategic evolution. Because we use raw document counts to reflect the intensity of strategic activity, fluctuations in data availability may partially confound our interpretations. Consequently, observed shifts in topic prevalence or apparent strategic inflection points might reflect archival density and data discontinuities rather than genuine, substantive changes in corporate strategy. To address this, future studies could employ normalization techniques or weighting schemes to better disentangle true strategic signals from artifacts of data availability.

In addition, this study attempted to identify strategies related to mitigating legal and regulatory risks within the opioid market. However, such strategies may not be explicitly articulated in corporate language and are often embedded in indirect or coded expressions. Consequently, the absence of clearly identified legal-avoidance strategies should not be interpreted as definitive evidence of their nonexistence. Future work may benefit from incorporating legal-domain expertise or alternative analytical frameworks designed to detect implicit regulatory positioning strategies.

Finally, the prompting framework employed in this study is designed to extract passages that explicitly describe strategies. While this improves precision, it may reduce recall by overlooking segments that imply strategies indirectly or require deeper contextual inference. With expert supervision or hybrid analytical pipelines, future research could allow LLM systems to flag potentially implicit or strategically sensitive content for subsequent human review, thereby enabling a more comprehensive and nuanced analysis.

8 Ethical Statement

We emphasize that the strategic categorizations and thematic summaries presented in this study are computational artifacts generated by Large Lan-

guage Models (LLMs) processing an archived text corpus. The outputs reflect the model's algorithmic abstraction of the text based on our defined prompts, rather than established legal, economic, or historical facts. Our framework is designed as an exploratory computational tool to process large-scale text data, not as an adjudicative mechanism. Consequently, the findings should not be interpreted as definitive representations of corporate intent or objective truth. Translating these computational signals into substantive conclusions about legal compliance, market manipulation, or economic strategy requires rigorous, independent evaluation by domain experts, including legal scholars, economists, and regulatory analysts.

References

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2025. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, New York, NY, USA. Association for Computing Machinery.

Alessio Faccia, Julie McDonald, and Babu George. 2024. Nlp sentiment analysis and accounting transparency: A new era of financial record keeping. *Computers*, 13(1).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hyewon Kang and Jinho Kim. 2022. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Applied Sciences*, 12(11).

Opioid Industry Documents Archive. n.d. Timeline of the opioid crisis. `https://timeline.oida-resources.jhu.edu/`. Accessed: 2 March 2026.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jantima Polpinij, Manasawee Kaenampornpan, Christopher S. G. Khoo, Wei-Ning Cheng, and Bancha Luaphol. 2026. A multi-stage nlp framework for knowledge discovery from crop disease research literature. *Mathematics*, 14(2).

Xuan Shen, Brian Wingenroth, Zichao Wang, Jason Kuen, Wanrong Zhu, Ruiyi Zhang, Yiwei Wang, Lichun Ma, Anqi Liu, Hongfu Liu, and 1 others. 2025. Oida-qa: A multimodal benchmark for analyzing the opioid industry documents archive. *arXiv preprint arXiv:2511.09914*.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

University of California, San Francisco and Johns Hopkins University. n.d. Opioid industry documents archive (oida). `https://www.industrydocuments.ucsf.edu/opioids/`. Accessed: 6 February 2026.

U.S. Department of Justice. 2019. Founder and four executives of insys therapeutics convicted of racketeering conspiracy. Accessed: 4 March 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. 2025. LLM reading tea leaves: Automatically evaluating topic models with large language models. *Transactions of the Association for Computational Linguistics*, 13:357–375.

A Selection of *K* in KNN

To determine the optimal value of K in the *K-Nearest Neighbors* (KNN) classifier, we employed *Repeated Stratified K-Fold Cross-Validation* (Pedregosa et al., 2011). Specifically, the evaluation procedure was configured with 5 folds and repeated 5 times (resulting in 25 total evaluations), in order to reduce variance induced by random data partitioning and to ensure the robustness of the performance estimates. The dataset consists of 300 text chunks, with each company contributing 150 randomly sampled chunks from a subset prefiltered using domain-specific keywords. These chunks were subsequently labeled in a binary manner ("yes"/"no") by a LLM, indicating whether the chunk is related to the company's commercialization strategy.

We evaluated the classifier across values of K ranging from 1 to 100 and used the F1-score as the primary performance metric, given the potential class imbalance in the dataset. The results are illustrated in Figure 5.

The model achieved its highest F1-score at $K = 5$. Compared to the baseline approach—where all keyword-filtered chunks were directly classified as "yes" (yielding an F1-score of 0.7277)—the KNN classifier with $K = 5$ improved the F1-score by

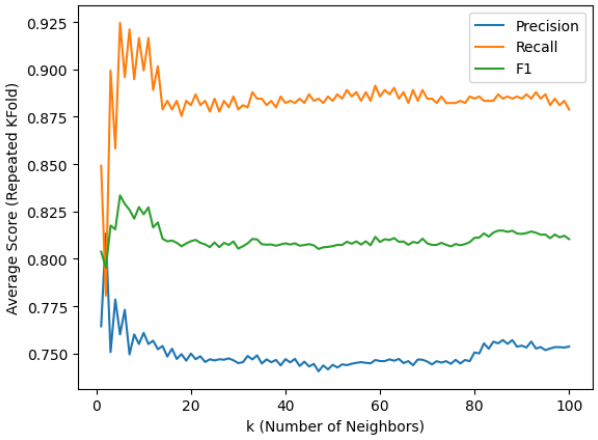


Figure 5: Model Performance over Different Values of K

0.1059. This improvement indicates that incorporating a supervised classification step via KNN effectively reduces noise introduced by keyword-based filtering and enhances the overall predictive performance of the system.

B Prompt Design and Model Configuration

We employed the `openai/gpt-oss-120b` large language model to classify document chunks according to whether they describe a corporate strategy. To ensure deterministic and reproducible outputs, the temperature parameter was set to 0.

B.1 Prompt for Strategy Classification

The following prompt was used to determine whether a given text chunk describes a coordinated corporate strategy or strategic action:

```
{
    "role": "system",
    "content": "You are a research analyst
        studying corporate strategic behavior
        in the pharmaceutical and opioid
        industry.

    Your task is to determine whether a given
        text chunk describes a COMPANY STRATEGY.

    A strategy is defined as a coordinated,
        intentional corporate action designed
        to achieve commercial, regulatory,
        legal, reputational, or market
        objectives.

    A TRUE strategy must:
    - Describe an intentional corporate action
        or coordinated effort
    - Aim to influence revenue, market share,
        regulation, litigation exposure, or
        public perception
    - Go beyond neutral reporting or operational
        background

    Non-strategies include:
    - Lists of advisors or experts
    - Neutral scientific discussion
    - Study design without corporate intent
    - Meeting logistics

    INSTRUCTIONS:
    1. Provide a brief explanation (2-5
        sentences).
    2. On a new line write exactly one of:
       LABEL: YES
       LABEL: NO
    3. Do not write anything after the LABEL
        line."
},
{
    "role": "user",
    "content": "TEXT CHUNK:
    {chunk_text}"
}
```

B.2 Prompt for Topic-Level Strategy Classification

The following prompt was used to classify whether a topic reflects explicit corporate strategic behavior in the opioid industry. The model was instructed to rely strictly on explicit textual evidence and to avoid inference or interpretation beyond the provided excerpts.

```
{
    "role": "system",
    "content": "You are an independent evaluator
        assessing whether a topic explicitly
        reflects corporate strategic behavior
        related to the opioid industry.
    Important:
    - The Topic Representation is contextual
        only.
    - The final decision MUST be based strictly
        on explicit statements in the Document
        Excerpts.
    - Do NOT infer intent.
    - Do NOT rely on background knowledge.
    - Do NOT interpret implications.
    - If it is not explicitly stated, it does
        NOT count.

    Only output the final structured answer.
    "
},
{
    "role": "user",
    "content": "Your task is to assign ALL
        applicable labels based ONLY on
        explicit textual evidence in the
        Document Excerpts.

    LABEL OPTIONS:
    - Sales Expansion & Promotion
    - Influence & Narrative Management
    - Regulatory Risk Management & Evasion
```

```
If none apply, return: None

DEFINITIONS:

Sales Expansion & Promotion requires
    explicit mention of:
- increasing prescription volume
- extending treatment duration
- revenue maximization
- aggressive sales targeting
- formulary positioning through rebates or
    contracting
- expanding into new or higher-risk patient
    populations
- undermining competitors through pricing,
    contracting, or messaging

Influence & Narrative Management requires
    explicit mention of:
- downplaying addiction risks
- shaping scientific evidence or publications
- funding or leveraging key opinion leaders
- sponsoring medical education aligned with
    promotion
- supporting advocacy groups to influence
    public opinion or healthcare policy

Regulatory Risk Management & Evasion
    requires explicit mention of:
- influencing FDA review or regulatory
    processes
- lobbying to shape opioid-related laws or
    policy
- structuring compliance to reduce legal
    exposure
- coordinating litigation defense or
    liability reduction
- managing adverse event reporting or safety
    data to limit regulatory consequences
STRICT RULES:
- Topic Representation provides context ONLY.
- The decision MUST rely on explicit wording
    in the excerpts.
- No inference allowed.
- If wording is vague, it does NOT count.
- Each assigned label MUST have its own
    supporting quote.
- If no label is explicitly supported,
    return None.
OUTPUT FORMAT:
Evidence:
Sales Expansion & Promotion:
- "<exact quoted sentence>"
Influence & Narrative Management:
- "<exact quoted sentence>"
Regulatory Risk Management & Evasion:
- "<exact quoted sentence>"
Justification:
2-4 sentences grounded ONLY in the quoted
    evidence.
Labels:
[comma-separated list of applicable labels
    OR "None"]
---
Topic Representation:
{topic_representation}

Document Excerpts:
{document_excerpts}
"
}
```

B.3 Prompt for Topic Interpretation

The following prompt was used to generate an analytical explanation of each topic identified by the topic modeling procedure. The model was instructed to describe only those promotional or strategic activities that are explicitly supported by the provided topic representation and representative document excerpt.

```
{
    "role": "system",
    "content": "You are analyzing the output of
        a topic modeling system
    for a research project on industry
        promotional strategies.

    Your task is to identify and explain any
        promotional or strategic
    activities that are explicitly described or
        strongly evidenced
    in the provided keywords (Representation)
        and representative document excerpt.

    Focus strictly on strategies, actions, or
        coordinated efforts that are
    clearly supported by the text. Do NOT infer
        hidden motives or intentions.
    Only describe strategic elements that can be
        directly traced to
    specific wording or content in the document.

    Requirements:
    - Write in clear, analytical language.
    - Limit the response to 3-5 sentences.
    - Do not list the keywords.
    - Do not copy phrases verbatim from the
        document.
    - Identify relevant actors (e.g., companies,
        clinicians, regulators, payers).
    - Describe concrete strategic actions only
        if explicitly mentioned.
    - Frame these actions as observable industry
        strategies without speculation.
    - If no clear strategic element is present,
        state that the material is primarily
        descriptive.
    - Ensure that every claim is directly
        traceable to the provided text."
},
{
    "role": "user",
    "content": "
    Representation:
    {representation}

    Representative_Doc:
    {representative_docs}
    "
}
```

Beyond Acoustics: Isolating Dialectal and Sociolinguistic Bias in Spanish ASR

Johnatan E. Bonilla
Humboldt-Universität zu Berlin
j.bonilla@hu-berlin.de

Abstract

Large-scale ASR systems such as Whisper achieve competitive aggregate Word Error Rate (WER) on multilingual benchmarks, but this aggregate conceals systematic disparities across speaker populations. We evaluate Whisper large-v3 on 276 recordings from the *Corpus Oral y Sonoro del Español Rural* (COSER), a dialectological archive of elderly rural speakers across all Spanish provinces. WER is computed separately for Informants and Interviewers within each recording, revealing that mixed-role evaluation underestimates Informant WER in the majority of provinces, with the largest corrections in southern areas. Negative Binomial regression with cluster-robust standard errors shows that Andalusia and Extremadura generate significantly more Informant errors than the Castilian heartland (Andalusia IRR = 1.20, $p < 0.001$; Extremadura IRR = 1.24, $p = 0.020$), while no geographic predictor reaches significance for Interviewers sharing the same recording environment. Male Informants generate 12.5% more errors than females after geographic adjustment ($p < 0.001$), consistent with differential vernacular retention in traditional rural communities. The geographic pattern aligns with established dialectological classifications of Peninsular Spanish. These results demonstrate that role-disaggregated evaluation is a necessary methodological prerequisite for fairness audits of ASR systems applied to sociolinguistically diverse corpora: aggregate benchmarks systematically suppress disparities that are borne disproportionately by the most underrepresented speaker populations, and their use in isolation constitutes both an allocative harm and a measurement failure.

1 Introduction

The adoption of large-scale ASR systems such as Whisper (Radford et al., 2023) for transcription of spoken-language archives has grown rapidly, yet aggregate performance figures carry an implicit assumption of demographic neutrality that does not survive empirical scrutiny. Word Error Rate (WER)—the proportion of reference words incorrectly transcribed, computed as the minimum edit distance between hypothesis and reference divided by total reference words—is the dominant evaluation metric for ASR, but its aggregate form masks population-level disparities. Systematic disparities have been documented across racial groups in English (Koenecke et al., 2020), across gender and dialect in automatic captioning (Tatman, 2017), across stigmatised regional varieties of British English (Markl, 2022), and at the intersection of non-standard phonology and gender (Harris et al., 2024). For elderly speakers, Vipperla et al. (2010) found WER increases of approximately 10 percentage points relative to younger adults. For Spanish, and for non-English languages more broadly, the systematic study of ASR bias remains sparse (Kantharuban et al., 2023).

The most relevant antecedent is San Martín et al. (2024), who evaluated Whisper and SeamlessM4T on the *Corpus Oral y Sonoro del Español Rural* (COSER; Fernández-Ordóñez 2005), over 1,700 hours of sociolinguistic interviews with elderly rural residents across all Spanish provinces. Their principal finding—Whisper large-v3 achieves a mean WER of 0.292 largely stable across dialect regions—leads them to conclude that the model is a viable transcription aid for rural corpus work. We identify two design limitations. First, WER is computed without separating the rural *Informant* from the *Interviewer*; without role-segregated evaluation, any disparity attributable to rurality is absorbed into a global average. Second, no multivariate analysis simultaneously controls for acoustic quality, geography, speaker age, and sex.

Our study provides a role-disaggregated evaluation of a large-scale ASR model on COSER. By computing WER separately for Informants and In-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 123–132

terviewers within the same recordings, we isolate performance differences associated with rural vernacular speech while holding recording conditions approximately constant. We then analyse these differences using a multivariate Negative Binomial regression that controls jointly for audio quality, geography, speaker sex, and age, providing the first geographically disaggregated account of ASR performance disparity in European rural Spanish.[1]

2 Background

2.1 The COSER

The *Corpus Oral y Sonoro del Español Rural* (COSER; Fernández-Ordóñez, 2005; Fernández-Ordóñez and Pato, 2020) comprises 1,947 hours of semi-directed sociolinguistic interviews covering 1,325 rural localities across all 52 Spanish provinces, with 2,574 registered Informants (mean age: 73; 52.4% female). Participants are elderly residents of low formal education who were born and have lived continuously in small rural communities—the sociolinguistic profile plausibly absent from the urban, broadcast, and web-sourced data that dominate ASR training corpora. Informants account for approximately 81.8% of total speaking time (SD = 7.56; San Martín et al., 2024). The Interviewer, generally younger and educated, speaks a variety close to the spoken standard that grounds Whisper's language model.

Transcriptions follow a semi-conventional orthographic norm (Fernández-Ordóñez and Pato, 2020) that encodes surface-level phonological reduction and morphological dialectal variants in surface-faithful orthography, but normalises the most salient phonetic features of southern varieties to standard orthography. This norm produces two competing and partially opposing effects on WER measurement.

Mechanism 1: WER inflation from encoded dialectal forms. Segment deletions and morphological variants are transcribed as produced: forms such as *comprao* (standard: *comprado*), *pa* (standard: *para*), *na* (standard: *nada*), *tá* (standard: *está*), and *to* (standard: *todo*) appear as reference tokens. Morphological variants—*marcharsen*, *traíba*, *tuviendo*—are preserved verbatim. Whisper, whose language model is grounded in standard written Spanish, systematically restores these forms to their standard equivalents: it hypothesises *comprado* where the reference reads *comprao*, *para* where the reference reads *pa*. Each such normalisation generates a substitution or deletion error in the WER computation, despite the fact that Whisper may have correctly identified the acoustic signal. The density of these reduced forms is substantially higher in southern and rural speech—where syllable-final consonant deletion and intervocalic /-d-/ deletion are most advanced—than in the Castilian heartland, producing a systematic WER gradient that is partly orthographic in origin rather than purely acoustic.

Mechanism 2: artificial WER suppression from normalised forms. Conversely, the most salient phonological features of southern varieties—*seseo* (merger of /s/ and the interdental fricative /θ/), *ceceo*, *yeísmo* (loss of the palatal lateral /ʎ/), and glotalisation of coda consonants—are explicitly *not* transcribed, being normalised to standard orthography. The reference always reads *caza* and *casa* with the same sibilant, and *pollo* regardless of whether the speaker produces a palatal lateral or a palatal fricative. When Whisper's output also defaults to standard orthography—whether because it correctly perceived the acoustic signal or because its language model overrides a non-standard input—the two transcriptions agree and no error is registered, masking what may be a genuine recognition failure at the acoustic level. The net effect is that WER *underestimates* Whisper's actual difficulty with southern phonology on normalised features while *overestimating* it on the features that COSER does encode. The WER disparities reported in § 4 therefore represent a conservative lower bound on the true performance gap for southern varieties.

2.2 Whisper's Training Distribution

Whisper's Spanish training data derives from 680,000 hours of weakly supervised web audio (Radford et al., 2023), filtered by language identification but not by speaker demographics. Although the exact composition is not disclosed, indirect evidence suggests systematic under-representation of non-standard varieties: Conneau et al. (2022) showed that multilingual ASR models trained on web-crawled data consistently underperform on low-resource language variants relative to high-resource standard registers; and Pratap et al. (2024) documented that even Massively Multilingual Speech models exhibit performance gaps on re-

[1] Scripts, per-speaker WER tables, and model code: https://github.com/johnatanebonilla/socio-asr-bias/.

Statistic	Value
Recordings analysed	276
Provinces covered	50
Total segments (after filtering)	1,321
Informant segments	530
Interviewer segments	791
Informant segments w/ sex metadata	505 (90.0%)
Informant segments w/ age metadata	447 (79.7%)
Informant mean age (COSER)	73 years
Informant % female (COSER)	52.4%
Recording-level mean WER	0.302

Table 1: Summary of the analysed dataset.

gional varieties absent from their training distributions. These findings concern different architectures (FLEURS, MMS) rather than Whisper directly, but the shared mechanism—web-crawled data over-representing standard registers—makes the inference plausible for Whisper as well. For Spanish specifically, the bulk of online audio plausibly consists of broadcast media, podcasts, and video content produced in urban, educated registers approximating the written norm—precisely the variety closest to the Interviewer's speech. The COSER Informants occupy the opposite pole of this distribution: elderly, rural, low-education speakers whose phonological and morphosyntactic surface forms diverge maximally from the written standard (§ 5.2). The performance gap between Informants and Interviewers, measured within the same recording environment, operationalises the distance between Whisper's training distribution and the target speech.

3 Methodology

3.1 Data and ASR Model

We use 276 COSER recordings with audio and verified transcriptions, spanning 50 provinces.[2] We evaluate Whisper large-v3 (Radford et al., 2023); recording-level mean WER is 0.302, closely replicating San Martín et al.'s result (0.292); the 1 pp difference is attributable to our larger sample (276 vs. 226 recordings), which includes more recently released files from peripheral and island provinces.

3.2 Speaker Segmentation and WER Computation

The COSER XML release includes speaker-turn timestamps, which would in principle allow direct acoustic segmentation by role: each timestamped interval could be extracted, transcribed independently by Whisper, and attributed to the corresponding speaker tag. To assess whether this approach was viable, we evaluated timestamp reliability by comparing, for each XML segment, the reference text falling within the declared boundaries against the word-level timestamps produced by Whisper's own decoder when processing the full recording—a measure of whether the XML boundary corresponds to the acoustic content Whisper actually finds there. Levenshtein similarity between the two text sequences showed substantial and systematic inconsistency: even after text normalisation, only 5,673 of 26,379 segments (21.5%) achieved a similarity score above 0.9, while over 5,000 segments fell below 0.5, indicating that the declared boundaries frequently do not correspond to the actual acoustic content of the recording at those positions. Because timestamp-based segmentation would therefore introduce uncontrolled boundary errors into the attribution procedure, we discarded the XML timestamps for acoustic segmentation entirely. Role attribution is instead achieved through the transcription-level tag system described immediately below, with the full WER computation procedure detailed in § 3.3.

COSER transcriptions encode speaker role through a structured tag system at the segment level. Tags of the form `I`n (`I1`, `I2`, ...) identify Informants and map directly to sociodemographic metadata entries (sex, age, birth year); tags `E`n identify Interviewers—university-trained fieldworkers for whom no demographics are recorded. Tags `IE`n and `II`n mark simultaneous speech involving at least one Informant and one Interviewer, or two Informants, respectively. All overlap segments are excluded from both WER computation and demographic attribution, since overlapping speech cannot be unambiguously attributed to a single speaker's acoustic footprint. Although this strict filtering reduces the analysed volume, it ensures that the WER measured for each role reflects exclusively that speaker's uninterrupted output.

Table 1 summarises the corpus as analysed. After filtering, 530 Informant and 791 Interviewer segments are retained across 276 recordings. The asymmetry—fewer Informant segments despite Informants contributing 81.8% of speaking time—reflects the interview structure: Informants produce long, uninterrupted narrative turns while Interviewers contribute many short question segments. Sex metadata is available for 505 segments (90.0%)

[2] Downloaded from `corpusrural.es`, 2025.

and age for 447 (79.7%); coverage is lower for higher-numbered Informants (I3–I5), who joined interviews opportunistically, as noted in the table caption.

3.3 WER Computation and Error Attribution

WER is computed with `jiwer` following San Martín et al. (2024): bracketed annotations are removed, text is lowercased, and punctuation stripped. Per-speaker attribution traverses the `jiwer.process_words` alignment: substitutions and deletions are attributed to the speaker of the aligned reference word; insertions to the nearest preceding reference word. Concretely, Whisper transcribes the complete audio file as a single linear text. The full reference word sequence is constructed by concatenating normalised text from all non-overlapping segments in order, maintaining a parallel array of per-word role labels. The `jiwer` alignment between Whisper's output and the concatenated reference is then traversed word by word, and each error is assigned to the role label of the corresponding reference position. This approach requires that Whisper's output preserves the temporal order of speech, which is satisfied by its autoregressive decoding.

To illustrate, consider a reference sequence of four words with roles I1: *fue*, I1: *pa*, E1: *para*, I1: *comprarlo*, against which Whisper hypothesises *fue para comprarlo*. The alignment produces: a match on *fue*, a deletion on *pa* (charged to I1), a match on *para* (E1, no error), and a match on *comprarlo*. The single deletion—Whisper's normalisation of the dialectally reduced form *pa* to its standard equivalent—is attributed exclusively to the Informant counter, not pooled into a global figure.

All values reported are *micro-averages* (total errors / total words) unless otherwise noted; this ensures that short segments with high WER do not inflate descriptive statistics relative to the predominant Informant contributions. Because COSER metadata are not uniformly available for all speakers, sociodemographic analyses are conducted on the subset of segments with available annotations; differences in sample size across models therefore reflect metadata coverage rather than sampling decisions.

3.4 Audio Quality

Three complementary objective metrics characterise acoustic conditions per recording, independently of Whisper's output. **SNR** (Signal-to-Noise Ratio, dB) measures the decibel difference between speech power and background noise power; it is a low-level signal measure that does not capture perceptual characteristics such as reverberation or loudness adequacy. **UTMOS** (Saeki et al., 2022) is a neural non-intrusive Mean Opinion Score predictor trained on naturalness judgements from human listeners (VoiceMOS Challenge 2022); it produces a scalar quality estimate on a 1–5 scale without requiring a clean reference signal, making it applicable to field recordings. **NISQA-MOS** (Mittag et al., 2021) is a multi-dimensional perceptual quality model that decomposes overall MOS into four sub-scores: Noisiness (NOI), Coloration (COL), Discontinuity (DIS), and Loudness (LOUD), each on a 1–5 scale.

The choice of quality covariate for multivariate modelling is determined empirically by Pearson and Spearman correlations between each metric and recording-level WER ($N = 276$). Contrary to intuition, raw SNR does not significantly predict WER ($r = -0.081$, $p = 0.156$), while UTMOS ($r = -0.235$, $p < 0.001$) and NISQA-Loudness ($r = -0.233$, $p < 0.001$) do. Noisiness, notably, does not predict WER ($p = 0.456$). This pattern indicates that Whisper is not sensitive to background noise per se but to signal level and perceptual naturalness—a distinction with direct implications for the SNR×dialect interaction reported in § 4.4. SNR is retained in the multivariate model given its wider dynamic range (SD = 10.27 dB) and interpretability. As a robustness check, we re-estimated all models replacing SNR with UTMOS; all geographic coefficients retained sign, magnitude, and significance, confirming that the choice of quality covariate does not drive the reported effects.[3]

3.5 Multivariate Modelling

We model raw error counts using **Negative Binomial GLMs** (NB2, log link), treating the number of transcription errors per speaker segment as the outcome and including $\log(N_{\text{words}})$ as an offset to account for differences in segment length. This formulation is equivalent to modelling error rate on the log scale while respecting the count nature of the data.

Informant predictors entered simultaneously are:

[3] Andalusia Informant IRR shifts from 1.201 (SNR model) to 1.198 (UTMOS model); Sex IRR from 1.125 to 1.122. Full UTMOS models available in the repository.

Autonomous Community (16 dummies, Castile and León as reference), centred SNR (continuous), sex (binary), and age cohort (categorical: 50–70 ref., 71–85, 86+, residual). Cluster-robust standard errors (sandwich estimator) grouped by recording (261 clusters) account for within-session correlation among segments from the same file, functionally equivalent to random intercepts per recording without distributional assumptions on the random effects (Abadie et al., 2022).

Separate but structurally identical models are estimated for Informants and Interviewers, both with cluster-robust standard errors. The parallelism of these two models is the central inferential strategy: a geographic coefficient that is positive and significant for Informants but absent for Interviewers—who occupy the same physical recording environment—is consistent with linguistic variety as the primary source rather than recording conditions. Coefficients are reported as incidence rate ratios (IRR $= e^{\hat{\beta}}$) with 95% confidence intervals and both standard and cluster-robust p-values.

4 Results

Global micro-averaged WER is 0.309 for Informants and 0.294 for Interviewers—a 1.5 pp gap that appears modest in aggregate but conceals pronounced geographic and sociodemographic structure, as the following subsections demonstrate.

4.1 Province-Level Geographic Distribution

Figure 1 presents Informant and Interviewer WER across all 50 provinces. Table 2 reports the 10 highest and 5 lowest Informant WER provinces.

The nine highest-WER provinces with positive Informant–Interviewer gaps are all southern (Andalusian: Almería, Sevilla, Cádiz, Málaga, Córdoba), western peripheral (Galician: Orense, Lugo), or Extremaduran (Cáceres). The exception is Soria, a northern province with small sample size. Albacete shows equally high Interviewer WER, suggesting shared acoustic difficulty rather than a linguistic effect. Orense shows an exceptionally high Interviewer WER (0.384), nearly matching its Informant WER, which may reflect shared recording-quality issues in the Galician sessions rather than a purely linguistic effect. The five lowest-WER provinces—all northern—show inverted gaps where Interviewers produce *higher* WER than Informants, consistent with rural varieties close to the Castilian standard.

Province	Inf.	Int.	Gap	N
Almería	.418	.295	+.123	4
Orense	.415	.384	+.032	5
Sevilla	.409	.315	+.095	4
Cádiz	.407	.278	+.129	5
Málaga	.401	.295	+.106	4
Lugo	.382	.255	+.127	4
Albacete	.381	.392	−.011	4
Cáceres	.380	.359	+.022	5
Córdoba	.372	.290	+.082	5
Soria	.371	.331	+.040	5
Álava	.236	.276	−.040	6
Gerona	.231	.233	−.002	6
Segovia	.227	.279	−.052	6
Valladolid	.223	.300	−.077	7
Vizcaya	.201	.282	−.082	7

Table 2: Provinces with the 10 highest and 5 lowest Informant micro-WER. Gap = Inf. − Int. Of the 50 provinces, 14 show gap $> +0.05$ (all southern or western); 8 show gap < -0.05 (all northern). Provincial estimates with $N = 4$ should be interpreted with caution given the small number of recordings.

4.2 Autonomous Community Aggregation

Table 3 and Figure 2 aggregate the provincial data to the 17 Autonomous Communities (*Comunidades Autónomas*), which serve as geographic predictors in the multivariate model.

Interviewer WER does not track the Informant gradient. Andalusia, the community with the largest sample ($N = 38$), shows Informant WER of 0.370 against Interviewer WER of 0.298—a gap of +0.072. Murcia (0.367 vs. 0.292, gap +0.075) and Extremadura (0.355 vs. 0.318, gap +0.037) follow the same pattern. Galicia shows a similar pattern (0.340 vs. 0.303, gap +0.038). Cantabria shows a large gap (+0.072) but with only $N = 6$ recordings. Castile and León (0.285), Navarre (0.258), and Basque Country (0.255) show inverted or near-zero gaps. In Castile and León the Interviewer WER (0.292) actually exceeds the Informant WER (0.285). The Canary Islands represent an interesting case: Informant WER (0.311) is elevated relative to the best WER peninsular communities but the gap (+0.064) is moderate, a pattern we discuss in § 5.2. Madrid ($N = 4$) and La Rioja ($N = 5$) show extreme inversions driven by short recording samples and should not be interpreted inferentially; note that Murcia ($N = 5$) similarly has a small sample, and its high IRR in the regression model should be interpreted with corresponding caution. Because both speakers share the same

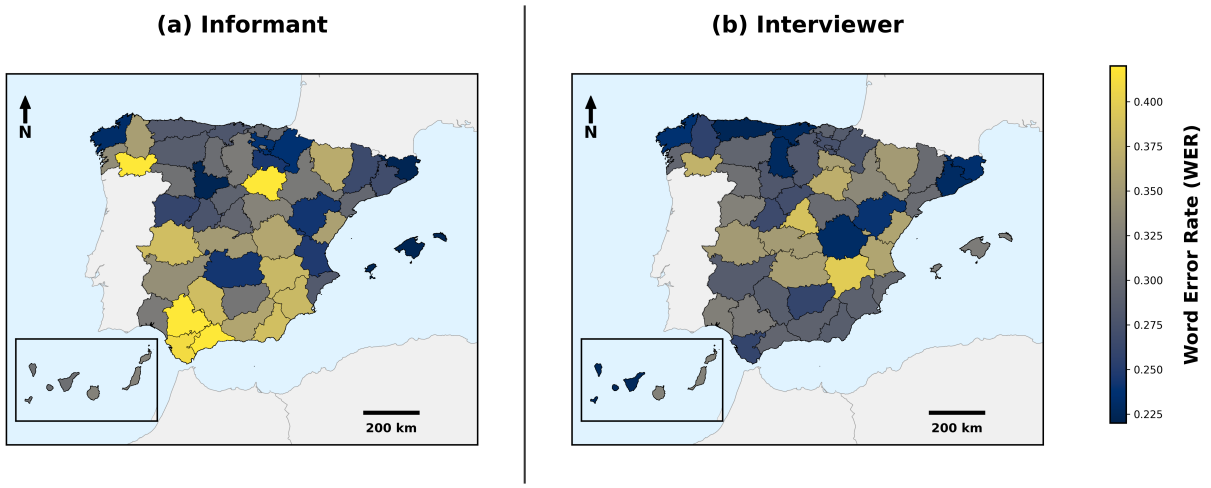


Figure 1: Province-level Informant (left) and Interviewer (right) micro-WER.

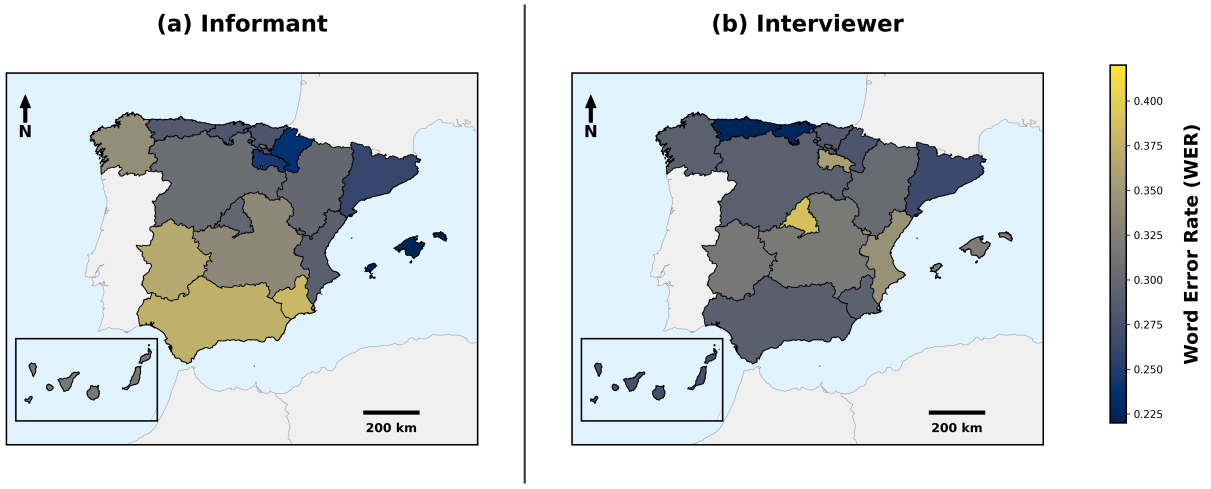


Figure 2: Autonomous Community-level Informant (left) and Interviewer (right) micro-WER.

recording environment within each recording, the systematic dissociation between the Informant gradient and the flat Interviewer pattern is consistent with linguistic variety as the primary source of the disparity.

4.3 Sociodemographic Effects

Among 505 Informant segments with sex metadata, males show consistently higher WER than females. The male micro-average is 0.320 versus 0.299 for females—a 2.1 percentage point gap. The difference is statistically significant (Mann-Whitney $U = 36{,}266$, $p = 0.003$, rank-biserial $r = 0.12$). The rank-biserial indicates a small effect size; the amplification to IRR = 1.125 (12.5%) in the multivariate model reflects the redistribution of variance after geographic adjustment.

Among the 447 Informant segments with age metadata, 422 correspond to speakers aged 50 or above. Three cohorts were defined: 50–70 ($n = 139$), 71–85 ($n = 237$), and 86+ ($n = 46$). No significant age difference emerges (Kruskal-Wallis $H = 1.37$, $p = 0.503$; means 0.310, 0.311, 0.303). The male–female gap is stable across all three cohorts at approximately +0.02, indicating that sex and age are orthogonal predictors in this population.

4.4 Multivariate Analysis

Table 4 reports the Negative Binomial GLM with both standard and cluster-robust p-values.

The Informant model reveals two robustly significant communities. Andalusia generates 20.1% more errors than Castile and León (IRR $= 1.201$, $p_{\text{clust}} < 0.001$), and Extremadura generates 24.0% more errors (IRR $= 1.240$, $p_{\text{clust}} = 0.020$). Galicia (IRR $= 1.199$, $p_{\text{clust}} = 0.117$) and Murcia (IRR $= 1.294$, $p_{\text{clust}} = 0.123$) show consistent positive

Aut. Community	Inf.	Int.	Gap	N
Andalusia	**.370**	.298	+.072	38
Murcia	.367	.292	+.075	5
Extremadura	**.355**	.318	+.037	10
Galicia	.340	.303	+.038	20
Castile-La Mancha	.314	.319	−.005	27
Canary Islands	.311	.246	+.064	11
Aragon	.307	.298	+.009	18
Cantabria	.303	.231	+.072	6
Castile & León	.285	.292	−.007	55
Valencian C.	.284	.307	−.022	16
Madrid	.282	.404	−.122	4
Asturias	.273	.232	+.041	7
Catalonia	.272	.272	.000	22
Navarre	.258	.290	−.032	7
Basque Country	.255	.281	−.026	20
La Rioja	.253	.373	−.120	5
Balearic Isl.	.251	.304	−.054	5
Global	*.309*	*.294*	*+.015*	*276*

Table 3: Micro-WER by Autonomous Community.

effects of similar magnitude that do not survive the more conservative cluster correction, likely reflecting the limited number of recording clusters in those communities ($N = 20$ and $N = 5$ respectively).

No geographic predictor is positive and significant in the Interviewer model: the largest Interviewer IRR is Extremadura (1.156), which does not reach significance ($p = 0.133$). The two significant Interviewer coefficients are both *negative*—Cantabria (0.739) and Asturias (0.805)—indicating that Interviewers in those communities generate *fewer* errors than the Castile and León reference.

Male Informants generate 12.5% more errors than females (IRR $= 1.125$, $p_{\text{clust}} < 0.001$), a robust effect that survives all geographic specifications. Neither age cohort reaches significance (71–85: IRR $= 0.971$, $p = 0.488$; 86+: IRR $= 0.961$, $p = 0.573$), confirming the bivariate null result. SNR is significant for Informants ($p_{\text{clust}} = 0.011$) but not for Interviewers ($p = 0.718$), indicating that audio quality affects recognition of vernacular speech more than standard speech.

The pseudo-R^2 values (Informant: 0.048, Interviewer: 0.013) indicate that geography, audio quality, sex, and age together explain a modest share of total WER variance; unmeasured variables such as speech rate, lexical density, and individual articulatory characteristics likely account for a substantial portion of the remaining variance. Nevertheless, the 3.6× ratio between models confirms that the measured predictors structure Informant performance far more than Interviewer performance—precisely the asymmetry predicted by a linguistic account of the disparity.

	IRR	95% CI	p_{std}	p_{clust}
Informant model (N=530 segments, 261 clusters)				
Andalusia	1.201	[1.07, 1.35]	.002	.0004***
Extremadura	1.240	[1.04, 1.48]	.019	.020*
Galicia	1.199	[1.05, 1.37]	.009	.117
Murcia	1.294	[1.01, 1.66]	.041	.123
Canary Isl.	1.087	[0.90, 1.31]	.373	.414
SNR (/dB)	0.995	[0.99, 1.00]	.005	.011*
Sex (male)	1.125	[1.05, 1.20]	.001	.0005***
Age 71–85	0.971	[0.90, 1.05]	.468	.488
Age 86+	0.961	[0.84, 1.09]	.541	.573
Interviewer model (N=791 segments, 261 clusters)				
Andalusia	1.022	[0.92, 1.13]	.677	.702
Extremadura	1.156	[0.96, 1.40]	.133	.158
Galicia	1.080	[0.95, 1.23]	.250	.289
Murcia	1.001	[0.81, 1.23]	.995	.996
SNR (/dB)	1.001	[1.00, 1.00]	.718	.734

α: Inf. 0.133, Int. 0.149. Pseudo R^2: Inf. 0.048, Int. 0.013 (3.6×). Ref.: Castile and León (geog.), 50–70 (age).
11 remaining CCAA non-significant ($p_{\text{clust}} > 0.05$) in both. Interviewer significant: Cantabria 0.739 (p=.009), Asturias 0.805 (p=.037)—both negative.
***$p < 0.001$; *$p < 0.05$. Cluster-robust SEs in both models: sandwich estimator grouped by recording (261 clusters).

Table 4: Negative Binomial GLM results for the Informant (top) and Interviewer (bottom) models.

5 Discussion

5.1 The Geographic Disparity is Linguistic

The within-recording Informant/Interviewer contrast is the core empirical contribution of this study. Two speakers sharing the same recording environment produce divergent WER values that track geography systematically. This within-recording contrast substantially reduces the plausibility of acoustic quality as the primary explanation for the geographic gradient, although we note that microphone distance and angle may vary between speakers within a session (§ 6).

The finding extends San Martín et al. (2024), who documented geographic variation in overall WER but could not isolate its source because their evaluation conflated Informant and Interviewer speech. Our role-disaggregated analysis reveals that the geographic gradient is entirely concentrated in the Informant channel: the Interviewer channel is geographically flat. The pseudo-R^2 ratio (3.6× for Informants vs. Interviewers) quantifies this asymmetry.

This pattern converges with findings across typologically distinct contexts. Koenecke et al. (2020) showed roughly double the error rate for African American English relative to white American English across five commercial ASR systems, attributing the gap to training-data composition. Markl (2022) extended this analysis to stigmatised British English varieties, arguing that performance gaps constitute both allocative and representational harms: speakers of non-standard varieties receive worse service from ASR and are implicitly positioned as deviations from a norm. Harris et al. (2024) showed that the interaction of gender and dialect is the primary driver of ASR error in non-standard American English. Our contribution extends this framework to a Romance language context where the relevant axis is not race but the rural–urban, vernacular–standard continuum structuring Peninsular Spanish dialectology.

	Group	Inf.	Int.	IRR
García Mouton (phonetic)				
Inf.	Northern	.290	.290	ref.
	Southern	.340	.301	1.131***
Int.	Northern	—	.290	ref.
	Southern	—	.301	$1.026^{\text{n.s.}}$
Fernández-Ordóñez (morphosyntactic)				
Inf.	North-Central	.296	.299	ref.
	Western	.330	.299	1.169**
	Southern	.346	.289	1.108*
	Eastern	.277	.290	$0.974^{\text{n.s.}}$
Int.	North-Central	—	.299	ref.
	Western	—	.299	$1.071^{\text{n.s.}}$
	Southern	—	.289	$0.981^{\text{n.s.}}$
	Eastern	—	.290	$1.027^{\text{n.s.}}$

Table 5: Auxiliary NB GLMs with dialectological classifications. Micro-WER (Inf./Int.) and Informant/Interviewer IRRs. All Informant models include SNR, sex, and age. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; n.s. not significant.

5.2 Alignment with Peninsular Spanish Dialectology

The communities and provinces with the highest Informant WER correspond to what established dialectological frameworks identify as the varieties most distant from the Castilian standard. We briefly describe the two classifications used and report auxiliary Negative Binomial models that replace the CCAA dummies with these classifications (Table 5).

García Mouton (1994) organises Peninsular Spanish along a phonetic axis, distinguishing a conservative *Northern* area characterised by maintenance of the *distinción* (contrast between alveolar /s/ and interdental fricative) and strong articulation of coda consonants, from an innovative *Southern* area—encompassing Andalusia, Canary Islands, Murcia, Extremadura, and the Valencian and Albacete transition zones—where three convergent phonological changes produce surface forms maximally distant from orthographic norms: yeísmo (loss of the palatal lateral), deletion of intervocalic /-d-/, and the progressive assimilation, neutralisation, and loss of coronal consonants in syllable coda. As García Mouton notes, the epicentre of these changes is western Andalusia, from which they radiate in successive stages northward and to the Canary Islands.

Fernández-Ordóñez (2016) departs from phonetic criteria by grounding the classification in grammatical evidence from the ALPI and the COSER itself. Her division identifies a *Western* area (Cantabria, both Castiles, Asturias, and Galician-Portuguese contact provinces) defined by the mass/count pronominal system (*neutro de materia*), leísmo, laísmo, and loísmo; an *Eastern* area characterised by inflected infinitives (*-sen*) and subjunctive-to-conditional displacement; and a *Southern* area where the etymological pronominal system predominates. The Western area—shaped by contact with Galician-Portuguese—cuts across the phonological north–south divide, capturing a morphosyntactic dimension that purely phonetic classifications miss.

Under García Mouton, Southern Informants generate 13.1% more errors than Northern Informants (IRR $= 1.131$, $p < 0.001$); the Interviewer contrast is 2.6% and non-significant ($p = 0.435$). Under Fernández-Ordóñez, both the Southern (IRR $= 1.108$, $p = 0.029$) and the Western area (IRR $= 1.169$, $p = 0.002$) show significant Informant effects.

A notable exception is the Canary Islands (Informant WER $= 0.311$, IRR $= 1.087$, $p_{\text{clust}} = 0.414$), classified as Southern yet showing no significant disparity. Canarian Spanish occupies a well-documented position as an interdialect between Peninsular and Latin American varieties (García Mouton, 1994): it shares with Caribbean Spanish generalised seseo and /-s/ aspiration—features plausibly well represented in Whisper's web-sourced training data given the large volume of Latin American audio online. This hypothesis—

that Canarian speech enjoys proximity to Whisper's training distribution that mainland Southern varieties do not share—is consistent with the observed pattern but remains speculative without access to the training composition.

5.3 The Sex Effect

The male–female disparity is robust across all specifications (Mann-Whitney $p = 0.003$; NB $p_{\text{clust}} < 0.001$ in the CCAA model; $p = 0.001$ under both García Mouton and Fernández-Ordóñez). Male Informants generate 12–13% more errors after geographic, acoustic, and age adjustment. The direction reverses the English pattern (Tatman, 2017; Feng et al., 2024)—a difference explained by the operative axis of variation. In English audits, the disparity is attributed to over-representation of male broadcast speech; in rural Spanish, the operative axis is variety proximity to the standard. In traditional rural communities, women adopt prestige phonological features at higher rates while men retain local vernacular phonology (Labov, 2001), producing speech further from Whisper's standard-oriented language model. This is consistent with Harris et al. (2024), who show that the gender–dialect interaction is the primary driver of ASR error in non-standard American English. We acknowledge that this interpretation is post-hoc: the same WER difference could partially reflect sex-linked differences in speech rate or articulatory precision, and direct evidence of the mediating mechanism (differential adoption of prestige features) would require a phonological error analysis that the current design does not provide.

6 Conclusions

This study demonstrates that aggregate WER conceals systematic sociolinguistic disparities in Whisper's performance on rural Spanish. Role-segregated evaluation reveals that mixed-role benchmarks underestimate the Informant WER in the majority of provinces, with the largest corrections in southern communities where dialectal divergence from the standard is greatest. Negative Binomial regression with cluster-robust standard errors identifies Andalusia and Extremadura as robustly generating 20–24% more Informant errors than the Castilian heartland, while no geographic predictor reaches significance for Interviewers sharing the same recording environment. Male Informants generate 12.5% more errors than females—a pattern consistent with differential vernacular retention and opposite to English audits. The disparity aligns with established dialectological classifications: both García Mouton's phonetic axis and Fernández-Ordóñez's morphosyntactic framework predict the geographic gradient, with the latter revealing a Western (Galician-Portuguese contact) effect invisible to administrative boundaries.

Limitations and Future Work

The within-recording design assumes approximately shared acoustic conditions for Informant and Interviewer. In practice, microphone distance and angle may vary: the Informant is typically seated facing the recorder while the Interviewer may move, consult notes, or sit at a different distance. This potential asymmetry cannot be measured from the audio alone and represents a residual confound.

Speaker attribution relies on transcription-level segmentation rather than time-aligned diarisation. In segments with very high WER—where alignment between Whisper's output and the reference degrades—error attribution to roles becomes less precise. This limitation affects Southern Informants disproportionately, since they exhibit the highest WER. Time-aligned diarisation of the COSER audio would strengthen the attribution procedure and enable fine-grained acoustic analyses at the speaker level.

All evaluations use Whisper large-v3; generalisation to wav2vec 2.0, MMS, or fine-tuned models requires further work. Galicia and Murcia show consistent positive effects that do not survive cluster correction (both $p \approx 0.12$); for Murcia ($N = 5$ clusters), the asymptotic properties of the sandwich estimator may not hold, and the cluster-robust p-value should be interpreted cautiously (Abadie et al., 2022).

Acknowledgments

This research was carried out within the Collaborative Research Centre SFB/CRC 1412 *Register: Language Users' Knowledge of Situational-Functional Variation*, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 416591334.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2022. When should you ad-

just standard errors for clustering?*. *The Quarterly Journal of Economics*, 138(1):1–35.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech and Language*, 84:101567.

Inés Fernández-Ordóñez. 2005. COSER: Corpus oral y sonoro del español rural. Universidad Autónoma de Madrid.

Inés Fernández-Ordóñez. 2016. Dialectos del español peninsular. In Javier Gutiérrez Rexach, editor, *Enciclopedia lingüística hispánica*, volume 2, pages 387–404. Routledge, London and New York.

Inés Fernández-Ordóñez and Enrique Pato. 2020. El COSER (Corpus Oral y Sonoro del Español Rural) y su contribución al estudio de la variación gramatical del español. In Ángel J. Gallego and Francesc Roca, editors, *Dialectología digital del español*, number 80 in Verba. Anexo, pages 71–100. Universidade de Santiago de Compostela, Santiago de Compostela.

Pilar García Mouton. 1994. *Lenguas y dialectos de España*. Arco Libros, Madrid.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 2001. *Principles of Linguistic Change, Volume 2: Social Factors*. Blackwell, Oxford.

Nina Markl. 2022. Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 521–534, Seoul, Republic of Korea. ACM.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-Self-Attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021)*, pages 2127–2131, Brno, Czechia.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Mirari San Martín, Jónathan Heras, Gadea Mata, and Sara Gómez. 2024. Is ASR the right tool for the construction of spoken corpus linguistics in European Spanish? *Procesamiento del Lenguaje Natural*, 73:165–176.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2010. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP J. Audio Speech Music. Process.*, 2010.

Who Speaks for Whom? LLM-Generated Survey Data as a Proxy for Public Opinion

Radhakrishnan Venkatakrishnan[1], **Travis Brodbeck**[1,2], **Michael D. Young**[1],
[1]University at Albany, [2]Siena University,
Correspondence: rvenkatakrishnan@albany.edu

Abstract

Technological advancements, such as Large Language Models (LLMs), offer a potential solution to the two-faceted problem facing social science researchers: rising costs and declining response rates. The use of artificial personas is a budding practice, where chatbots are given the demographic characteristics of the person they are supposed to role-play as and answer questions for researchers. Before scholars and practitioners augment or replace the data created by interviewing humans, it is essential to understand how well models perform in generating accurate, reliable, and robust data, with concerns that the training of LLMs results in a bias towards the norms of WEIRD cultures. We present a procedure for practitioners to use to evaluate the quality of their synthetic data by measuring Intra Class Correlation (ICC), Earth Mover Distance (EMD), Variance, Hedging, and demographic drivers of LLM output. We find that the models may generate plausible results in the aggregate, but these synthetic data do not exhibit the depth or nuance of human respondents. Secondarily, we find that despite having generated definitive answers on a ten-point scale, the reasoning provided by the LLM exhibited varying degrees of hedging that do not consistently align with the LLM's answer. The distortion of the results was not uniformly distributed; instead, the effects were more extreme for some demographic groups. Our findings suggest that the technology generating synthetic survey data may not be mature enough to address the increasing challenges of interviewing humans for public opinion research. Code and data are available in Github.[1]

1 Introduction

The evolution of LLMs underlying Artificial Intelligence (AI) tools suggests that the technology may be approaching the limits of the Turing Test (Reinbold, 2020; Bhatnagar, 2026), moving from simple imitation to sophisticated impersonation. Researchers face the question of how AI will change social science research, specifically in how it is conducted. Recent scholarship indicates that humans are already struggling to distinguish between human-authored and LLM-generated text (Kreps et al., 2022). This blurring of lines presents a fundamental challenge for social science researchers: if general audiences cannot discern the origin of content - human or AI, researchers may soon face an increasingly difficult task in distinguishing synthetic and authentic data. Before social science researchers embark on the utilization of synthetic or manufactured data, it must be reliably comparable to human responses, and we must understand how those results are generated.

Research budgets are stretched due to declining survey response rates (Eggleston, 2024) attributed to a variety of technological changes, such as the adoption of answering machines (Oldendick and Link, 1994), caller ID (Link and Oldendick, 1999), cell phones (Brick et al., 2007), and call screening conducted by AI agents on modern smartphones (Markus, 2025). Technological change is not the only cause for declining response rates as non-researchers, such as telemarketers, contributed to public's aversion to answering the phone (Link et al., 2006). These factors combined with the public's eroding trust in institutions like pollsters (Johnson et al., 2024) create an environment where conducting the "gold standard" of probabilistic telephone research is both more difficult and more expensive. As researchers moved to online surveys, they faced challenges in data quality between collection modes (Couper and Miller, 2008), similar to the differences in data between self-administered paper interviews and telephone interviews (Vannieuwenhuyze et al., 2010). Against the backdrop of rising costs and the ubiquity of internet access, non-probability panels and opt-in surveys are more commonly used for survey experiments and data

[1]rvenka31/llm-proxy-public-opinion-surveys

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 133–148

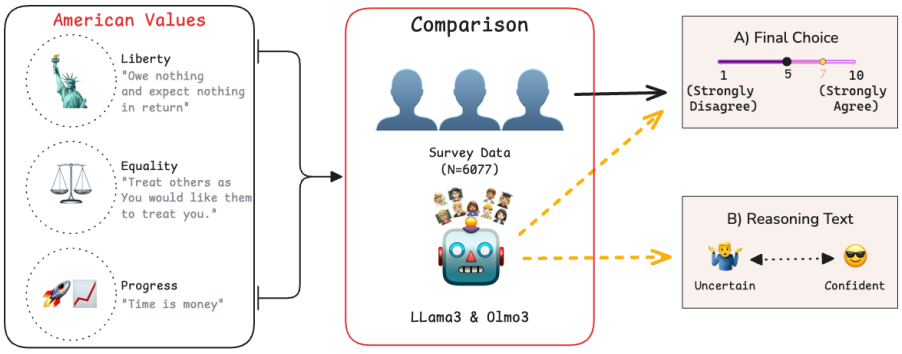


Figure 1: A visual overview of the research framework, illustrating the relationship between demographic inputs and representational fidelity compared on the American Values Survey between human responses and LLM responses. Human responses include only their Final choice, marked on a Likert scale from 1 to 10. LLM responses also include inference-time tokens called reasoning traces.

collection (Callegaro and DiSogra, 2008). Due to the costs of probability sampling designs, non-probability panels are increasingly used, especially when trying to interview hard-to-reach populations, often young people, men, and people of color (Pyrooz et al., 2025).

Just as widespread internet access once made online and non-probability sampling more common, the release of ChatGPT 3.5, along with other similar models, has given researchers the opportunity to leverage new means to answer questions in new ways (Hayashi, 2024). This opportunity puts researchers in a tough position: between the rock of methodological rigor and the hard place of rising costs fueled by declining response rates. Like online surveys, a method for collecting data more cheaply and quickly is tempting. Given this temptation, this paper investigates the viability of using synthetic responses generated by LLMs to measure public opinion.

Synthetic data are manufactured data designed to mimic real-world data by using techniques like deep learning and generative AI (Joshi et al., 2024). It is also referred to as "silicone samples", when tasked to mimic human participants in public opinion surveys (Argyle et al., 2023). Using synthetic data is appealing for several reasons. Unlike humans, LLMs are unlikely to suffer from interview fatigue where interpreting and answering a few dozen questions weighs on one's cognitive ability, potentially biasing future answers in the survey (Ghafourifard, 2024). Related to fatigue, humans satisfice by skipping optional questions, saying they don't know, or providing incomplete answers to finish the interview faster (Krosnick, 1991; Krosnick et al., 2002). LLMs are programmed to complete interviews as instructed, whereas humans may be interrupted, uninterested, skeptical of the prompts, and decide to terminate mid-interview. Humans can be difficult to reach at certain times of day or the week (Weeks et al., 1987), whereas LLMs can be summoned at any time. With the exception of computational costs, LLMs do not ask for financial incentives to participate in research, providing ample opportunity for experimentation. Humans can experience difficulty reading or hearing, whereas LLMs are not restricted by the senses. Humans either decide to participate in an interview out of their own personal motivations, whether that is to help the researcher, to advance research, or to receive an incentive (Hjortskov et al., 2023). LLMs, on the other hand, are programmed to follow instructions and act upon demand.

LLMs exhibit similar problematic behaviors to humans. As humans will lie, LLMs lie or hallucinate the facts (Farquhar et al., 2024). As humans may experience acquiescence bias, being more agreeable to minimize conflict (Davis et al., 2019), LLMs have shown affirmative or positivity bias, saying yes or agreeing with the human prompting the tool (Fanous et al., 2025). LLM outputs are subject to multiple sources of variability: prompt formulation, chain-of-thought instructions, and inference parameters such as temperature, top_p, and top_k — all of which control output randomness and can reduce reliability (Li et al., 2025; Wei et al., 2022). Similarly to noticing social cues and norms, LLMs often gravitate to the mean and can show less variation than would be observed with real data (Xie and Xie, 2025). Like humans having blind spots, the training data for underrepresented groups could bias output and lead to inaccurate syn-

thetic data that could harm research of marginalized groups (Foka et al., 2025; Santurkar et al., 2023). As humans learn throughout the day, obtaining new information, LLMs gain new information to update their foundational model and training data, or through fine-tuning and prompting, creating possible knowledge gaps or incomplete datasets that result in less accurate output.

Incorporating synthetic samples creates many opportunities for cost savings and experimentation, but also introduces significant risks regarding data quality and representativeness. This paper seeks to navigate this "new world" of survey methodology, specifically exploring the rise of using LLMs to create synthetic or artificial data to measure public opinion. Looking to elicit values and morals embedded in an LLM to compare against human respondents is an evolving research direction (Pistilli et al., 2024; Jiang et al., 2024). The World Values Survey (Haerpfer et al., 2022) collects responses from human participants worldwide and has been adapted for LLM evaluation in works such as (Zhao et al., 2024), which evaluated LLMs on the implicit and explicit values they express across different test settings. The OpinionQA dataset captures misalignment in steering LLMs with given persona on ATP questionnaires. (Santurkar et al., 2023) While there is growing evidence that LLMs exhibit WEIRD (Western, Educated, Industrialized, Rich, and Democratic) alignment (Zhou et al., 2025), very few works (Santurkar et al., 2023) have looked into this alignment. Yet the focus of these works was broad and not solely on American values, despite their centrality to the Western dimension of WEIRD. Since LLMs tools offer the potential for richer, conversational data collection, they require a rigorous framework for measurement and evaluation. For synthetic data to be viable, we must be able to: **(1) Quantify Model Behavior:** Develop metrics to measure the tendency of models to provide cautious, non-committal answers—and other stylistic artifacts. **(2) Define Appropriateness:** Establish benchmarks for when synthetic data is an acceptable proxy for human opinion and when it introduces unacceptable error. **(3) Assess Output:** Apply established metrics to ensure the validity of social science research without compromising quality.

Ultimately, while LLMs offer innovative paths for pre-testing and imputing missing data, their role in representing the collective "voice" of the public must be scrutinized. We must determine if we are accurately measuring opinion or simply reflecting a distorted version of the past. Therefore, before synthetic data can be considered a valid proxy, we must develop strict frameworks to measure its failures, focusing on its demographic stereotyping and artificial confidence. To that end, this research critically evaluates model behavior through the following questions

RQ1: To what extent do LLM responses align with human respondents across demographic groups?

RQ2: To what extent do demographic variables influence LLM responses relative to human responses?

RQ3: How does model selection, scale, prompting, and reasoning influence LLM performance?

We will analyze LLM responses across different demographic inputs to identify patterns of stereotyping and misrepresentation, establishing when synthetic data serves as an acceptable proxy for human opinion and when it does not. By systematically mapping failure points, we intend to help with the adoption decision of these tools in survey-based social science research.

2 Methodology

2.1 Data

To answer these questions, we simulate synthetic responses for individual human profiles based on the actual survey. Our source dataset is the American Values Survey ($N = 6,077$), comprising 34 ten-point Likert-scale value statements aggregated into three subscales: *Liberty*, *Equality*, and *Progress* (Gibson and Lipinski, 2021). The survey was conducted in 2021 by Sienna University [2] and it is detailed in Appendix §A.1. Among the demographic information captured, we select 9 demographic variables (Age, Race, Ethnicity, Gender, Employment, Education, Political Affiliation, State, and Voter Registration) for our experiment.

2.2 Models

We evaluate four open-access instruction-tuned models: Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Olmo-3-7B-Instruct, and Olmo-3.1-32B-Instruct (Olmo et al., 2025). They were selected to explain two effects: parameter count (8B/7B vs 32B/70B) to test if size increases demographic sensitivity, and model

[2] https://sri.siena.edu/the-american-values-study/

family (Llama vs Olmo) to assess the generalization within open-access US-based models across architectures. For each human respondent h_i in the dataset, we construct a persona-based system prompt $P(h_i)$ that includes all 9 demographic attributes and generate $k = 5$ responses using a temperature of $\tau = 0.1$, which introduces minimal stochastic variation. Detailed prompt and hyperparameter settings are provided in the Appendix §A.2.

2.3 Prompts

We experiment with two prompt variants to assess the impact of "chain-of-thought" (CoT), which prompts the model to think step-by-step (Wei et al., 2022; Kojima et al., 2023).

Final Choice First (FCF): Prompting the model to provide a numerical score for each value question before articulating its reasoning trace.

Reasoning First (RF): Prompting the model to first generate a detailed reasoning trace before arriving at a numerical score for the value question.

This design allows us to quantify the effects of CoT and its order on the model's behavior. The language of the two prompts is identical, except for the order of the reasoning and final choice. The complete prompt is provided in the appendix Table 4.

2.4 Evaluation

We define three complementary metrics to quantify the model's performance in survey simulation.

1. **Behavioral Consistency:** Using Variance σ^2 and Intra Class Correlation(ICC), we capture the consistency of the model's score across iterations and for the same demographic profile. We use ICC(1,1) to assess the reliability of individual iterations and ICC(1,k) to assess consistency across the full set (Shrout and Fleiss, 1979).

2. **Alignment Quality:** We use a combination of *Wasserstein Distance/Earth Mover Distance (EMD)* and *Variance ratio* to assess how LLM scores align with human respondents and whether they vary in a similar vein to the human respondents for a given question. Using EMD, we measure the model's simulated score distribution against human ground truth for value constructs, thereby capturing representational distortion, as in (Zhao et al., 2024). Variance ratio, unlike iteration variance (captured by ICC), assesses whether the LLM differentiates between demographic groups to the extent that human respondents naturally differ on a value question. Together, these two measures characterize each question along two dimensions: how far the LLM deviates from human responses (EMD) and whether it responds to demographic variation proportionally (variance ratio).

3. **Hedging% ($H\%$):** Quantifies the model's epistemic uncertainty by analyzing the frequency of hedging language in the reasoning traces for the simulated scores. We quantify the model's uncertainty by analyzing the reasoning trace. Using a verified lexicon of hedging markers L_{hedge} (e.g., "assume", "even though", "appears to") from (Islam et al., 2020), we calculate the Hedging% $H\%$ as the proportion of words that are hedging words across the k iterations. There are 657 unique hedging markers grouped into three categories: hedge words, booster words, and hedging phrases. For this analysis, we group the three categories into a single category, i.e., hedging. Using Spacy's tokenization, we identify hedging in the reasoning trace by grouping words and phrases.

Together with the three dimensions, we can assess the model's behavioral rigidity, representational distortion, and epistemic amplification or suppression across different demographic profiles. The first two dimensions (consistency and alignment) capture the model's rigidity, ensuring that LLMs are sufficiently replicative, and explain representational distortion by focusing on the score distributions and their divergence from human responses. The third metric (Hedging%) captures the model's reasoning uncertainty during generation, which may reflect its confidence in the assigned scores and its awareness of potential biases or limitations in its knowledge.

2.5 Regression

Using metrics derived from the evaluations and additional targets, we specify LLM score, Absolute Error (|LLM score - human score|), Signed Error (LLM score - human score), and Hedging% as dependent variables in the Ordinary Least Squares (OLS) regression analysis to identify the demographic drivers of the above dependent variables.

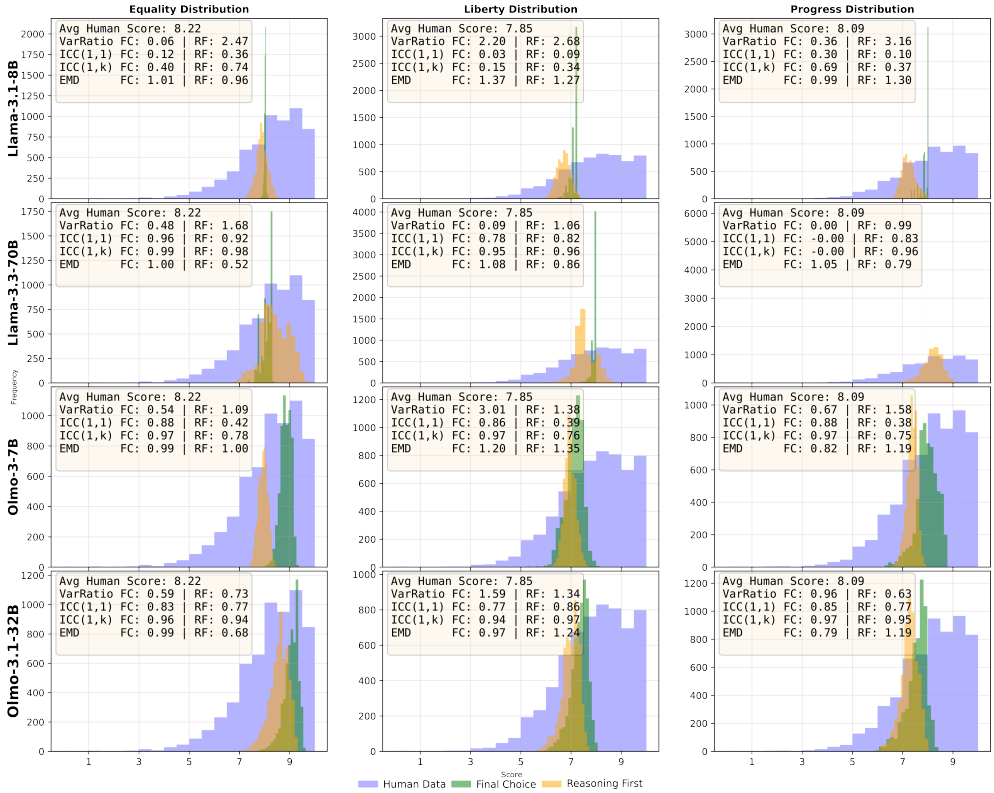


Figure 2: Score distribution between 4 LLMs in 2 Prompt Styles (FC and RF) and Human scores with metrics ICC(1,1), ICC(1,k), Wasserstian distance (EMD) and variance ratio

The regression model is specified as follows:

$$Y_{i,q} = \beta_0 + \sum_{d=1}^{D} b_d X_{i,d} + \alpha_q + \epsilon_{i,q} \qquad (1)$$

where $Y_{i,q}$ is the dependent variable (either LLM score, Abs. Error, Sign. Error or $H\%$) for respondent i and the question q pair. $X_{i,d}$ represents the d^{th} demographic feature for the respondent i, b_d is the corresponding unstandardized coefficient, α_q is the fixed effect of the value questions and $\epsilon_{i,q}$ is the error term. Categorical variables are encoded using one-hot encoding. Age and State are aggregated into groups for reducing dimensionality. Age is categorized into age_groups (18-29, 30-44, 45-64, 65+), and States are converted into regions (Northeast, Midwest, South, West). We report partial R^2 and coefficients (b) to quantify the explained variance in the outcome attributable to demographics after accounting for question fixed effects, and to quantify the magnitude and direction of the effects of demographic characteristics on the outcome. Additionally, we include the LLM's numeric output (final_choice) as a predictor of the dependent variable $H\%$ to assess whether the model's language correlates with its internal certainty. A significant effect would suggest the LLM's *persona* maintains stylistic consistency with its scores; specifically, a negative coefficient would indicate the model hedges less as its assigned scores increase.

3 Results

3.1 Behavioral Rigidity

ICC(1,1), ICC(1,k) that captures the single run and iterative reliability are shown in the Figure 3. The FCF strategy demonstrates high reliability overall. Llama-3.1-8B achieved the highest consistency, with ICC(1,1) = 0.94 and ICC(1,k) = 0.99, indicating near-perfect agreement across runs. Olmo-3-7B and Olmo-3.1-32B also showed strong reliability (ICC(1,1) = 0.96 and 0.93, respectively). Examining individual dimensions, reliability varied considerably. For the Liberty dimension, Llama-3.1-8B showed very low reliability (ICC(1,1) = 0.03). For the Equality dimension, most models performed well, with Llama-3.3-70B reaching ICC(1,1) = 0.96. The Progress dimension was more variable: Llama-3.3-70B yielded a near-zero ICC(1,1), indicating no reliable agreement, while Olmo-3-7B and Olmo-3.1-32B maintained good reliability (ICC(1,1) of 0.88 and 0.85, respectively). Under the RF strategy, overall reliability

was generally lower than in the FCF condition. Dimension-level patterns echoed those observed in the FCF condition. Overall, results indicate that the FCF prompting strategy yields higher inter-rater reliability than the RF strategy across models and dimensions. Larger models (Llama-3.3-70B, Olmo-3.1-32B) tend to produce more consistent scores regardless of strategy.

EMD, and variance ratio for each model, prompt style, and value dimension are shown in Figure 2. The prompt style affects the LLama and Olmo models differently. With FCF Llama-3.1-8B and Llama-3.3-70B has the highest EMD of 1.12 and 1.04. In contrast, Olmo-3.1-32B (EMD = 1.01) and Olmo-3-7B (EMD = 0.0.92) showed considerably lower overall deviation. With RF, the pattern is partially reversed. Llama-3.3-70B showed a marked 30% reduction in overall EMD (EMD = 0.72) compared to its FCF value. The Olmo models, however, increased with RF: Olmo-3-7B (EMD = 1.18) and Olmo-3.1-32B (EMD = 1.03) both rose substantially from their FCF values. For a breakdown by model family, construct, and prompt style, refer to the Appendix Tables 6 to 10.

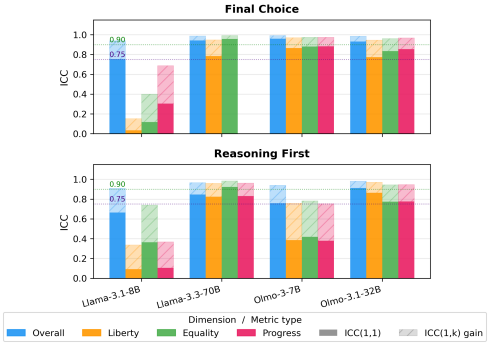


Figure 3: ICC(1,1) and ICC(1,k) where k=5 showing the overall and invidual dimension reliability

3.2 Regression

The regression analysis in Table 1 reveals that demographic variables explain LLM score outcomes at an average of 68% (range: 38%–91% across models and prompt combinations), compared to just 9% for human responses. For the outcome $H\%$, demographic variables explain on average 71% of the variation (range: 48%–89%). Absolute Error (15%) and Signed Error (21%) are more modestly explained by demographics alone. Across all models and prompt styles, predictors that appear in the top 20 in at least 6 out of 8 models and prompt combinations are included in 1. Political affiliation emerged as a consistent and substantively meaningful predictor across all four outcomes. Compared to Democrats, Republicans received lower LLM Scores, higher Hedge%, and greater Abs. Error, suggesting that LLMs assign less favorable scores, hedge more frequently, and deviate more from human scores when evaluating for a republican persona. Being a registered voter was associated with higher LLM Scores ($\beta = .23$, 100%), lower Hedge%, and lower Abs. Error ($\beta = -.13$, 100%). Employment status showed consistent effects. Unemployed respondents received substantially lower LLM Scores ($\beta = -.24$, 100%), higher Hedge%, and greater Abs. Error ($\beta = .14$, 100%). Part-time employed respondents similarly received lower LLM Scores. Education effects were notably uneven. Notably, respondents with only a grade-school education exhibited the largest coefficient Abs. Error and Sign. Error ($\beta = .45$, 100%;$\beta = .31$, 100%), despite not appearing to be consistent predictors of LLM Score or Hedge%. High-school educated demographic received lower LLM Scores and greater Absolute Error, but negative Signed Error ($\beta = -.22$, 100%), indicating systematic underestimation relative to human scores. Non-binary respondents showed substantially higher Abs. Error and Sign. Error ($\beta = .26$, 100%;$\beta = .33$, 100%), suggesting more distributional deviation in model responses, while transgender respondents were associated with notably over estimation in Sign. Error ($\beta = .53$, 100%) and lower Hedge% ($\beta = -.16$, 100%). Compared to the reference youngest age group (18–29), older respondents (30–44, 45–64, 65+) received lower Sign. Error scores indicating under estimation ($\beta = -.34$, 100%; $\beta = -.43$, 100%; $\beta = -.57$, 100%). Age was not a consistent predictor of Hedge%.

4 Discussion

4.1 Demographic Influence on Model Size and Prompt Style

Model size and prompt style interact — larger models (Llama-3.3-70B, OLMo variants) achieve good reliability (ICC(1,1) and ICC(1,k)) under FCF, but only larger models maintain acceptable reliability under RF. Llama-3.1-8B fails the reliability threshold under RF on most dimensions and should be treated as unsuitable for consistent scoring regardless of prompt style. Both OLMo-3-7B and OLMo-3.1-32B sit above 0.75 on most dimensions under

Table 1: Consistent significant demographic predictors of LLM response outcomes and their average coefficients

Predictor	**LLM Score**	**Abs. Error**	**Sign. Error**	**Hedge%**
R^2 *[min,max]*	*.68 [.38,.91]*	*.15 [.04,.22]*	*.21 [.06,.33]*	*.71 [.48,.89]*
Political affiliation (ref: Democrat)				
Republican	$-.12^{*}$ (88%)	$.13^{*}$ (100%)	$.01^{*}$ (100%)	$.01^{*}$ (100%)
Independent	$-.08^{*}$ (100%)	$.09^{*}$ (100%)	—	$.05^{*}$ (75%)
Other	$-.10^{*}$ (100%)	$.22^{*}$ (100%)	$.13^{*}$ (88%)	—
Voter registration				
Registered (Yes)	$.23^{*}$ (100%)	$-.13^{*}$ (100%)	—	$-.04^{*}$ (88%)
Employment (ref: Full-time)				
Unemployed	$-.24^{*}$ (100%)	$.14^{*}$ (100%)	—	$.05^{*}$ (75%)
Part-time	$-.11^{*}$ (100%)	—	—	$.03^{*}$ (100%)
Other	—	—	—	—
Education (ref: Bachelor's)				
Grade school	—	$.45^{*}$ (100%)	$.31^{*}$ (88%)	—
High school	$-.09^{*}$ (88%)	$.12^{*}$ (100%)	$-.22^{*}$ (100%)	$.01^{*}$ (75%)
Some coll/trade	—	$.06^{*}$ (88%)	$-.15^{*}$ (100%)	$.04^{*}$ (100%)
Graduate/Prof.	$.07^{*}$ (100%)	—	—	$.03^{*}$ (88%)
Race (ref: Other/non-listed)				
White/Cauc.	$.08^{*}$ (88%)	—	—	$.07^{*}$ (88%)
Asian	$.01^{*}$ (88%)	—	$-.26^{*}$ (100%)	$.06^{*}$ (75%)
Native American	$-.09^{*}$ (88%)	$.22^{*}$ (100%)	—	—
Black/Afr. Amer.	—	—	$-.20^{*}$ (100%)	—
Age group (ref: 18–29)				
30–44 years	—	$-.06^{**}$ (88%)	$-.34^{*}$ (100%)	—
45–64 years	$-.01^{*}$ (100%)	$-.12^{*}$ (100%)	$-.43^{*}$ (100%)	—
65+ years	$-.01^{*}$ (100%)	$-.16^{*}$ (100%)	$-.57^{*}$ (100%)	—
Gender (ref: Female)				
Male	$-.01^{*}$ (88%)	—	—	$.01^{*}$ (88%)
Non-binary	—	$.26^{*}$ (100%)	$.33^{*}$ (75%)	—
Transgender	—	—	$.53^{*}$ (100%)	$-.16^{**}$ (100%)
Hispanic ethnicity				
Hispanic (Yes)	—	—	$-.14^{*}$ (100%)	$-.04^{*}$ (75%)

Positive coefficients, Negative coefficients $^{*}p < .001$. $^{**}p < .01$.
(%) indicates occurance of the predictor in top 20 significant predictor for the 8 cases (4 Models and 2 Prompt types).
— not in top 20 in <75% or 6 out of the 8 cases.

FCF, and largely above 0.75 on Overall under RF 2. They're the most reliable scorers across both prompt styles.

Forcing the model to generate a "Chain of Thought" introduces additional context beyond the direct lookup of stereotypes and lowers the artificially high R^2. However, even with context, the LLM R^2 remains roughly 7x higher than the human baseline (See Appendix Table 2), indicating the bias is deep-seated.

LLaMA models exhibit a clear scaling effect, with the larger 70B model showing lower R^2 values than the smaller 8B model across both prompt styles, suggesting that larger models allow for greater response variance. The smaller OLMo 7B shows an extremely high R^2 (0.86–0.96), indicating responses are almost entirely driven by demographic inputs. While the larger OLMo 32B reduces this determinism under the FCF prompt style, it does not do so consistently under the RF prompt, suggesting that chain-of-thought style prompting may lead models to construct explicit demographic rationales, reinforcing rather than moderating stereotype-driven responses. Taken together, these patterns suggest that LLMs broadly simulate demographic archetypes rather than individual variation, with model size offering only partial mitigation. This tendency is most pronounced for specific constructs: for the Liberty scale, LLaMA-3.1-8B reached an R^2 of 0.96 (Refer Appendix Table 2), indicating that for questions involving freedom and government constraint, smaller models collapse almost entirely into stereotypical responses, leaving virtually no room for within-group variance.

4.2 Dimension effect

Liberty emerged as the most divergent dimension between human respondents and LLMs, with EMD values peaking at 1.17 and the highest variance ratio of 1.67 (Figure 2). Equality generally yielded the lowest EMD values (0.89) and variance ratio (0.96), particularly for larger models within the same family. Progress showed intermediate EMD (1.02) and variance ratio (1.05), though notable spikes were observed for Llama-3.1-8B (EMD = 1.30 under RF), suggesting that reasoning strategies do not uniformly reduce distributional bias and may amplify it for certain model-dimension combinations. While Llama-3.3-70B demonstrated the highest overall alignment under RF (EMD = 0.72), models such as Olmo-3.1-32B showed that reasoning can exacerbate divergence in specific dimensions, particularly Liberty and Progress. The lack of inferential variation across dimensions raises concerns about adopting a single model family or prompt style with confidence.

4.3 Hedging

Hedging is a known LLM tendency to avoid commitment, which would naturally distort survey simulation. The variation in hedging across demographics was an unexpected result, yet has semblance to the findings from (Santurkar et al., 2023). The bias was most pronounced across a few demographics and select questions. Whenever the model is prompted with "Non-binary" or "Transgender," it immediately enters a "Safety Mode" characterized by high hedging. This is evident from the strong negative coefficients for these predictors in the hedging regression. This could be due to the post-training safety fine-tuning that penalizes the model for making strong statements about marginalized identities, leading to a default response of hedging when these identities are present in the input. This could highlight the cautionary process, which differs from typical model behavior arising from under-representation in the training data or confusion about the topic. Instead, it reflects a deviant behavior often associated with sensitive topics in

fairness and safety research. The model's response is not necessarily driven by confusion about the topic itself, but rather by a learned behavior to avoid making definitive statements about certain identities, which results in increased hedging. The quality of synthetic data is distorted when LLM output is influenced by functions such as "Safety Mode".

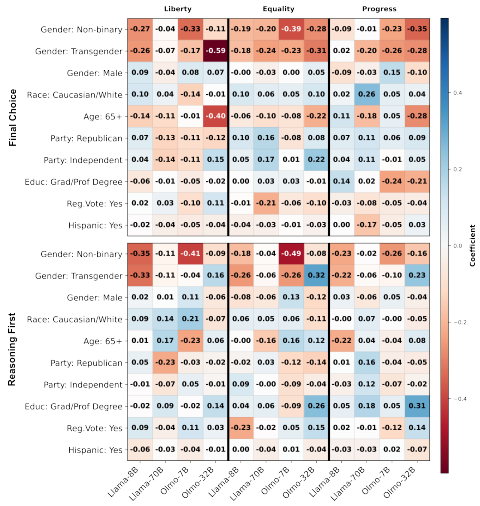


Figure 4: Hedging predictors by demographics. Including significant (p<0.05) predictors excluding states.

4.4 Demographic Caricature

LLMs act as mean-finders, generalizing toward the average member of the population. Our results from Table 1 suggest the WEIRD claim is only partially justified for educated individuals . Similar to (Santurkar et al., 2023), the response caricatures are aligned well with liberal and educated. Additionally, the LLM does not follow the edges of opinion but consistently seeks the center. Despite the inclusion of demographic context in the artificial persona, LLMs remain fixated on certain perceived monoliths (Unemployed, Grade School, Non-binary, and Transgender). The models mostly ignore the other demographic traits that make up the intersectionality of the individual, thereby restricting potential variation in synthetic output.

The results suggest that the LLM treats "Registered Voter" as a proxy for "Good/High Value Citizen," mechanically boosting the scores across the board. LLMs underweight the environmental and longitudinal factors like geography and age, opting instead to amplify "identity signals" like voter registration or employment status that align with the most frequent and often most biased patterns in their training sets. Individual constructs show pronounced stereotypes. Regarding the Progress construct, the demographic indicator educ_Grade school emerges as a substantial negative driver (coefficients ranging from -0.43 to -0.50). The model appears to use "low education" as a heuristic, effectively decreasing scores on complex topics such as economics and infrastructure. Similarly, partyid_Republican serves as a dominant negative predictor for Equality ($\beta = -0.40$)). While empirical human data often show lower scores for this group, the model fails to replicate the substantial internal variance found in human populations. Instead, the LLM flattens the diverse spectrum of Republican thought—from Libertarianism to Populism—into a monolithic "Anti-Equality" coefficient. This highlights a critical failure in representational fidelity: the model lacks the granular capacity to distinguish between specific ideological motivations, such as fiscal opposition to taxes versus social opposition to equity programs, reinforcing the findings of (González Barman et al., 2025) in an experimental setting focused on eliciting diverse opinions.

5 Conclusion

This study provides both a procedure and substantive results of evaluating the performance of select open-source LLMs as synthetic respondents in survey research. While LLMs offer ample experimental opportunities and methods for generating hypotheses, the results of this study suggest that they are currently inadequate substitutes for data collected from human respondents. On the one hand, the results generated by the different models and prompting strategies were highly consistent across iterations, whereas the substantive output from the artificial data was less robust. The synthetic data from the models and prompting styles used resulted in extremely narrow variances in the data that fail to capture the gradation of actual human responses. Beyond the numerical results, the reasoning provided by the LLMs constituted a noncommittal, unclear rationale that did not consistently support their answers to the question. We approach the adoption of LLM respondents in public opinion survey data with deep skepticism, positing that what appears to be human-like opinion is often just a highly probable text completion, stripped of genuine human nuance. If researchers indiscriminately adopt these tools and incorporate synthetic data in published results, we risk measuring a distorted version of training data and other noise rather than

dynamic public sentiment.

Limitations

The aggregate results and questions associated with this study were made publicly available in 2021 and may have been included in the models' training data, potentially contaminating the synthetic data and confounding its quality. This risk could not be directly measured or controlled for in our analysis. Furthermore, our prompt does not explicitly account for the effect of this 2021 data, which was not included in the replication survey. The use of a ten-point integer scale (0–10) is another limitation, as it differs from more conventional response formats such as binary, four-point, or three-point scales, which may limit the generalizability of our evaluation metrics. Another limitation identified during the analysis was the conversion of respondents' age values from integers to age groups for regression analysis, which limits the specificity for making claims. Additionally, we don't isolate the post-training effect by comparing our Instruct models with their base-model counterparts, which induces a noticeable performance shift, as noted by (Santurkar et al., 2023). We encourage researchers to consider these elements in future replications of this study.

6 Acknowledgments

We acknowledge the support of Siena Research Institute in allowing us to use the American Values Survey. We acknowledge the support of Social Science Automation in facilitating the presentation of this research. We thank Eddie Smith and Pierce Johnson for their valuable observations on the LLM output responses. We thank the reviewers for pointing out related works that strengthened our findings.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Yash Bhatnagar. 2026. Breaking the turing test: Testing the relevance of the turing test against modern llms. *International Journal for Research in Engineering Application Management*, pages 1–.

J Michael Brick, Pat D Brick, Sarah Dipko, Stanley Presser, Clyde Tucker, and Yangyang Yuan. 2007. Cell phone survey feasibility in the us: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71(1):23–39.

Mario Callegaro and Charles DiSogra. 2008. Computing response metrics for online panels. *Public opinion quarterly*, 72(5):1008–1032.

Mick P Couper and Peter V Miller. 2008. Web survey methods: Introduction. *Public opinion quarterly*, 72(5):831–835.

Rachel E Davis, Timothy P Johnson, Sunghee Lee, and Christopher Werner. 2019. Why do latino survey respondents acquiesce? respondent and interviewer characteristics as determinants of cultural patterns of acquiescence among latino survey respondents. *Cross-Cultural Research*, 53(1):87–115.

Jonathan Eggleston. 2024. Frequent survey requests and declining response rates: evidence from the 2020 census and household surveys. *Journal of Survey Statistics and Methodology*, 12(5):1138–1156.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Anna Foka, Gabriele Griffin, Dalia Ortiz Pablo, Paulina Rajkowska, and Sushruth Badri. 2025. Tracing the bias loop: Ai, cultural heritage and bias-mitigating in practice. *AI & SOCIETY*, 40(8):5835–5847.

Mansour Ghafourifard. 2024. Survey fatigue in questionnaire based research: The issues and solutions. *Journal of caring sciences*, 13(4):214–215.

Chris Gibson and Daniel Lipinski. 2021. Americans, deeply divided, yet share core values of equality, liberty progress.

Kristian González Barman, Simon Lohse, and Henk W de Regt. 2025. Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives? k. gonzález barman et al. *Philosophy & Technology*, 38(2):35.

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*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard

Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven-country-pooled datafile version 5.0.

Yoichi Hayashi. 2024. Prospects for revolutionary and popular ai technology following the launch of chatgpt in 2023.

Morten Hjortskov, Christian Bøtcher Jacobsen, and Anne Mette Kjeldsen. 2023. Choir of believers? experimental and longitudinal evidence on survey participation, response bias, and public service motivation. *International Public Management Journal*, 26(2):281–304.

Jumayel Islam, Lu Xiao, and Robert E Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can Language Models Reason about Individualistic Human Values and Preferences? *Preprint*, arXiv:2410.03868.

Timothy P Johnson, Henning Silber, and Jill E Darling. 2024. Public perceptions of pollsters in the united states: experimental evidence. *Social Science Quarterly*, 105(1):114–127.

Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. 2024. Synthetic data in human analysis: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4957–4976.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. *arXiv preprint*. ArXiv:2205.11916 [cs].

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236.

Jon A Krosnick, Allyson L Holbrook, Matthew K Berent, Richard T Carson, W Michael Hanemann, Raymond J Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A Ruud, V Kerry Smith, and 1 others. 2002. The impact of" no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66(3):371–403.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lujun Li, Lama Sleem, Geoffrey Nichil, Radu State, and 1 others. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251.

Michael W Link, Ali H Mokdad, Dale Kulp, and Ashley Hyon. 2006. Has the national do not call registry helped or hurt state-level response rates? a time series analysis. *International Journal of Public Opinion Quarterly*, 70(5):794–809.

Michael W Link and Robert W Oldendick. 1999. Call screening: Is it really a problem for survey research? *The Public Opinion Quarterly*, 63(4):577–589.

Andy Markus. 2025. AT&T tests new AI digital receptionist. https://about.att.com/blogs/2025/ai-digital-receptionist.html.

Robert W Oldendick and Michael W Link. 1994. The answering machine generation: who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58(2):264–273.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1132–1144.

David C Pyrooz, James A Densley, and Jose Antonio Sanchez. 2025. Are online opt-in panels viable data sources on hard-to-reach populations? population and relational inferences on gang membership in the united states. *International Criminology*, pages 1–17.

Patric M Reinbold. 2020. Taking artificial intelligence beyond the turing test. *Wis. L. Rev.*, page 873.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Jorre Vannieuwenhuyze, Geert Loosveldt, and Geert Molenberghs. 2010. A method for evaluating mode effects in mixed-mode surveys. *Public opinion quarterly*, 74(5):1027–1045.

Michael F Weeks, Richard A Kulka, and Stephanie A Pierson. 1987. Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, pages 540–549.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yu Xie and Yueqi Xie. 2025. Variance reduction in output from generative ai. *arXiv preprint arXiv:2503.01033*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025. Should llms be weird? exploring weirdness and human rights in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2808–2820.

A Appendix

A.1 American Values Survey Data

The American Values Survey was conducted by Siena College Research Institute (SCRI) in 2021. The respondents were from 50 states and the District of Columbia. Respondents were asked how much they agree with 34 value statements covering core American values on Liberty, Equality, and Progress. For our experiment we only kept questions related to the three value constructs and removed questions not relevant to the values. The survey was administered online, and initial attention checks were conducted to ensure high-quality responses. Irrespective of demographics and inclinations, human respondents express a core sense of embodying American values. This dataset offers a unique opportunity to evaluate LLMs, as it combines rich demographic information with clearly demarcated group identities that, despite strong internal affiliations, are united by a shared commitment to a nation's core values. The actual value statements are shown in the Table 3.

A.2 Prompting LLMs

The LLM responses were forced to be in a structured format $JSON$ to facilitate downstream analysis of the reasoning traces and final scores. The detailed prompts are provided in Table 4. The prompts are also similar to the actual survey questions, with only the input and output instructions appended at the beginning and end of the survey text for validity. The actual survey did not collect reasoning traces, so we do not have human ground truth for the reasoning component. Since our experiment requires large-scale inference, we use the VLLM library (Kwon et al., 2023). To generate multiple responses to the same prompt, we use a temperature of $\tau = 0.1$, since the ideal deterministic temperature $\tau = 0$ doesn't conveniently allow multiple iterations for the same input prompt through the VLLM library. We dont limit the number of tokens generated by model. Apart from the temperature $k = 5$ i.e. number of samples per input is the only hyperparameter we modified. Rest of the values were kept default. All four LLMs are loaded from Hugging Face's official model weight collection. We don't use any quantization. All the models are run on NVIDIA A100 GPUs. We use 2 GPUs for the smaller 7B and 8B models and use 4 GPUs for the larger 32B and 70B models.

A.3 Evaluation Metrics

We tabulate the metrics used to compare the LLM responses with human responses. The ideal and poor values for each metric and their resultant interpretations are tabulated for easier understanding in Table 5. Behavioral consistency is a prerequisite for alignment evaluation — only models with sufficient ICC reliability are meaningfully assessed on alignment quality metrics. They could indicate consistent distortion or consistent alignment. A high ICC, alongside a low σ^2, is a signature of consistency and confidence, treating the demographic profile as a fixed caricature rather than an opinion distribution. We include a summary of the 5 metrics (ICC(1,1), ICC(1,k), variance ratio, EMD, and Hedging%) used in the experiment, broken down by Model, Construct, and prompt styles.

Table 2: Mean R^2 [min, max] by scope.

Scope	LLM	Abs.Error	Sign.Error	Hedge%	Human
Overall	.68 [.38, .91]	.15 [.04, .22]	.21 [.06, .33]	**.71** [.48, .89]	.09
By value construct:					
Liberty	**.76** [.22, .96]	.21 [.03, .38]	.26 [.04, .46]	.61 [.28, .87]	.06
Equality	.68 [.45, .91]	.08 [.03, .16]	.12 [.06, .30]	**.72** [.39, .93]	.12
Progress	.67 [.00, .90]	.10 [.02, .21]	.17 [.05, .38]	.62 [.45, .83]	.07

Note. Mean R^2 across dimensions; [min, max] in brackets. Abs.Error is Absolute Error and Sign.Error is Signed Error. Hedge% is where the average hedging token present in the reasoning text is the target. Human R^2 has zero variance within each construct. **Bold** indicates $R^2 \geq 0.70$.

Label	Value Statement
Equality_1	All people are equal, regardless of race, ethnicity, gender, physical appearance, or any other personal characteristic.
Equality_2	Treat others as you would like them to treat you.
Equality_3	No one is above the law.
Equality_5	The religious beliefs and practices of all people should be both protected and respected.
Equality_6	Any injustice to a single person is an injustice to all.
Equality_7	We are all, all of us, in this life together and we should look out for the well-being of everyone else.
Equality_8	People may disagree but that is no excuse for being disagreeable.
Equality_9	No person is complete if they do not give of themselves in service to others.
Equality_10	Before making a judgement about someone else, try to walk a mile in their shoes.
Equality_11	In order for us all to live together, each of us has to make concessions.
Equality_12	Not everyone starts off with the same set of tools or skills, sometimes we need to level the playing field by giving some people a head start.
Equality_13	Because we only have one planet, protecting our environment is a priority.
Equality_14	Steps must be taken to protect people from those who lie and cheat.
Equality_15	Each of us should have an equal chance to be successful.
Equality_16	Every American has the right and responsibility to vote.
Liberty_1	No one, not even the government, should be able to restrict another's pursuit of happiness.
Liberty_4	The benefits of investing capital and hard work rightfully belong to the entrepreneur that accepted the risk.
Liberty_5	You only live once: seek to experience all that life has to offer.
Liberty_6	No one should tell me how to live, how to love or what to think.
Liberty_11	Each of us has the power to pull ourselves up by our bootstraps, that is, to take control of our own destiny.
Liberty_12	What you know is more important than who you know.
Liberty_13	Each of us is free to follow our own unique path in life.
Liberty_14	Stand on your own two feet without reliance on other people, organizations or the government.
Liberty_15	Owe nothing and expect nothing in return.
Liberty_16	Everyone can speak their mind in public regardless of the viewpoint without fear of punishment.
Progress_1	Success comes to those that dedicate themselves to making the most of their abilities.
Progress_2	Advances in areas like health, technology, business, or personal development, rely on the careful application of science.
Progress_3	Give a person a fish, and you'll feed them for a day. Teach a person to fish, and you've fed them for a lifetime.
Progress_4	A penny saved is a penny earned.
Progress_5	Having specific goals, whether those goals involve personal, economic, artistic or societal achievements, is the best way to proceed in life.
Progress_6	It is important to achieve something specific and measurable each and every day.
Progress_7	Every problem has a solution.
Progress_8	Tomorrow always holds the possibility of being a better day.
Progress_9	Time is money.

Table 3: Survey constructs and full value statements used in the evaluation.

Final Choice First (FCF) Prompt Template

Your demographic information is given in JSON format
{'state': 'Rhode Island', 'gender': 'Female', 'age': 60.0, 'hispanic': 'No', 'race': 'Caucasian/White', 'education': 'Graduate or Professional degree', 'registered to vote': 'Yes', 'party id': 'Democrat', 'employment': 'Employed full-time'}

First, you will see a statement that may be familiar to you. Take a moment and consider it in terms of how you live your life. Are these words that you live by? All the time? Or some of the time? Would others say that you embody these statements, that your actions, or that the things you say are a reflection of these words often, sometimes, not very often, or perhaps not at all? We are all different people. Some statements may be completely us, others partially, and some not us at all. Try as best as you can to evaluate yourself, your thoughts, your actions, as well as your beliefs on each statement. As you evaluate yourself on each statement, you can score yourself anywhere between 0 and 10.

0: You disagree with the statement, are not guided by it, and no one would ever say that the statement reflects how you live your life.
5: The statement may be one that you endorse, but you don't always live your life with it in mind.
10: You believe the statement, and not only aspire to live that way, but you do.

STATEMENT:
Each of us is free to follow our own unique path in life.

INSTRUCTION:
Return ONLY valid JSON with exactly these keys:
- "final_choice": integer (0–10)
- "reasoning": a first-person justification for how you arrived at your score. Do NOT write placeholders like "..." or "[insert reasoning here]".

OUTPUT:

Reasoning First (RF) Prompt Template

Your demographic information is given in JSON format
{'state': 'Rhode Island', 'gender': 'Female', 'age': 60.0, 'hispanic': 'No', 'race': 'Caucasian/White', 'education': 'Graduate or Professional degree', 'registered to vote': 'Yes', 'party id': 'Democrat', 'employment': 'Employed full-time'}

First, you will see a statement that may be familiar to you. Take a moment and consider it in terms of how you live your life. Are these words that you live by? All the time? Or some of the time? Would others say that you embody these statements, that your actions, or that the things you say are a reflection of these words often, sometimes, not very often, or perhaps not at all? We are all different people. Some statements may be completely us, others partially, and some not us at all. Try as best as you can to evaluate yourself, your thoughts, your actions, as well as your beliefs on each statement. As you evaluate yourself on each statement, you can score yourself anywhere between 0 and 10.

0: You disagree with the statement, are not guided by it, and no one would ever say that the statement reflects how you live your life.
5: The statement may be one that you endorse, but you don't always live your life with it in mind.
10: You believe the statement, and not only aspire to live that way, but you do.

STATEMENT:
Each of us is free to follow our own unique path in life.

INSTRUCTION:
Return ONLY valid JSON with exactly these keys:
- "reasoning": a first-person justification for how you arrived at your score. Do NOT write placeholders like "..." or "[insert reasoning here]".
- "final_choice": integer (0–10)

OUTPUT:

Table 4: FCF and RF prompt templates. *Italicised* text denotes the demographic precursor prepended to the prompt. The only difference between the two styles is the order of the JSON keys in the instruction.

Metric	Range / Value	Interpretation
ICC	Ideal: Closer to 1	The model's responses demonstrate high stability and are consistent and reliable for that specific persona.
	Poor: Values below 0.5	The model is generating inconsistent responses and appears to be sensitive to random noise in the prompt.
EMD	Ideal: Closer to 0	The model's score distribution shows high representational accuracy and closely matches human score distributions.
	Poor: Higher values	The model score distribution is significantly distorted from that of the human group's actual recorded values.
Variance Ratio	Ideal: Near 1.0	The model differentiates between personas to the same natural extent that humans do in real-world data.
	Poor: Below 1 or Above 1	The model either gives similar scores regardless of demographics or exaggerates differences across those demographics.
Hedging%($H\%$)	Ideal: Lower (Contextual)	The model provides a clear and decisive stance with very little ambiguity regarding its reasoning.
	Poor: High percentages	The model displays high uncertainty and is playing it safe to avoid taking a firm stance on the topic.

Table 5: LLM Evaluation Metrics for demographic consistency and alignment with human respondents on public opinion survey

Table 6: **Intraclass Correlation Coefficient (ICC(1,1)) Reliability Summary**
Higher ICC(1,1) values indicate greater single-run reliability across simulated responses. Values near 1.0 suggest highly uniform (potentially homogenized) outputs across individual ratings. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. ICC(1,1)
By Model	
Olmo-3.1-32B	**0.84**
Llama-3.3-70B	0.76
Olmo-3-7B	0.69
Llama-3.1-8B	0.30
By Construct	
Equality	**0.66**
Liberty	0.58
Progress	0.52
By Reasoning Style	
Final Choice	**0.68**
Reasoning First	0.62

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.3-70B	0.67	0.86
Olmo-3.1-32B	0.85	0.83
Olmo-3-7B	**0.90**	0.48
Llama-3.1-8B	0.30	0.31

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.3-70B	**0.94**	0.80	0.41
Olmo-3.1-32B	0.80	0.82	0.81
Olmo-3-7B	0.65	0.62	0.63
Llama-3.1-8B	0.24	0.06	0.20

Table 7: **Intraclass Correlation Coefficient (ICC(1,k)) Reliability Summary**
Higher ICC(1,k) values denote greater internal consistency and reliability among profiles for average simulated responses. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. ICC(1,k)
By Model	
Olmo-3.1-32B	**0.96**
Olmo-3-7B	0.89
Llama-3.3-70B	0.85
Llama-3.1-8B	0.57
By Construct	
Equality	**0.85**
Liberty	0.75
Progress	0.71
By Reasoning Style	
Reasoning First	**0.83**
Final Choice	0.80

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.3-70B	0.73	**0.97**
Olmo-3.1-32B	0.96	0.96
Olmo-3-7B	0.98	0.81
Llama-3.1-8B	0.54	0.59

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.3-70B	**0.99**	0.95	0.48
Olmo-3.1-32B	0.95	0.96	0.96
Olmo-3-7B	0.88	0.86	0.86
Llama-3.1-8B	0.57	0.24	0.53

Table 8: **Variance Ratio Summary**
Variance Ratio of the LLM responses to human baseline. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. Var Ratio
By Model	
Llama-3.1-8B	**1.83**
Olmo-3-7B	1.42
Olmo-3.1-32B	1.05
Llama-3.3-70B	0.73
By Construct	
Liberty	**1.67**
Progress	1.05
Equality	0.96
By Reasoning Style	
Reasoning First	**1.60**
Final Choice	0.92

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	0.89	**2.78**
Olmo-3-7B	1.47	1.36
Olmo-3.1-32B	1.13	0.97
Llama-3.3-70B	0.20	1.27

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	1.26	**2.44**	1.76
Olmo-3-7B	0.82	2.19	1.13
Olmo-3.1-32B	0.66	1.47	0.80
Llama-3.3-70B	1.08	0.57	0.50

Table 9: **EMD Summary**
EMD between LLM responses and reference human distribution. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. EMD
By Model	
Llama-3.1-8B	**1.15**
Olmo-3-7B	1.09
Olmo-3.1-32B	0.98
Llama-3.3-70B	0.88
By Construct	
Liberty	**1.17**
Progress	1.02
Equality	0.89
By Reasoning Style	
Reasoning First	**1.03**
Final Choice	1.02

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	1.12	**1.18**
Llama-3.3-70B	1.04	0.72
Olmo-3-7B	1.01	1.18
Olmo-3.1-32B	0.92	1.03

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	0.99	**1.32**	1.14
Olmo-3-7B	1.00	1.27	1.01
Olmo-3.1-32B	0.83	1.11	0.99
Llama-3.3-70B	0.76	0.97	0.92

Table 10: **Hedging Percentage Summary**
Hedging % measures the proportion of hedging tokens in the generated reasoning traces indicating uncertain or non-committal language. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. Hedging %
By Model	
Llama-3.1-8B	**6.73**
Llama-3.3-70B	6.51
Olmo-3-7B	6.50
Olmo-3.1-32B	5.22
By Construct	
Equality	**6.59**
Liberty	6.14
Progress	5.94
By Reasoning Style	
Final Choice	**6.73**
Reasoning First	5.75

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	**6.94**	6.52
Llama-3.3-70B	6.94	6.08
Olmo-3-7B	6.23	6.77
Olmo-3.1-32B	6.82	3.62

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	7.10	**7.24**	5.77
Llama-3.3-70B	7.09	6.07	6.30
Olmo-3-7B	6.81	6.34	6.32
Olmo-3.1-32B	4.89	4.89	5.36

Documenting Corporate Harm: A Semantic Action Trajectories Approach to the Opioid Industry Document Archive Shared Task

Benjamin Miller
University of Canterbury
Ōtautahi Christchurch, Aotearoa New Zealand
benjamin.miller@canterbury.ac.nz

Abstract

This paper presents a method for modeling change in the possibility space of actors over time as represented in the Opioid Industry Document Archive (OIDA). The approach treats documents as a structured field of actor–action relations and models these relations as *semantic action trajectories* across time. Semantic role labeling (SRL) using the Emory Language and Information Toolkit (ELIT) is applied to extract subject–predicate structures from a corpus of internal industry documents. Subjects are normalized and grouped into actor categories using a combination of rule-based heuristics and constrained language model adjudication. Predicate vocabularies associated with these actors are mapped to psycholinguistic categories using the LIWC lexicon, and random forest feature selection with principal component analysis is used to construct a low-dimensional representation of discourse structure across periods.

The resulting discourse space reveals systematic shifts in how corporate actors, regulators, clinicians, and patients are positioned over time. In particular, corporate entities and the opioid products they produce follow nearly identical semantic trajectories, suggesting that companies and drugs occupy interchangeable roles in the archive's discourse. This method provides a way to analyze changing institutional behavior at scale across heterogeneous litigation and historical archives.

1 Introduction

Large litigation archives provide an unusually detailed record of institutional communication and practice. The Opioid Industry Document Archive (OIDA) contains millions of previously undisclosed corporate documents produced during litigation concerning the opioid crisis. An additional, larger Public Document Repository, mandated by orders issued in the Purdue Pharma bankruptcy proceedings in the United States Bankruptcy Court for the Southern District of New York promises to expand these holdings related to one of the most significant breaches of the public trust in US history (Vadivelu et al., 2018). These documents describe the marketing, sales, research and development, compliance, and regulatory details, along with call notes, procedural descriptions, and trial material that underlie a crisis that nearly tripled the reported drug overdose death rate in the US (Vadivelu et al., 2018). In scope, the court order describes more than 100 million pages of material to be added to a public repository. The OIDA materials and this yet to be released repository provide an important evidentiary record describing an institutional project that contributed to a global health crisis. However, their scale and heterogeneity make systematic analysis difficult.

This paper introduces a computational method for analyzing changes in actor behavior over time as represented by discourse in the OIDA corpus. The central premise is that a heterogeneous corpus like OIDA can be modeled as a set of structured relations linking actors and actions at times. By extracting subject–predicate structures at time from text and aggregating them into subject groups, it becomes possible to model the changing possibility space of actor groups. Following (Mehran et al., 2025), a possibility space refers to the set of allowable actions associated with a subject. In this study, that space is operationalized as the predicates attached to a subject group.

The method builds on prior work applying computational analysis of subject–predicate relations to ideological and institutional discourse (Mehran et al., 2025). More broadly, computational approaches have been used to model narrative and discourse structure across large document collections (Miller et al., 2015). Here we extend those approaches in two directions. First, we model discourse diachronically by associating actor–action relations with historical periods. Second, we intro-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 149–158

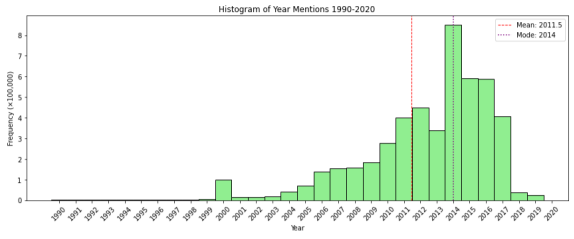


Figure 1: Distribution of year mentions extracted from the corpus. The sample is heavily concentrated in the litigation period of the opioid crisis, with a mean referenced year of 2011.5 and a modal year of 2014.

duce an automated actor grouping procedure combining rule-based classification with constrained large language model adjudication.

The result is a representation of institutional discourse as a set of *semantic action trajectories* that describe how actors move through a conceptual discourse space over time.

2 Task and Data

The NLP+CSS shared task focuses on computational analysis of the Opioid Industry Document Archive. The archive includes internal corporate communications, regulatory materials, litigation documents, and investigative records produced during the development of the opioid crisis.

For this study, a corpus of 10,000 OCR documents was randomly sampled from the public Opioid Industry Documents Archive (OIDA), a repository of corporate documents released through opioid-related litigation (Alexander et al., 2022). This sampled sub-corpus contains approximately 238 million tokens and includes heterogeneous document types such as emails, reports, legal transcripts, and regulatory submissions. The sample was constructed through repeated random traversal of the archive structure, with each traversal yielding one document. The resulting sub-corpus reflects the temporal, topical, and source distributions inherent in OIDA.

Temporal references extracted from documents show a strong concentration in the later years of the archive. Figure 1 shows the distribution of year mentions across the corpus.

Documents were grouped into six temporal periods spanning 1980–2019.

3 Method

After sampling, analysis proceeds in four stages: predicate extraction, actor grouping, semantic feature construction, and discourse space modeling.

Statistic	Value
Predicate instances	246,073
Subject–predicate pairs	184,831
Normalized subject expressions	52,305
Unique predicates	4,548

Table 1: Summary statistics from the SRL predicate extraction stage.

3.1 Predicate Extraction

Semantic role labeling (SRL) has long been used to extract predicate–argument structures from text (Gildea and Jurafsky, 2002). For this project, SRL was applied using the Emory Language and Information Toolkit (ELIT) (He et al., 2021). The pipeline performs tokenization, part-of-speech tagging, dependency parsing, and semantic role extraction.

While recent work increasingly relies on transformer-based representations, the present approach intentionally uses a non-neural SRL implementation to produce explicitly structured actor–action relations. This choice prioritizes interpretability and analytical transparency: subject–predicate structures can be directly aggregated into actor-level distributions and inspected without post hoc probing or attribution methods. At the same time, recent work shows that SRL remains a challenging task for large language models, particularly in settings without pre-identified predicates, where performance degrades substantially relative to structured approaches (Li et al., 2025). In such settings, even strong LLM-based methods require retrieval augmentation and task-specific scaffolding to achieve competitive performance.

This motivates the use of a structured SRL pipeline in the present study, where the goal is not maximal benchmark accuracy but a stable and interpretable representation of actor–action relations in noisy, heterogeneous archival data. Table 1 summarizes the resulting predicate extraction statistics.

The corpus contains 5.0 million year mentions, 246,073 SRL predicates, approximately 215k valid predicates, and roughly 185k rows containing subject–predicate pairs. The resulting subject extraction rate is approximately 75%, with a parser error rate of 1.6%. After preprocessing and filtering, the final dataset contains 86,414 subject–predicate–year triples distributed across six temporal periods.

Because many OCR-derived sentences contain

Period	Observations
1980–1994	165
1995–1999	291
2000–2004	3,930
2005–2009	20,074
2010–2014	49,488
2015–2019	12,466

Table 2: Number of cleaned subject–predicate observations by time period.

copular or auxiliary constructions, a subject recovery procedure was applied to identify fallback subjects in otherwise incomplete parses. This process yielded an additional 13.9% subject recoveries, producing a final inventory of 52,305 unique normalized subjects and 4,548 unique predicates across 60,451 sentences.

Year mentions were extracted to associate actor–action relations with the historical periods referenced within documents rather than relying solely on document creation dates(Pustejovsky et al., 2003). Because litigation archives frequently contain retrospective discussion of earlier events, this approach enables temporal indexing of discourse about past regulatory actions, marketing practices, and clinical developments.

Subjects were grouped into a controlled actor ontology consisting of 17 groups that fall broadly into the categories of organizational actors, individual actors, discourse artifacts, and referential placeholders. Table 3 shows the distribution of subject groups after normalization and cleaning. The distribution follows a typical long-tailed pattern, with a small number of high-frequency discourse roles accounting for a large share of predicate instances. Subject groups are defined as follows: clausal_or_artifact_subject (non-agentive linguistic or document artifacts, e.g., clauses, sections), addressee (second-person or recipient roles), individual_actor (named or generic persons), corporate_self (first-person corporate voice, e.g., “we”), information (abstract informational entities, e.g., data, reports), referential (pronouns and discourse placeholders), commercial_products (drug or product names), external_actor (third-party organizations or actors outside the focal firm), patients_consumers (patients or end-users), medical_status (conditions or diagnoses), medical_professionals (clinicians and healthcare providers), corporate_entities (named firms), regulators_government (regulatory or state

Subject Group	Count
clausal_or_artifact_subject	19,226
addressee	15,751
individual_actor	13,250
corporate_self	12,602
information	6,554
referential	5,109
commercial_products	3,359
external_actor	3,303
patients_consumers	1,640
medical_status	1,407
medical_professionals	1,306
corporate_entities	1,194
regulators_government	515
commercial_partners	430
other_actor	334
indefinite_actor	325
interrogative_actor	109

Table 3: Distribution of normalized subject groups after cleaning and actor grouping.

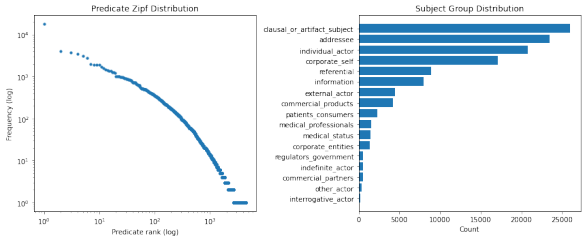


Figure 2: Zipf distribution of predicate frequencies extracted from SRL. The heavy-tailed distribution confirms typical lexical structure and suggests the extraction pipeline preserved natural predicate usage patterns.

actors), commercial_partners (distributors or business partners), other_actor (miscellaneous actors), indefinite_actor (non-specific agents, e.g., “someone”), and interrogative_actor (questioned or unknown agents).

Predicate frequency follows a heavy-tailed Zipf distribution typical of natural language corpora (Zipf, 1949).

Each extracted predicate instance forms a minimal actor–action relation that can be associated with document metadata and temporal references.

3.2 Actor Grouping

Raw subject expressions exhibit substantial lexical variation. Subjects were therefore normalized through a multi-stage procedure consisting of lexical normalization, rule-based classification, and language-model adjudication of ambiguous cases.

The rule-based stage captures high-frequency actors and organizational references. Residual subjects were evaluated using a constrained language model prompt designed for closed-set categoriza-

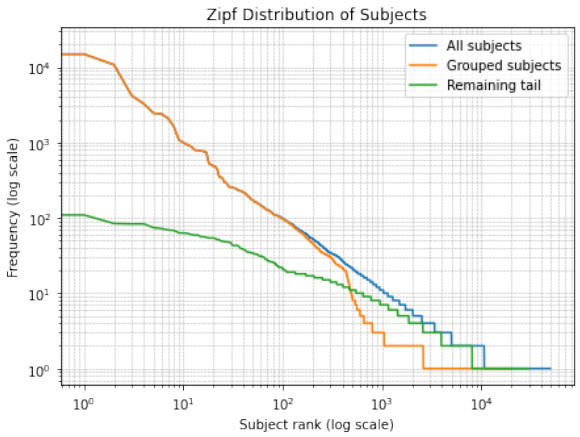


Figure 3: Zipf distribution of normalized subject expressions. The high-frequency head is largely captured by rule-based grouping, while mid-frequency subjects are resolved through LLM-assisted adjudication. The remaining long tail consists primarily of rare expressions and OCR artifacts.

tion. For each of the 267 candidate subjects, the model was shown the subject string, its corpus frequency, and eight example sentence contexts, optionally including the extracted predicate. The model was instructed to assign the subject to the single best label from the predefined actor ontology, to avoid inventing new labels, and to prefer `No Group Found` when evidence was weak or the extraction appeared malformed. Outputs were returned in a structured JSON format including a recommended group, confidence score, optional secondary group, mixed-use flag, and short rationale. A confidence cutoff of 0.9 was used. This stage recovered 267 subject types, including branded opioid products such as Kadian and Eluxadoline, and yielded 10,086 additional predicate–argument–time triples.

Approximately 15% of previously ungrouped subjects were recovered through this procedure.

Figure 3 shows the frequency distribution of subject expressions, which follows a heavy-tailed Zipf distribution typical of natural language. A small number of high-frequency subjects account for a large share of instances, while the majority occur only once or twice.

After removing high-frequency auxiliary predicates and boilerplate discourse markers, 86,414 predicate instances remained.

These distributions provide a semantic representation of actor discourse.

3.3 Semantic Feature Construction

For each actor group and time period, predicate vocabularies were mapped to meaningful psychological and social conceptual categories using the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2015; Tausczik and Pennebaker, 2010).

3.4 Discourse Space Modeling

Random forest classification was used to identify semantic features that distinguish actor groups. A one-versus-rest classification setup was used for each actor group, using LIWC category frequencies as input features. Permutation importance was estimated across 200 bootstrap samples, and features exceeding one standard deviation above the mean feature importance were retained for discourse space analysis. 13 features remained from the initial 118 provided by LIWC.

Principal component analysis (PCA) was applied to the resulting feature matrix. The first two principal components explain 38.0% of the variance in the semantic feature space (PC1: 21.6%, PC2: 16.4%). A third component explains an additional 12.8% of variance but was not included in the analysis in order to preserve a two-dimensional discourse space suitable for visualizing actor trajectories. While 3D visualization is a natural extension, we prioritize a 2D projection for interpretability and leave higher-dimensional visualization as future work.

4 Results

The result of the pipeline is a time-indexed mapping from actor groups to distributions over predicate-linked semantic categories, which can be interpreted as an empirical approximation of each group's "possibility space," or the set of actions attributed to that group within the archive.

Before modeling actor movement in semantic discourse space, we first examine the temporal distribution of subject groups in the corpus. Figure 4 shows period-wise deviations in subject-group frequency relative to each group's overall mean, expressed as z-scores. Positive values shown in red indicate periods in which a subject group is overrepresented relative to its overall distribution, while negative values shown in blue indicate underrepresentation.

The figure suggests three broad phases. First, early periods are characterized by product-,

Feature	PC1	PC2
Cognition	0.472	0.317
cogproc	0.449	0.380
insight	0.391	0.128
perception	0.284	-0.445
allure	0.278	-0.205
motion	0.235	-0.346
attention	0.232	-0.253
focuspresent	0.097	-0.296
cause	0.001	0.358
acquire	-0.079	-0.121
reward	-0.092	-0.242
need	-0.160	-0.016
work	-0.326	0.162

Table 4: Top LIWC feature loadings for the first two principal components of the discourse space. Positive and negative values indicate opposing semantic poles along each component.

medical-, and patient-centered discourse. Second, a middle period shows increasing prominence of corporate entities and external actors, suggestive of branding, distribution, and marketing activity. Third, later periods emphasize corporate, regulatory, and addressee-centered discourse, reflecting both increasing regulatory scrutiny and a greater prevalence of directive communication (e.g., "you will").

Our findings suggest that actor groups occupy distinct regions of the resulting discourse space, and their trajectories across periods reveal systematic changes in their possibility spaces.

Figure 5 plots the positions of selected actor groups from the first period (1980–1994) to the final period (2015–2019). Movement in this space reflects shifts in the semantic framing of actor discourse as captured by LIWC feature distributions.

Two principal semantic dimensions structure this space. Table 4 lists the LIWC features with the highest loadings on the first two principal components.

The first component along the x-axis (PC1) contrasts cognitive and perceptual processing language (Cognition, cogproc, insight) with goal-oriented organizational discourse (work). The second component along the y-axis (PC2) contrasts causal analytic reasoning (cogproc, cause) with experiential and perceptual language (Perception, motion, focuspresent). Together these axes differentiate discourse oriented toward explanation and reasoning from discourse oriented toward operational coordination, experiential language, and immediate activity.

No subject group remains semantically stable across the full temporal span of the archive. Several groups—including `patients_consumers`, `regulators_government`, `commercial_partners`, and most dramatically `medical_professionals`—shift substantially within the discourse space. Across these groups, discourse moves away from causal and analytic reasoning toward greater emphasis on experiential and perceptual language.

The `medical_professionals` group exhibits the largest displacement. This suggests a substantial shift in how clinicians are positioned within internal corporate communication across the periods represented in the archive.

Two additional groups display a particularly striking pattern: `corporate_entities` and `commercial_products`. The first includes firms such as Cephalon, McKesson, Endo, Insys, and Teva, while the second includes drug entities such as Exalgo, Opana, and Xartemis XR. The trajectories of these two groups move almost identically through the discourse space. Predicates associated with the companies change in the same direction, and to nearly the same degree, as predicates associated with the drugs themselves.

This parallel movement suggests that corporate actors and the pharmaceutical products they produce are treated almost interchangeably within the predicate structures of the archive. In effect, the companies producing these drugs and the drugs themselves occupy nearly identical semantic roles in the discourse.

Across both groups, discourse shifts away from goal-directed action language toward perceptual processing, while also moving slightly from experiential language toward explanatory reasoning. In practical terms, this corresponds to a reduction in action-oriented framing and a greater emphasis on explanation and interpretation.

This coupling suggests that corporate responsibility and product behavior are linguistically co-constructed within the archive, with actions attributed to drugs mirroring those attributed to the firms that manufacture them.

The only group that moves in the opposite direction is `corporate_self`. Over time this group shifts away from experiential and perceptual language and toward more goal-directed discourse, accompanied by a modest increase in reasoning-oriented language.

To clarify the behavior of the principal institutional actors, Figure 6 shows

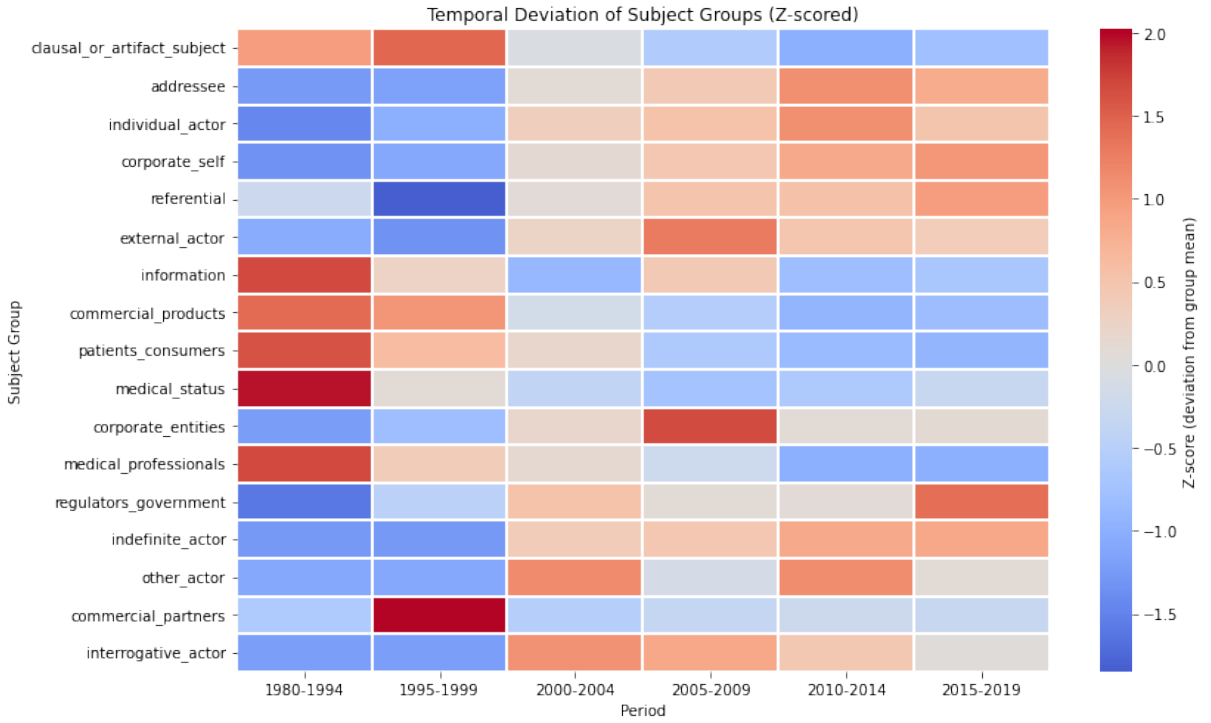


Figure 4: Temporal deviation of subject groups across periods, shown as within-group z-scores relative to each group's mean frequency. Positive values indicate periods in which a subject group is overrepresented relative to its overall distribution; negative values indicate underrepresentation. The figure highlights a temporal shift from product- and patient-centered discourse in earlier periods toward corporate, regulatory, and interactional subject positions in later periods.

the trajectories of five key subject groups across periods: `corporate_self`, `patients_consumers`, `commercial_partners`, `medical_professionals`, and `regulators_government`. Focusing on these actors highlights the most substantial movements in the discourse space and reveals that semantic change across the archive is not uniform across time.

The trajectories support the earlier interpretation that the corpus reflects three broad discursive phases: an early period oriented toward medical discussion, a middle period emphasizing distribution and commercial coordination, and a later period dominated by regulatory scrutiny and investigative discourse.

For example, the `corporate_self` group begins in a region of the discourse space associated with experiential and cognitive language. During the second period (1995–1999) it shifts further toward experiential framing, before moving sharply away from this modality in later periods. This later movement corresponds to the increasing prevalence of corporate email communication and the transition toward distribution and regulatory investigation.

By contrast, the `patients_consumers` grouping initially moves toward more experiential action language during the 1995–1999 period, before shifting toward more cognitive and perceptual discourse. This pattern is consistent with the later emphasis on retrospective patient testimony and the collection of investigative evidence.

The `commercial_partners` group exhibits the greatest overall movement, though not the greatest displacement from its starting position. This group undergoes the most pronounced shifts in discourse framing as its institutional role changes across the periods represented in the archive.

Taken together, these trajectories illustrate how actor positions within the discourse space evolve and drift across time, revealing which institutional actors occupy the most unstable or rapidly changing semantic roles. In this sense, trajectory instability provides a quantitative indicator of shifting institutional roles within the evolving discourse of the opioid crisis.

4.1 Feature Trajectories

Finally, temporal feature trajectories across periods by group highlight which semantic dimensions

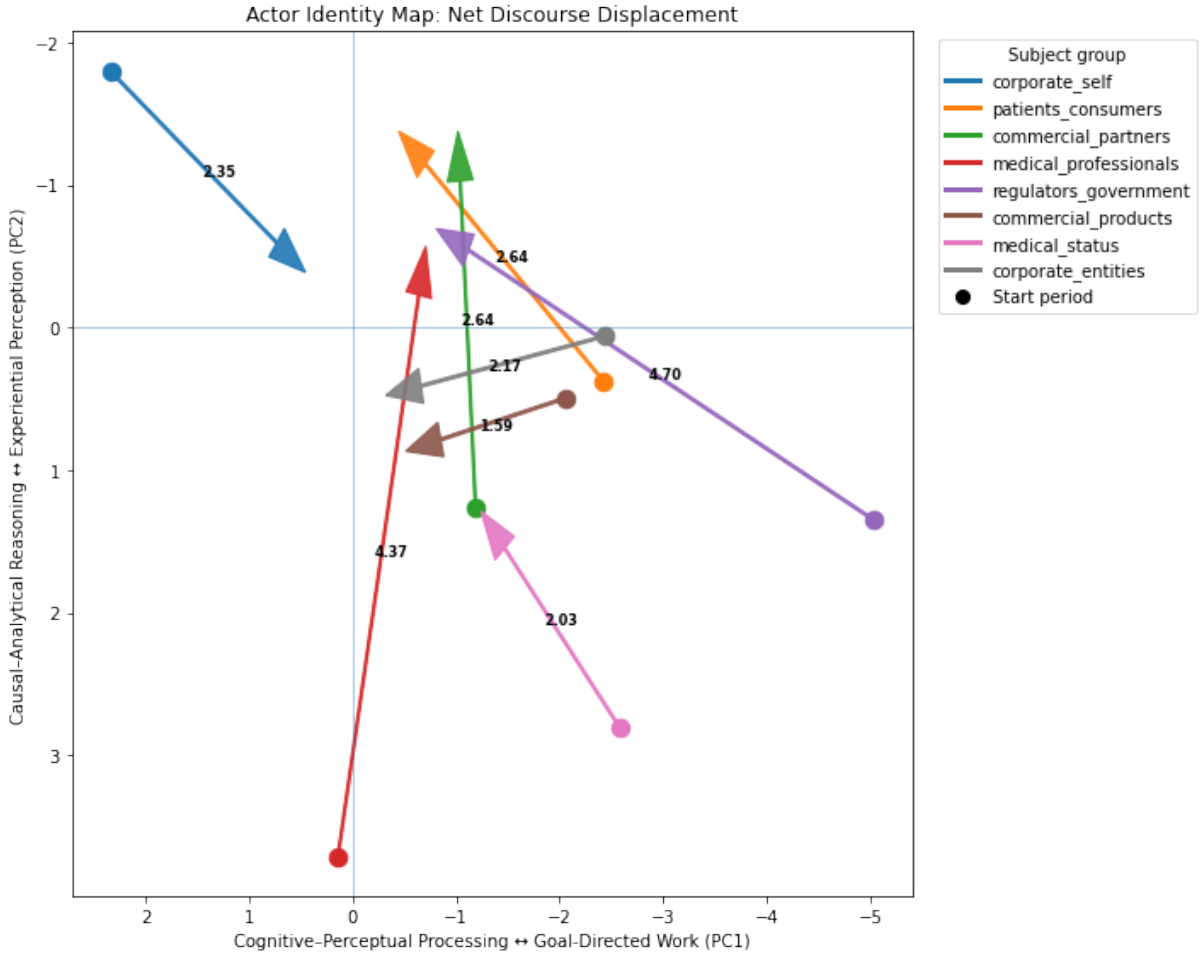


Figure 5: Actor trajectories in discourse space across historical periods.

drive actor movement in the discourse space. For example, reward language drove early variation in the 1995–1999 period for the `corporate_self` grouping, while need did the same in the 2000–2004 period for the `medical_status` group. Some features reveal relatively little change over time despite notable early spikes, such as attention, while others reveal interesting colinear pairings, such as `medical_professionals` and `patients_consumers` in the 2010–2014 and 2015–2019 periods relative to actions LIWC labeled as *acquire*. In effect, this figure provides a quantitative perspective on the actions ascribed to and undertaken by broad classes of subjects, people, and institutions as they navigated a developing catastrophic public health crisis.

5 Analysis

The results suggest a gradual reorientation of discourse across the archive. Earlier documents tend to situate discussion around products and patients, while later communication increasingly centers on corporate entities and regulatory actors.

This shift corresponds partly to changes in document type, particularly the emergence of internal corporate email around 2000. However, the semantic structure of discourse also changes, suggesting a broader transition in institutional communication as the crisis developed.

Within the discourse space, corporate actors exhibit comparatively stable positioning, while medical and regulatory actors show larger movement across periods. The near-identical trajectories of `corporate_entities` and `commercial_products` suggest that firms and the drugs they produce occupy closely aligned semantic roles within the archive. In practice, similar types of actions are attributed to both companies and their products across time. One interpretation is that product behavior and corporate behavior are linguistically co-constructed in the documents, such that actions described in relation to drugs (e.g., efficacy, risk, usage) mirror those attributed to the firms themselves (e.g., development, marketing,

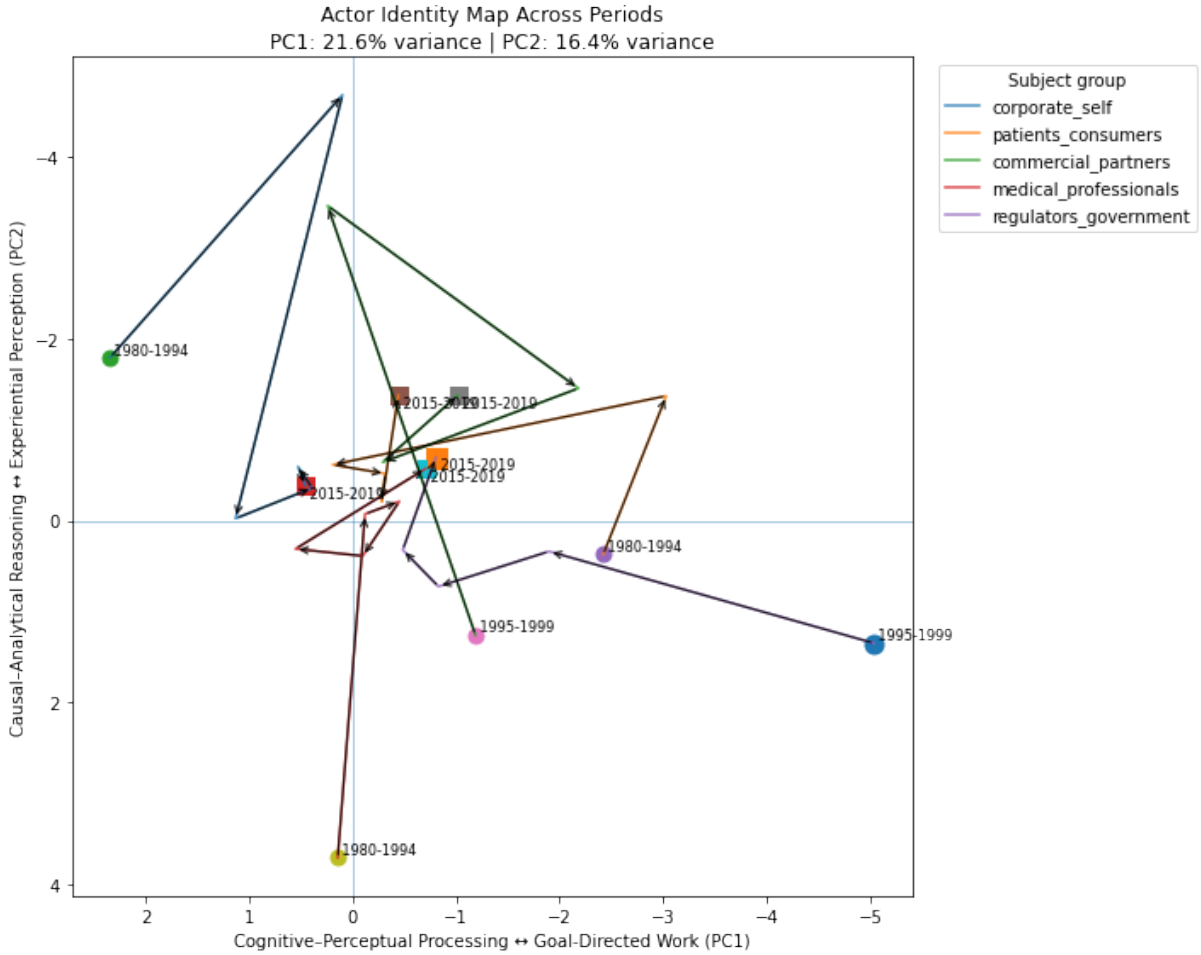


Figure 6: Trajectories of selected actor groups across periods in the PCA discourse space. Each line traces the movement of an actor group's semantic position across six temporal periods.

distribution). This pattern suggests that responsibility, agency, and outcomes are distributed across both entities and products, rather than sharply distinguished between them. Further work could disaggregate these subject positions into individual actors, clarifying which actors are foregrounded or backgrounded in the attribution of responsibility.

A key interpretive caveat is the shift in corporate communication associated with the adoption of email. To assess whether the observed discourse shifts could be driven primarily by communication medium, we approximate the prevalence of email-style documents using synonym-based detection of header markers. Table 5 summarizes opioid-term frequency and the estimated prevalence of email markers. The results show a sharp increase in email-style communication after 2000, indicating that part of the observed shift reflects changes in archival composition. However, the persistence of structural changes in actor–action relations suggests that the trajectories capture more than a simple medium effect.

Period	Opioid Mentions	Share of Documents	Email Markers
1980–1994	72	14–19%	0.0%
1995–1999	196	∼24%	3.3%
2000–2004	796	∼9%	9.7%
2005–2009	7,041	∼17%	11.5%
2010–2014	9,632	∼10%	18.8%
2015–2019	2,159	∼8%	15.3%

Table 5: Temporal distribution of opioid-term mentions and email markers across periods.

Email markers include header elements such as *From*, *To*, *Subject*, and timestamp fields.

The sharp shift after 2000 partly reflects the increasing presence of internal communication such as email, but the change in subject roles cannot be explained solely by communication medium. Instead, the corpus reveals a broader transition from product- and patient-centered discourse toward organizational coordination, corporate entities, and

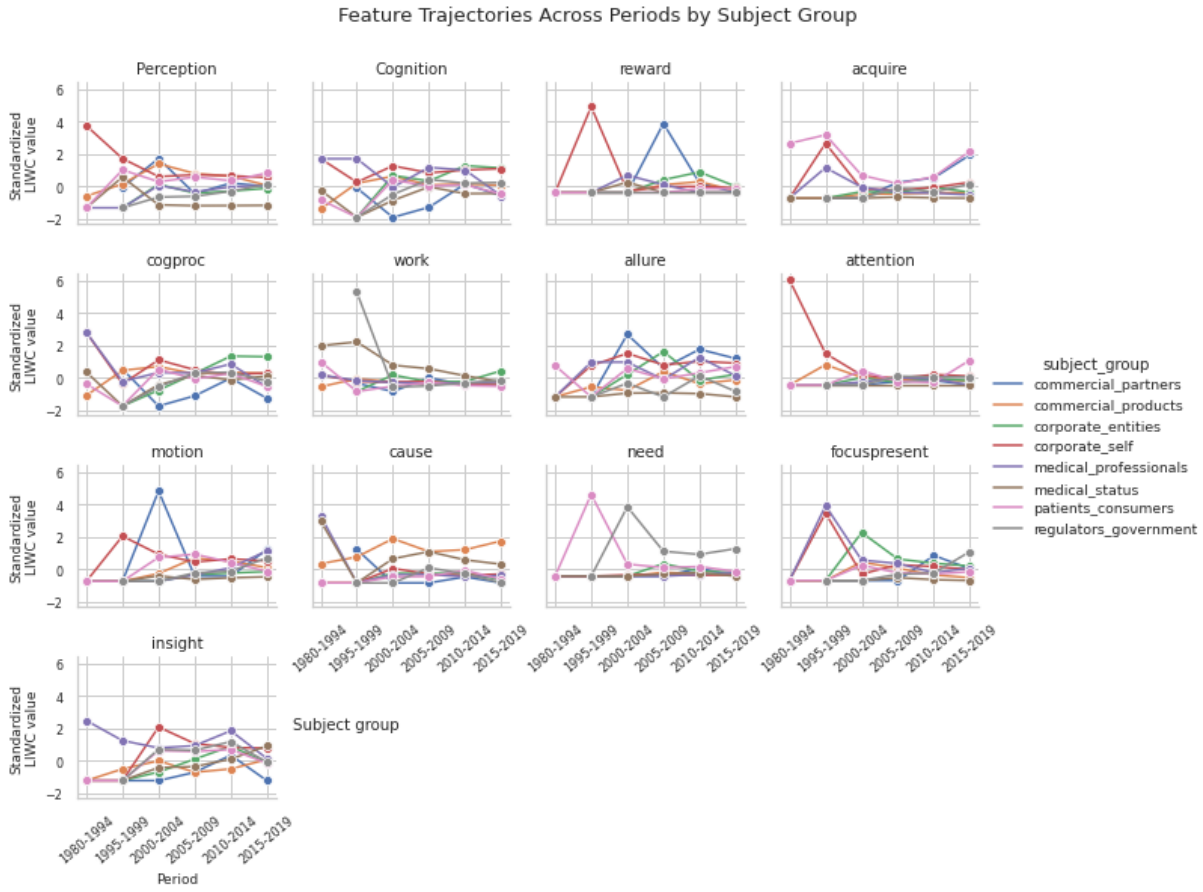


Figure 7: Temporal trajectories of selected LIWC features across actor groups.

regulatory actors.

As a second validation test, the frequency of mentions of opioids as a category was also measured. Early periods contain a higher proportional share of direct references to opioid products and medical conditions, but later periods increasingly center on corporate actors, regulatory institutions, and directive communication among organizational members. This suggests a shift in the archival record from product- and patient-focused discourse toward internal coordination and regulatory engagement.

Taken together, these results and tests indicate that the observed semantic trajectories reflect changes in what actions are attributed to actor groups, rather than shifts in latent topic or document similarity alone.

6 Conclusion

This paper introduced a computational pipeline for modeling actor discourse trajectories in the Opioid Industry Document Archive. By extracting predicate structures, grouping actors, and constructing a semantic discourse space, the method produces interpretable representations of institutional communication and semantic change of subject groupings over time.

The resulting actor trajectories reveal systematic differences in how corporate actors, regulators, clinicians, and patients are positioned within internal communication and how these positions evolve across historical periods.

These results demonstrate how structured actor–action representations can reveal shifts in institutional discourse that are not captured by topic-based or document-level analyses alone, and provide a foundation for more granular analysis of responsibility, agency, and role attribution within corporate structures.

Limitations

A key limitation is the uneven temporal distribution of documents in the archive. Early periods (pre-2000) contain substantially fewer documents than later periods, which are dominated by internal corporate email. The archive also contains hetero-

geneous document types, including reports, emails, and legal transcripts, as would be normal for any comprehensive corporate archive. As a result, the observed diachronic shifts cannot be interpreted purely as changes in that underlying corporate behavior; they also reflect changes in document production, preservation, and legal disclosure. The analysis and signal therefore also captures changes in the archival representation of institutional discourse, rather than a fully controlled sample of communication across time.

To mitigate the effects of temporal imbalance, actor–action relations are indexed using within-period normalization, and analysis focuses on relative deviations (z-scores) rather than raw counts. However, comparisons between early and late periods should be interpreted cautiously, with greater confidence placed on within-period structure and post-2000 trends where document density is higher.

Explicit modeling of semantic action trajectories in pre- and post-email corporate regimes could help disentangle the effects of communication medium from underlying institutional change. Additionally, early periods remain relatively sparse despite targeted sampling. A supplemental sampling strategy for pre-1990 documents was explored, but the combination of OCR noise, data sparsity, and deviations from the random sampling design led to its exclusion. Future work could focus on improving SRL robustness for noisy OCR data.

Acknowledgements

I thank the organizers of the NLP+CSS OIDA Shared Task; the United States Bankruptcy Court for the Southern District of New York, whose orders in the Purdue Pharma bankruptcy proceedings require the creation of a Public Document Repository on the subject; and the University of California, San Francisco and Johns Hopkins University for maintaining and providing access to the Opioid Industry Documents Archive. I am grateful to Lawrence Fogelman for discussions of the legal context and interpretive considerations relevant to this work. Portions of the analysis pipeline code were developed with the assistance of a large language model (GPT-5.3). The model was used for programming assistance, debugging, and LaTeX formatting; all methodological decisions, analyses, and interpretations were conducted by the author.

References

G. Caleb Alexander, Lisa A. Mix, Sayeed Choudhury, Rachel Taketa, Cecilia Tomori, Mehdi Mooghali, Andrew Fan, Sarah Mars, Daniel Ciccarone, Michael Patton, Dorie E. Apollonio, Laura Schmidt, Michael Steinman, Jeremy Greene, Pamela Ling, Andrew K. Seymour, and Stanton Glantz. 2022. The opioid industry documents archive: A living digital repository. *American Journal of Public Health*, 112(8):1126–1129.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Han He, Liyan Xu, and Jinho D Choi. 2021. Elit: Emory language and information toolkit. *arXiv preprint arXiv:2109.03903*.

Xinxin Li, Huiyao Chen, Chengjun Liu, Jing Li, Meishan Zhang, Jun Yu, and Min Zhang. 2025. Llms can also do well! breaking barriers in semantic role labeling via large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23162–23180.

Weeda Mehran, Ben Miller, and Stephen Herron. 2025. Nothing in common? analysis of moral, psychological, and social factors in the identity construction of far-right and violent jihadi extremists. *Studies in Conflict & Terrorism*, pages 1–24.

Ben Miller, Jennifer Olive, Shakthidhar Gopavaram, Yanjun Zhao, Ayush Shrestha, and Cynthia Berger. 2015. A method for cross-document narrative alignment of a two-hundred-sixty-million word corpus. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1673–1677. IEEE.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Nalini Vadivelu, Alice M Kai, Vijay Kodumudi, Julie Sramcik, and Alan D Kaye. 2018. The opioid crisis: a comprehensive overview. *Current pain and headache reports*, 22(3):16.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*.

Toward Unsupervised Conceptual Metaphor Discovery: A Case Study in Online Immigration Discourse

Alexandria Leto
University of Colorado Boulder
alexandria.leto@colorado.edu

Maria Leonor Pacheco
University of Colorado Boulder
maria.pacheco@colorado.edu

Abstract

In Conceptual Metaphor Theory (CMT), a metaphor is a systematic mapping from a concrete source domain (e.g., physical load) to a more abstract target domain (e.g., taxes), so that reasoning about concepts in the target domain is guided by inferences from the source domain (Lakoff, 1993). In this work, we propose that since different source domains can frame the same target in starkly different ways, the conceptual mappings evidenced by metaphorical expressions can guide computational political discourse analysis. We present a proof-of-concept for an unsupervised method that uncovers salient conceptual mappings from a corpus. Prior work in computational political metaphor analysis has drawn on CMT, but it typically requires a predetermined inventory of focused source and target domains. In contrast, we introduce a simple LLM-based method that detects metaphorical expressions from a corpus with strong performance, then clusters them to approximate source domain categories. We demonstrate its utility through a case study on online immigration discourse, showing that the resulting metaphor clusters provide context for frame analysis. We conclude by outlining future work needed to develop a robust framework for conceptual metaphor discovery in political discourse.

Code

1 Introduction

In Conceptual Metaphor Theory (CMT), a theoretical framework in cognitive linguistics that originated with Lakoff and Johnson (1980), metaphors are viewed as a fundamental structure of human thought. They allow us to understand abstract or unfamiliar concepts (the target domain) in terms of more concrete or familiar ones (the source domain). For example, in the metaphorical expression "Taxes burden the middle class," the source word "burden" invokes the *physical load* source domain (Figure 1).

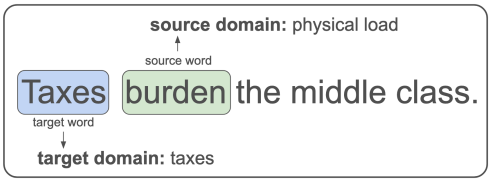


Figure 1: Example of a metaphorical expression inspired by Lakoff (2004).

Correspondences between concepts in the source domain and concepts in the target domain, referred to as conceptual metaphors or conceptual mappings (Lakoff, 1993), are thus formed: taxes are a physical load and the middle class is the carrier.

The implications of these mappings provide an understanding of how authors frame an issue, defined as "selecting some aspects of a perceived reality and making them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (Entman, 1993). In our example, the problem definition is the economic "burden" that the middle class is "carrying," with the implied solution being to "lift" it through reduced taxation. Conceptual metaphors therefore offer a useful lens for political framing analysis. In this work, we propose an unsupervised framework for discovering them in a corpus, recovering the metaphorical expressions used and the conceptual mappings they are drawn from.

While prior frameworks have used CMT as the scaffolding for large-scale discourse analyses, they require a pre-defined list of expected source concepts (Mendelsohn et al., 2020; Card et al., 2022; Mendelsohn and Budak, 2025). The source concepts explored are generally focused on a relatively narrow view of metaphor, honing in on "dehumanizing" metaphors, in which a person or group of people is the target concept, and the source concept (such as "vermin" or "animals") serves to strip the

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 159–175

group of human qualities. Additionally, the target is often fixed to a particular group such as "immigrants" (Wang, 2024), precluding the discovery of other targets that may be equally relevant. For example, in immigration discourse, one might also be interested in how politicians, law enforcement, or immigration policy are metaphorically framed.

Our framework, in contrast, is more flexible than prior approaches, enabling the discovery of conceptual metaphors in novel datasets or discourse types for which no established inventory of conceptual mappings exists. This paper serves as a proof of concept for its feasibility and usefulness in political framing analysis. Below, we outline our main contributions.

1. An unsupervised method for discovering conceptual metaphors that first extracts candidate word pairs with grammatical relationships associated with metaphor use, identifies which pairs invoke a metaphor, generates rich textual descriptions for the resulting *(source word, target word)* pairs, and then clusters these descriptions so that expressions sharing a conceptual mapping are grouped together.
2. A case study demonstrating that the discovered conceptual mappings provide signal for frame prediction and reveal how immigration is framed in in online political discourse.

These contributions constitute a proof-of-concept toward an automated conceptual mapping discovery framework for frame analysis. We close with a discussion of the main challenges and opportunities for fully realizing this vision.

2 Background

Metaphor In their book *Metaphors We Live By*, Lakoff and Johnson (1980) introduced CMT, arguing that metaphor is not limited to figurative language, but is instead a fundamental structure undergirding our conceptual system. Metaphors, also referred to as conceptual metaphors or conceptual mappings, are a sets of ontological correspondences between concepts in a target domain and concepts in a source domain, allowing patterns of inference in the source domain to be applied to the target (Lakoff, 1993). The theory spurred an influx of work that, over the years, built on, criticized, and altered the original theory (Grady, 1997; Charteris-Black, 2016; Kövecses, 2017, 2000, 2020).

With this growing volume of metaphor work in linguistics, Group (2007) noted a resulting "[v]ariability in intuitions, and lack of precision about what counts as a metaphor." In response, they developed the Metaphor Identification Procedure (MIP), a step-by-step guide for identifying metaphorical lexical unit in text. In MIP, an annotator first establishes a lexical unit's meaning in the given context, then determines whether the unit has an alternative "basic" meaning (one that is more concrete, embodied, precise or historically older) in other contexts. If it does, the lexical unit is metaphorical. We adopt the operationalization of Lakoff and Johnson (1980)'s conceptual metaphors from the Language Computer Corporation (LCC) dataset (Mohler et al., 2016). In LCC, a metaphorical expression is a two-term unit within a sentence consisting of a lexical unit invoking a target domain and another invoking a source domain, and both novel and conventionalized metaphors are included. We also use LCC to benchmark our metaphor expression identification component.

Computational Metaphor Detection and Interpretation Early computational work for detecting and interpreting metaphors (Fass, 1991; Martin, 1990; Narayanan, 1999; Feldman and Narayanan, 2004; Agerri et al., 2007) relied domain-specific, hand-annotated metaphor knowledge bases such as the Master Metaphor List (Lakoff et al., 1991), MetaBank (Martin, 1994), and the Mental Metaphor Databank (Agerri et al., 2007), among others. While these hard-coded approaches were powerful, they lacked broad coverage (Shutova, 2010). As a result, practitioners turned to unsupervised methods for metaphor processing such as clustering (Mason, 2004; Shutova, 2010; Shutova and Sun, 2013; Mohler et al., 2013; Shutova et al., 2017) and topic modeling (Heintz et al., 2013), sometimes supplementing these approaches with knowledge from broader lexical resources such as WordNet (Miller, 1994).

Since, practitioners have approached metaphor analysis using supervised neural network and BERT-based models (Do Dinh and Gurevych, 2016; Swarnkar and Singh, 2018; Pedinotti et al., 2021; Li et al., 2024). However these approaches, like early work, tend to require large hand-labeled datasets for training and do not necessarily generalize well to out-of-domain data (Yang et al., 2023).

More recently, Large Language Models (LLMs) have been assessed for their ability to detect metaphors in an unsupervised setting (Dankin et al., 2022; Ichien et al., 2024; Tong et al., 2024;

Sanchez-Bayona and Agerri, 2025). For example, Puraivan et al. (2024) evaluate a set of prompts for obtaining a binary metaphorical classification for a single Spanish verb, yielding high accuracy on their test set. Tian et al. (2024) propose an explainable approach to word-level binary metaphor detection, in which they guide the LLM's reasoning with CMT-based "scaffolding." Fuoli et al. (2025) formulate the task at the phrase level, prompting LLMs to identify metaphorical phrases in movie reviews. They evaluate a range of methods, including Retrieval Augmented Generation (RAG), prompt engineering, and model fine-tuning for their ability to improve performance. This body of work shows that LLMs represent a promising avenue for unsupervised metaphor detection. We follow this line of work, using an LLM to identify metaphorical expressions and generate rich descriptions for them, then cluster these descriptions to recover conceptual mappings from a corpus.

Political Metaphor and Frame Analysis Metaphor is a powerful tool for framing political issues, as the source domain chosen emphasizes certain aspects of a target while backgrounding others (Entman, 1993; Lakoff, 2004). For example, when immigrants are described as "flooding" the border, the evoked *water* source domain casts them as an uncontrollable natural force, foregrounding threat and overwhelm. When the same target is described as "seeking refuge," a very different source domain (one of vulnerability and safety) is evoked, inviting sympathy instead.

A substantial body of computational work has operationalized this insight. Mendelsohn et al. (2020) and Card et al. (2022) each include "dehumanizing metaphor dimensions" in their respective large-scale framing analyses. Wang (2024) presents a method for extracting metaphors from politically-charged news articles by filtering word pairs which can grammatically invoke a metaphor, then using a fine-tuned RoBERTa classifier to predict a "metaphor score." Most recently, Mendelsohn and Budak (2025) released a framework for obtaining scores for seven metaphor categories associated with immigration (including "water" and "war"). However, most of these approaches rely on a predetermined set of source and target concepts (Mendelsohn et al., 2020; Card et al., 2022; Mendelsohn and Budak, 2025), restricting their analyses to metaphors that are already well-documented for a particular topic. While Wang (2024) addresses this with an unsupervised method designed to uncover new source concepts from metaphorical verbs, they limit their consideration to target nouns referencing immigrants. We present a first step toward addressing these limitations with an unsupervised, automated framework that generalizes to new topics without a pre-conceived notion of expected target or source concepts.

3 Inducing Metaphors

In CMT, metaphorical expressions in natural language serve as evidence for the conceptual mappings which underlie them. Building on this, we present an unsupervised method for identifying metaphorical expressions, paired with a clustering approach to group those that share a conceptual mappings. Following Mohler et al. (2016), we treat metaphorical expressions as within-sentence `(source word, target word)` pairs that grammatically invoke a metaphor, where the source word has an alternative "basic" meaning (one that is more concrete, embodied, precise or historically older) in other contexts (Group, 2007).

Although other parts of speech can grammatically invoke a metaphor, we follow prior work (Shutova et al., 2010) in limiting our scope to metaphors composed of a source verb and a target noun. This choice serves three purposes. First, it makes our framework more computationally tractable. Second, focusing on target nouns aligns our work with prior work on politically-charged dehumanizing metaphors (whose targets are typically nouns), giving us a meaningful point of comparison. Third, limiting source words to verbs lets us take advantage of recent work demonstrating high LLM performance on identifying metaphorical verbs (Puraivan et al., 2024).

3.1 Extracting Candidate Metaphorical Expressions

Prior work shows that metaphors occur in a limited set of grammatical patterns (Petruck and Dodge, 2016). We use this knowledge to extract a set of candidate `(source verb, target noun)` pairs from each sentence in a given corpus. We follow the procedure in Wang (2024), identifying all `(verb, noun)` pairs in each sentence, then use the *spaCy* python library[1] to determine the shortest dependency path (SDP) between each pair. If the SDP matches one of the pre-defined metaphor-

[1] `https://spacy.io/`

ical construction patterns (shown in Table 7), the corresponding target and source word pair are considered a metaphor candidate.

3.2 Metaphorical Expression Detection

To confirm whether a target and source word pair are in fact metaphorical, we prompt an LLM in a zero-shot setting to generate a "metaphor salience score" between 0 and 1, representing how likely a human would be to recognize the source verb as metaphorical in its sentence context. Because LLMs have been shown to over-classify metaphors in binary classification settings (Hicke and Kristensen-McLachlan, 2024), the continuous scores let us calibrate predictions to the task. We also prompt the LLM to generate a rich natural-language explanation for each identified metaphor, which we later use to cluster expressions with similar conceptual mappings. The prompts for this step are adapted from Puraivan et al. (2024) and can be found in the Appendix, Figure A.1.

3.3 Grouping Metaphors of Similar Domains

Source Domain Groups To identify groups of metaphors likely to invoke the same source domain, we cluster the LLM-generated explanations of each metaphorical expression (Table 1) using the K-means algorithm with cosine similarity. We embed each metaphor explanation with `SBERT` using the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). While this model was originally designed to produce sentence-level embeddings, it has been successfully used to produce document-level embeddings (Zhang et al., 2025). We test a variety of cluster counts, incrementing by 25 where the minimum is 25 and the maximum is 300.

Target Domain Groups To identify metaphors that map to the same target domain, we first establish a set of canonical target domains through a human-in-the-loop process. We begin by categorizing each metaphorical target noun as a "Person," "Place," or "Thing" using a zero-shot LLM prompt. Within each category, we embed target nouns using SBERT (`all-MiniLM-L6-v2`) and apply K-means to cluster semantically similar nouns, selecting k using the elbow method. We then hand-annotate the resulting groups to arrive at a canonical set of target domains per noun category, producing a hierarchical set of canonical target families such as those shown in Table 4. Finally, we prompt an LLM in a zero-shot setting to map each target noun in context to its canonical target domain. All prompts are provided in Appendix A.4.

4 Evaluation

In this section, we introduce the datasets used for evaluation and assess the quality of each pipeline component, including candidate extraction, metaphor detection, and source and target grouping.

4.1 Datasets

We evaluate our metaphor detection component (Section 3.2) using a portion of the English LCC dataset (Mohler et al., 2016). The LCC is a collection of general-domain excerpts annotated with source and target span tags and a metaphoricity score between 0 and 3 (inclusive). Each excerpt is also associated with target domains spanning a wide range of topics such as "guns," "migration," and "abortion". The full dataset contains $\sim 78,000$ annotated excerpts. However, because our method centers on identifying metaphorical (`source verb`, `target noun`) pairs, we focus on instances where the target and source spans include only a single word and where the source span is a verb, and the target span a noun. This results in a set of $\sim 10,000$ excerpts, with $5,733$ postive and $4,285$ negative examples. The distribution of metaphor scores for this set is shown in Figure 6. We include additional information, including a full list of target domains and the metaphoricity score distribution, in Appendix A.2.

To evaluate the quality of the resulting metaphor groups (Section 3.3), we use a dataset of English tweets about immigration released by Mendelsohn and Budak (2025). This dataset has two splits; the first split contains $\sim 1,600$ tweets annotated with tweet-level metaphoricity scores between 0 and 1 and a label mapping it to one of eight source domains (animal, commodity, parasite, pressure, vermin, war, water, and domain-agnostic). The second split contains $\sim 35,000$ tweets that do not include labels for metaphoricity.

4.2 Quality of Candidate Pairs

We recruit three graduate students to annotate 100 metaphor (`source verb`, `target noun`) pairs automatically extracted from the LCC dataset to evaluate the quality of our candidate extraction. Two are computer scientists and the third is a linguist. Annotators were instructed to determine whether

Score	Explanation
0.0	*enslaved* is used literally. It refers to the historical reality of African slaves in European Colonies.
0.7	*used* is applied metaphorically to the concept of *anti-blackness*, framing it as a tool.
1.0	*support* is applied metaphorically, framing a *political stance* as structural reinforcement.

Table 1: Examples of Qwen 3 metaphoricity scores and explanations.

	LCC Score Thresholds		
	1	**2**	**3**
Random	0.495	0.496	0.499
Llama3.2	0.590	0.603	0.619
Qwen3	0.794	0.783	0.781

Table 2: ROC-AUC scores for our metaphoricity scores with different LCC metaphoricity score thresholds used to determine positive samples.

		precision	**recall**	**f1**
Random	thresh = 0.2	0.570	0.792	0.663
	thresh = 0.3	0.568	0.690	0.623
RoBERTa	5-fold	0.817	0.809	**0.812**
	generalize	0.780	0.785	0.781
Llama3.2	binary prompt	0.580	0.991	0.732
	score prompt	0.588	**0.975**	0.733
Qwen3	binary prompt	**0.828**	0.752	0.788
	score prompt	0.802	0.782	0.792

Table 3: Results of baseline and our models on the LCC dataset. Our metaphor score prompt with `Qwen3` outperforms all other models.

each pair constitutes a valid metaphor candidate. Pairs were doubly-annotated with a third annotator breaking ties. We obtained a Krippendorf's α of 0.893 and determined an accuracy of 0.891.

4.3 Quality of Metaphor Identification

We evaluate the quality of our metaphor identification component using the LCC dataset. We assess both the correlation of our continuous metaphoricity scores with human annotations and the performance of our method on a binary classification task, comparing against supervised methods and a random baseline.

Metaphoricity Thresholds A key advantage of our continuous metaphoricity score is its flexibility. By adjusting the threshold, practitioners can tune the system toward capturing subtle metaphorical language or restricting it to only the most salient instances. We evaluate this claim by computing the Spearman correlation between our continuous metaphoricity scores and the hand-annotated four-point scores provided with the LCC dataset. We find that `Qwen3-8B` (Team, 2025) produces scores that correlate moderately well with human judgements ($\rho = 0.564$), while `Llama3.2-3B` (Dubey et al, 2024) shows only weak correlation ($\rho = 0.211$). This suggests that our continuous score is informative when paired with a capable model.

We also report ROC-AUC scores at various LCC metaphoricity score thresholds (Table 2). We find that `Qwen3` salience scores are moderately predictive of the binary metaphor label at every LCC threshold, significantly outperforming both `Llama3.2` and the random baseline. This supports our claim that practitioners can set a threshold suited to their task.

Binary Classification Since our scores are on a different scale from the LCC annotations (a continuous score from 0 to 1 versus a four-point scale), evaluation is not straightforward. Following Wang (2024), we reframe evaluation as a binary classification task, assigning negative labels to excerpts with a metaphoricity score of 0 and positive labels to those with a score of 1 or greater. Because we are interested in capturing even subtle metaphorical instances, this mapping aligns well with our goals. We report results at the threshold achieving the highest F1 score in Table 3. A full analysis with various thresholds is included in Appendix A.4.

We compare our prompting method against two alternative methods: a random baseline that assigns uniform random scores and a supervised RoBERTa classifier following Wang (2024), which fine-tunes the model on the LCC dataset using source span embeddings as input to the classification layer. For RoBERTa, we use 5-fold cross-validation and report micro-averaged results. To assess generalization across domains, we also evaluate RoBERTa in a leave-one-domain-out setting, fine-tuning on all but one target domain and testing on the held-out domain, and micro-averaged across all 32 domains.

The results are shown in Table 3. As expected, the supervised model achieves the highest macro F1, benefiting from in-domain training data. Our

Category	Canonical Groups
Person Groups	"immigrants", "non-immigrant u.s. citizens", "politicians", "law enforcement", "government", "other"
Place Groups	"U.S Country", "U.S. State", "U.S. City", "Non-U.S.", "unclear"
Thing Groups	"immigration", "money", "idea/opinion", "abstract concept", "other"

Table 4: Canonical target domains identified in hand-annotated set of tweets about immigration.

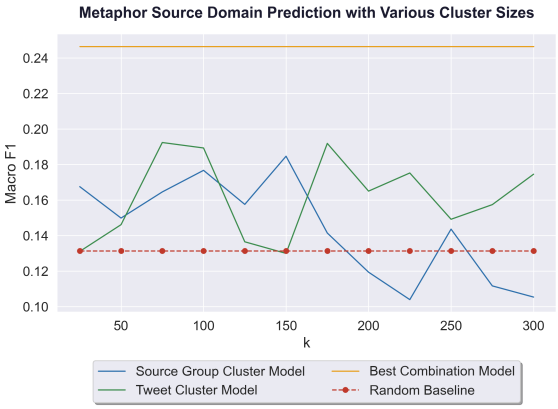


Figure 2: Performance of logistic regression models trained with various feature vectors. Best performance is achieved when metaphor source group cluster ($k = 50$) and tweet cluster ($k = 75$) information is combined (yellow).

prompting method with `Llama3.2` significantly over-classifies metaphor, achieving high recall at the cost of precision. However, our binary and score-based prompting methods with `Qwen3` perform competitively, and given that our approach requires no labeled data, we expect it to generalize more reliably to new domains and corpora.

To account for the possibility that the LCC dataset appeared in the pretraining data for `Qwen3` and inflated our results through data leakage, we additionally evaluate on a held-out set of examples with no publicly-available metaphor labels. We sample 100 `(source verb, target noun)` pairs from the unlabeled corpus of immigration-related tweets, balanced across LLM-generated metaphoricity scores, and have an author of the paper independently annotate each as "Metaphorical" or "Literal" using MIP (Group, 2007). We obtain an F1 score of 0.696 for this set, confirming the relatively high performance of the `Qwen3` score-based prompting method.

4.4 Quality of Domain Groups

We assess our target and source domain grouping methods using the immigration dataset. For the source domains, we measure the predictive signal of our induced groups for Mendelsohn and Budak (2025)'s source domain labels, accompanied by an intrusion test and qualitative analysis of the resulting groups. For the target domains, we conduct an annotation study to evaluate the quality of the group assignments.

4.4.1 Source Domain Groups

Predictive Signal Analysis To assess whether our induced source domain groups capture meaningful information, we frame the evaluation as a classification task—given a metaphorical tweet, predict the source domain labels of Mendelsohn and Budak (2025). To isolate the signal contributed by the groups themselves, we represent each tweet solely by its source domain group memberships, without incorporating any textual information, and train a logistic regression classifier on this representation, restricting to tweets with a metaphoricity score exceeding 0.3. To do this, we construct feature vectors of length k, where k is the number of clusters used to group LLM-generated metaphor explanations. All values of the feature vector are initialized to 0. If the tweet contains a metaphorical `(source verb, target noun)` pair that belongs to a given cluster i, the feature at index i is set to the metaphor salience score for the pair. If a single tweet has multiple pairs belonging to the same cluster, we take a mean of the metaphor scores. We note that our source domain groups, induced at the verb level from general metaphorical language, are not directly comparable to Mendelsohn and Budak (2025)'s tweet-level labels, which are assigned from a pre-defined set of dehumanization-focused domains. We therefore expect a performance ceiling, and instead focus on whether our induced groups carry meaningful predictive signal independent of the underlying text.

We compare against two baselines: a random baseline, and a *Tweet Cluster Model*, a logistic regression classifier using the same feature vector approach but clustering tweet context embeddings directly rather than metaphor explanations, serving as a metaphor-agnostic point of comparison. We also report the best performance obtained by combining the *Source Domain Cluster* and *the Tweet Cluster* feature vectors for all values of k.

Macro F1 scores for this predictive signal anal-

Source Domain	Metaphor Explanations
"physical object"	(1) *exhibited* is used metaphorically, framing *patriotism* as a tangible object that can be displayed.
	(2) *shows* is used metaphorically, framing *madness* as something that can be physically displayed.
	(3) *expose* is used metaphorically, framing the *political intentions* as the physical act of uncovering.
"money"	(1) *paying* is used metaphorically, framing the act of *allocating attention* as a financial transfer.
	(2) *fund* is used metaphorically, framing *policy support for immigration* as financial backing.
	(3) *cost* is used metaphorically, framing *losing lives* as a monetary expenditure.
"war"	(1) *won* is used metaphorically, framing a *political struggle* as a physical battle with a clear victor.
	(2) *fought* is used metaphorically, framing the *struggle for sovereignty* as physical combat.
	(3) *win* is used metaphorically, framing a *political or strategic struggle* as a physical battle to be won.

Table 5: Qualitative analysis of source group clusters from online immigration discourse.

ysis are shown in Figure 2. The results show that the models trained on the metaphor and tweet cluster feature vectors each outperform the random baseline across various values of k. Maximum performance is achieved by combining the two feature vectors, suggesting that the verb-level metaphor groups and text-based tweet groups capture complementary information for identifying Mendelsohn and Budak (2025)'s tweet-level source domains.

Intrusion Test We also evaluate the resulting source domain groups with an intrusion test. We evaluate the Metaphor Level clusters with $k = 50$ (see Figure 2). We construct triplets of LLM-generated metaphor explanations where two samples are from a given cluster and the third is from a different cluster. Annotators are asked to identify the intruder, i.e., the sample that does not fit with the others. We construct 100 triplets under three difficulty settings by ranking samples within each cluster by proximity to the centroid. In the "easy" setting (33 triplets), the two matching samples are drawn from the top 25% of the cluster, making them more semantically similar and easier to distinguish from the intruder. In the "medium" setting (33 triplets), they are drawn from the top 50%, and in the "hard" setting (34 triplets), from the top 75%. Each triplet is annotated by two annotators. We obtain a Krippendorff's α of 0.939, indicating high agreement and suggesting that the clusters are cohesive across difficulty levels.

Qualitative Analysis We inspect clusters for cohesiveness and identify source domains that are well documented CMT. Table 5 shows three examples: "physical object," "money," and "war" (Lakoff and Johnson, 2024).

4.4.2 Target Domain Groups

We identify 16 canonical target domains, shown in Table 4. To evaluate the quality of the target domain assignments, we recruit three students (same as Section 4.2) to annotate 100 randomly selected `(source noun, target group)` pairs with a metaphoricity score greater than 0.3. For each pair, annotators determine whether the example maps to the specified target noun and domain groups.

The annotators achieve a Krippendorff's α of 0.728 on noun group membership and 0.849 on target domain membership. `Qwen3` assigns noun groups with an accuracy of 0.913 and target domains with an accuracy of 0.717.

5 Case Study on Immigration Discourse

In this section, we outline our case study on immigration discourse. We use the larger split of $\sim 35,000$ immigration tweets, which include induced general policy frame, issue-specific frame, and narrative frame labels. These frames represent media frames, i.e., interpretive lenses used to emphasize particular aspects of an issue and shape how it is understood in public discourse. Our analysis is designed to ascertain whether the metaphor source domain clusters contribute signal for predicting frame labels. A detailed explanation for each frame can be found in Mendelsohn et al. (2021).

Predictive Signal We train and evaluate three logistic regression classifiers for frame label prediction: (1) *Source Group Cluster Model* (2) *Tweet Cluster Model* and (3) *Best Combination Model*. Features for these models are described in Section 4. We select k, the number of clusters for each model, based on the results shown in Figure 2,

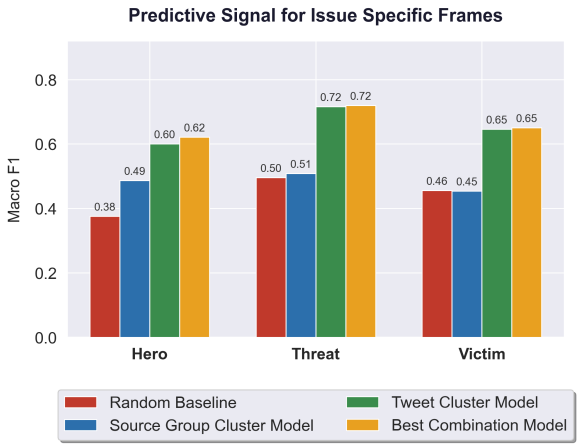


Figure 3: Performance of logistic regression models trained with various feature vectors to detect Issue Specific Frames in a dataset of tweets about immigration.

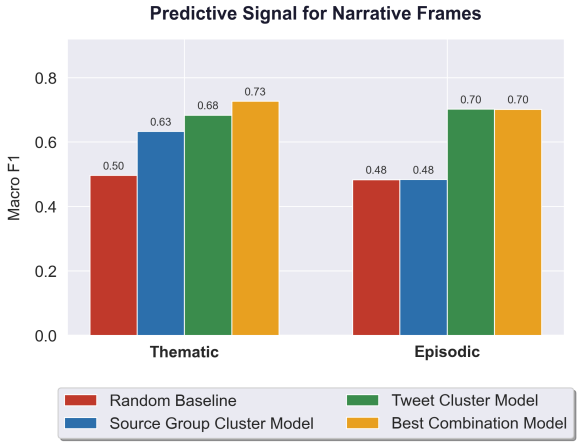


Figure 4: Performance of logistic regression models trained with various feature vectors to detect Narrative Frames in a dataset of tweets about immigration.

where maximum performance is achieved when metaphors are clustered with $k = 50$ and tweets are clustered with $k = 75$. Macro F1 results for predicting issue-specific frames are shown in Figure 3, narrative frames are shown in Figure 4, and general frames are shown in Figure 5.

We find that each of the three models typically outperforms the random baseline in predicting frame labels. An exception is that the *Source Group Cluster Model* fails to outperform the random baseline to predict "Episodic," "Security and Defense," and "Crime and Punishment" frames. The *Tweet Cluster Model* outperforms the *Source Group Cluster Model* for all frames. Consistently, the *Best Combination Model* achieves the best Macro F1, outperforming the *Tweet Cluster Model* in predicting the "Hero," "Thematic," "Health and Safety," "Policy Prescription and Evaluation," and "Crime and Punishment" frames. This shows that while our induced metaphors do not provide more signal for frame labels than the tweet context, they *enhance* it, suggesting that metaphors provide complementary information beyond the tweet text. These results are statistically significant at $p = 0.05$ after Nadeau-Bengio corrections, calculated with a paired t-test.

Qualitative Analysis We identify and describe source domain clusters with high feature importance for predicting a selection of frames in Table 6. This analysis shows relatively intuitive links between frames and clusters. For example, a cluster encompassing metaphorical uses of "build" is associated with the "Hero" frame, while metaphorical uses of "kill" are associated with both the "Health and Safety" and "Crime and Punishment" frames. However, this analysis also reveals some weaknesses in our clustering process. For example, cluster (14), associated with "make" is not cohesive. The physical act of creating something is applied to a wide range of abstract processes.

Qualitative analysis of Dehumanizing Metaphors We examine the resulting clusters for evidence of the dehumanizing metaphorical source domains widely studied in immigration discourse (Card et al., 2022; Mendelsohn and Budak, 2025). We find that metaphorical expressions which link to common dehumanizing source domains were not sorted cleanly into individual clusters. Instead, many of the dehumanizing metaphors congregated in the same "dehumanization" cluster, encompassing verbs such as "flooding" and "drowning" which maps to the "natural disaster domain" as well as "caged," which maps to the "animal" source domain. Prior work shows that these two source domains tend to be used by speakers with opposing stances on immigration issues, with the "natural disaster" domain favored by the right, and the "animal" domain favored by the left (Mendelsohn and Budak, 2025). Conflating them in a single cluster therefore obscures politically meaningful distinctions, motivating a more sophisticated grouping process for frame analysis in politically charged discourse.

6 Conclusion and Future Work

In this work, we developed an unsupervised method informed by CMT for identifying metaphorical expressions and grouping them to surface conceptual mappings that produce similar framings of

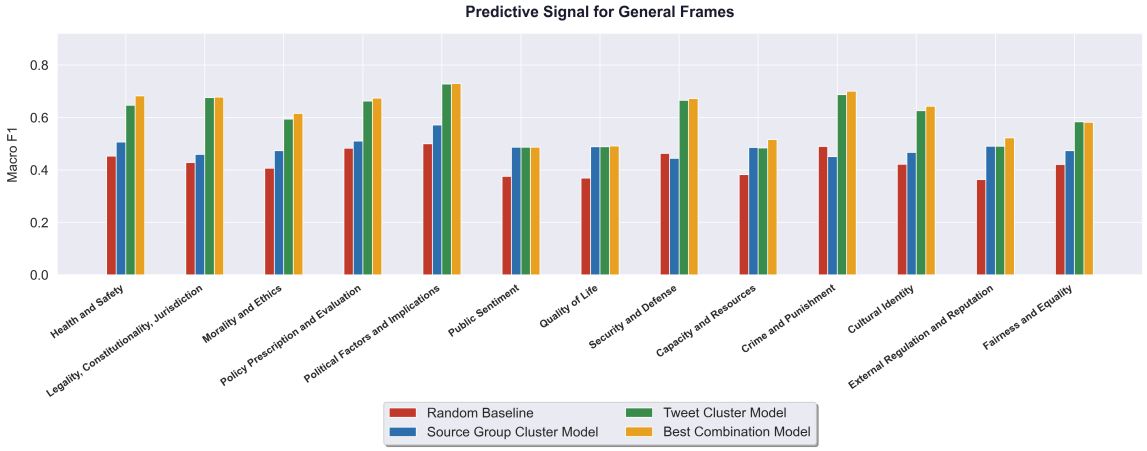


Figure 5: Performance of logistic regression models trained with various feature vectors to detect General Frames.

Frame	Important Cluster Descriptions
Hero	(34) *Vote, count, elect* and *cast* applied to contexts where literal voting is impossible. It is extended metaphorically to imply influence, endorsement, support, or removal.
	(24) *Build*, *create*, *construct*, and *manufacture* applied to abstract outcomes (e.g., nations, crises, policies) Physically bringing something into existence is mapped onto causing or establishing.
	(14) *Make* across a wide range of constructions applied to abstract outcomes. The physical act of making something is mapped onto a sprawling set of non-physical processes.
Health and Safety	(16) *Kill* and *murder* applied to people, groups, policies, institutions. Causing physical death is mapped onto the abstract processes of damaging, undermining, endangering, etc.
	(25) *Hit, attack, target, shoot, blast* (physical violence verbs) are mapped onto criticism, opposition, and strategic effort directed against people, policies, and institutions.
	(6) *Hurt* applied to countries, economies, communities, etc. Physical suffering is mapped onto being adversely affected by policies, economic forces, social conditions, or interpersonal actions.
Crime and Punishment	(16) described above
	(21) *Fix* applied to immigration systems, laws, crises, policies. Mending an object is mapped onto addressing, reforming, or resolving systemic dysfunction in laws, institutions, and social conditions.
	(45) *Break*, *cut*, *tear*, and *rip* applied to laws, systems, families, promises, countries. Destroying objects is mapped onto transgressing rules, dismantling relationships, and reducing resources.

Table 6: Descriptions of clusters important for predicting various frames.

an issue. Our framework achieves strong performance on unsupervised metaphorical expression detection, and the induced conceptual mappings are both quantitatively coherent (as measured by an expression-intrusion task) and qualitatively meaningful, capturing several well-documented source domains from the CMT literature. A case study on online immigration tweets demonstrates that the induced conceptual mappings provide predictive signal for frame labels and offer an interpretable way to study the relationship between conceptual metaphors and framing.

These results constitute a proof-of-concept for an automated conceptual mapping discovery framework. We identify three main directions for fully realizing this vision. First, we require a more thorough evaluation to conclude that our framework is domain-agnostic. Here we present a case study on immigration discourse, a domain where polarized metaphor usage is well-documented. In future efforts, we must determine if we can uncover interesting insights for other contentious domains where metaphor is understudied, such abortion or gun control. Second, we must develop and evaluate a method for mapping source words to distinct, named source concepts (rather than unnamed clusters) for more explicit conceptual mapping discovery. Third, for richer frame analysis, we must develop a method to induce Entman (1993)'s frame elements from the discovered conceptual mappings.

Limitations

This work has two main limitations. First, further evaluation is needed to determine how well our metaphorical phrase detection component generalizes to new domains and types of documents. In addition, we present only one case study on immigration discourse, which offers a limited view of the broader applicability of our approach.

Second, while we use a clustering approach to approximate source domain groups, we do not explicitly map metaphorical source words to distinct source concepts. Similarly, we do not explicitly explore the connection between the induced source and target concepts, which could provide additional insights and potentially guide improvements in metaphor induction.

Ethical Considerations

We recognize that using an LLM-based approach in a politically-charged domain may necessarily result in some biases and thus acknowledge a degree of uncertainty in reporting all results. We leave exploration of the specific biases of the `Qwen3` model in this area to important future work. Additionally, while one goal of this paper is to call attention to and condemn the use of dehumanizing metaphors, we recognize the potential for this line of work to reinforce such biases or for it to be utilized by bad actors.

Acknowledgments

This work was partially supported by a Research & Innovation Office (RIO) Seed Grant from the Office of the Provost and Executive Vice Chancellor for Academic Affairs, and RIO at the University of Colorado Boulder. Any opinions, findings, conclusions, or recommendations expressed in this manuscript are those of the authors and do not necessarily reflect the views of the University or the Funding Offices.

This work utilized the Blanca high performance computing resources at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

References

Rodrigo Agerri, John Barnden, Mark Lee, and Alan Wallington. 2007. Metaphor, inference and domain independent mappings. In *RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING*, page 7.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Jonathan Charteris-Black. 2016. *Fire Metaphors: Discourses of Awe and Authority*, 1 edition. Bloomsbury Publishing.

Lena Dankin, Kfir Bar, and Nachum Dershowitz. 2022. Can yes-no question-answering models be useful for few-shot metaphor detection? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 125–130, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and language*, 89(2):385–392.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Preprint*, arXiv:2509.24866.

Joseph Grady. 1997. Theories are buildings revisited. *Cognitive Linguistics*, 8(4):267–290.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with LDA topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia. Association for Computational Linguistics.

Rebecca M. M. Hicke and Ross Deans Kristensen-McLachlan. 2024. Science is exploration: Computational frontiers for conceptual metaphor theory. *Preprint*, arXiv:2410.08991.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Preprint*, arXiv:2308.01497.

Zoltán Kövecses. 2000. The scope of metaphor. *Topics in English linguistics*, 30:79–92.

Zoltán Kövecses. 2017. *The Routledge handbook of metaphor and language*, chapter Conceptual metaphor theory. Routledge.

Zoltán Kövecses. 2020. *A Brief Outline of "Standard" Conceptual Metaphor Theory and Some Outstanding Issues*, page 1–21. Cambridge University Press.

G. Lakoff and M. Johnson. 2024. *Metaphors We Live By*. University of Chicago Press.

George Lakoff. 1993. The contemporary theory of metaphor.

George Lakoff. 2004. *Don't Think of an Elephant!: Know Your Values and Frame the Debate: the Essential Guide for Progressives*. Chelsea Green Publishing Company.

George Lakoff, Jane Espenson, and Alan Schwartz. 1991. Master metaphor list (technical report). *Cognitive Linguistics Group University of California, Berkeley*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Yu Xi Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. EmbodiedBERT: Cognitively informed metaphor detection incorporating sensorimotor information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16868–16876, Miami, Florida, USA. Association for Computational Linguistics.

James H Martin. 1990. *A computational model of metaphor interpretation*. Academic Press Professional, Inc.

James H Martin. 1994. Metabank: A knowledge-base of metaphoric language conventions. *Computational Intelligence*, 10(2):134–149.

Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Julia Mendelsohn and Ceren Budak. 2025. When people are floods: Analyzing dehumanizing metaphors in immigration discourse with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8079–8103, Vienna, Austria. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, Georgia. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Srinivas Narayanan. 1999. Moving right along: A computational model of metaphoric reasoning about events. *Aaai/iaai*, 121127.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miriam R L Petruck and Ellen K Dodge. 2016. MetaNet: Repository, identification system, and applications. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany. Association for Computational Linguistics.

E. Puraivan, I. Renau, and N. Riquelme. 2024. Metaphor identification and interpretation in corpora with ChatGPT. *SN Computer Science*, 5:976.

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.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477, Vienna, Austria. Association for Computational Linguistics.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697, Uppsala, Sweden. Association for Computational Linguistics.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 978–988, Atlanta, Georgia. Association for Computational Linguistics.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, Beijing, China. Coling 2010 Organizing Committee.

Krishnkant Swarnkar and Anil Kumar Singh. 2018. DiLSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, New Orleans, Louisiana. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. 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 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Yunxiao Wang. 2024. Metaphorical framing of refugees, asylum seekers and immigrants in UKs left and right-wing media. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 18–27, St. Julians, Malta. Association for Computational Linguistics.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

Jintao Zhang, Guoliang Li, and Jinyang Su. 2025. Sage: A framework of precise retrieval for rag. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 1388–1401.

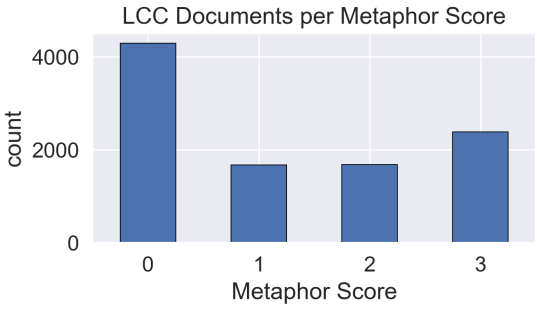


Figure 6: Distribution of metaphor scores in LCC split used to benchmark our metaphor classification method. When we consider all documents with a score of 1 or greater to be metaphorical,

A Appendix

A.1 Additional Metaphor Induction Details

In Table 7 we present the five constructional patterns used to filter `(verb, noun)` pairs which can grammatically induce a metaphor. In Table 7 we show a sample of LLM metaphor salience score explanations, which are clustered to group metaphors with similar source domains.

A.2 Additional Dataset Details

Table 8 and Figure 6 shows the distribution of hand-annotated metaphoricity scores for the 10,018 documents from the LCC dataset used to evaluate our metaphor classification method. Table 9 shows the hand-annotated target domains for this set.

A.3 Additional Binary Metaphor Classification Results

Table 10 shows a detailed classification report of our implementation of the supervised metaphor classification component presented in (Wang, 2024). Our classifier achieves an F1 score of 0.809 on all positive classes (metaphor score greater than 0), resulting in a macro F1 score of 0.810 when the task is viewed from the binary classification perspective. These results are very similar to those achieved by (Wang, 2024), who presented an F1

Construction Pattern	Example
S_VERB-*dobj*-T_NOUN	"I gave you that idea."
T_NOUN-*nsubj*-S_VERB	"Ideas travel quickly."
S_VERB-*agent*-ADP-*pobj*-T_NOUN	"That idea gnaws at my mind."
S_VERB-*amod*-T_NOUN	"That idea sings."
T_NOUN-*nsubjpass*-S_VERB	"My idea was shot down."

Table 7: Constructional patterns used to filter SDPs which may invoke a metaphor. Italics in the "Construction Pattern" column represent relation types between tokens. Source verbs and target nouns are underlined in the "Example" column.

LCC Score	Count
0	4,285
1	1,671
2	1,677
3	2,385
1 $\leq$	5,733
total	10,018

Table 8: Distribution of metaphor scores for the portion of the LCC dataset used for evaluation.

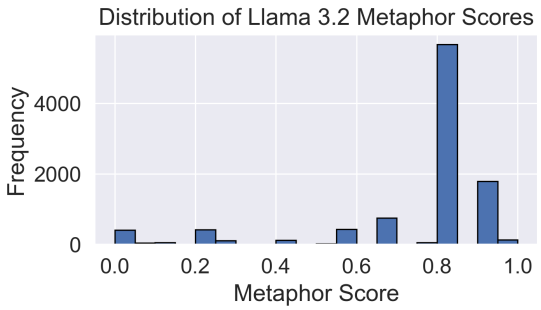


Figure 7: Llama 3.2-generated metaphor salience scores for (`source verb, target noun`) pairs extracted from the LCC dataset.

score of 0.86 on samples with a metaphor score of 0 and 0.83 on those with a score greater than 0.

A.4 Additional Metaphoricity Threshold Analysis

Tables 7 and 8 show the distribution metaphor salience scores for (`source verb, target noun`) pairs in the LCC dataset. Table 11 shows the performance of the scores on the LCC dataset for metaphor classification at various thresholds.

LCC Target Domain	Count
guns	2013
mental concepts	1577
government	1025
democracy	673
elections	613
bureaucracy	603
poverty	581
taxation	498
wealth	430
religion	409
money	313
disease	231
migration	152
intellectual property	125
taxpayers	104
taxes	99
islamic	86
terrorism	85
gun debate groups	79
gun rights	78
politicians	78
marriage	45
drug trafficking	40
welfare	23
debt	16
climate change	15
abortion	13
demographics	11
control of guns	3
total	10018

Table 9: Hand-annotated target domains for all excerpts in the English LCC dataset used for evaluation. The dataset spans a wide variety of topics, making it an optimal resource for evaluating computational methods for general metaphor detection.

	precision	recall	f1-score	support
0	0.778	0.848	0.812	2403
1	0.438	0.347	0.387	654
2	0.399	0.403	0.401	705
3	0.718	0.668	0.692	1218
accuracy	0.675	0.675	0.675	0.675
macro avg	0.583	0.566	0.573	4980
weighted avg	0.665	0.675	0.669	4980

Table 10: RoBERTa classifier performance trained and tested on the complete LCC dataset. Here the task is formulated as a multi-class classication task to predict metaphor scores 0-3.

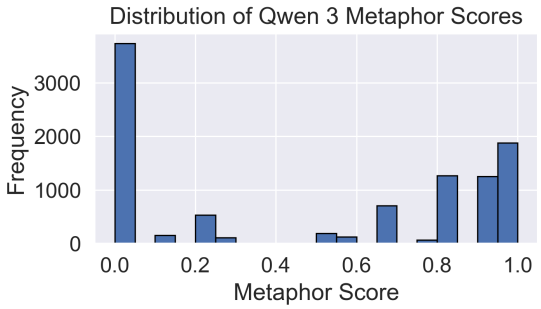


Figure 8: Qwen 3-generated metaphor salience scores for (`source verb, target noun`) pairs extracted from the LCC dataset.

Threshold	**0.1**	**0.2**	**0.3**	**0.4**	0.5	0.6	0.7	0.8	0.9	1
Random Baseline	**0.699**	0.663	0.623	0.579	0.525	0.462	0.386	0.291	0.158	0
Llama 3.2	0.732	**0.733**	0.732	0.732	0.731	0.731	0.725	0.705	0.333	0.027
Qwen 3	0.787	0.790	**0.792**	0.791	0.791	0.786	0.778	0.733	0.613	0.426

Table 11: Macro F1 performance of our method on the LCC dataset at different score thresholds. LCC documents with a metaphorical score of 1 or higher are considered metaphorical. Documents with an LLM score above the threshold are predicted to be metaphorical.

System Prompt

You are an annotator who is developing a dataset for measuring metaphors. Your response should be in JSON format with the key 'metaphor_salience_score' and a value between 0 and 1 that indicates how salient the metaphor invoked by the specified verb is. Additionally, in the JSON, you should indicate with the key 'explanation' the justifications for your decision.

Desired JSON: {'metaphor_salience_score': float, 'explanation': 'your reasoning for the answer'}.

Do not generate anything else.

User Prompt

Analyze the use of the specified verb in the sentence provided. Focus only on determining whether this verb, in its specific use within the sentence, is used literally, that is, describing the physical action, or whether it is used metaphorically, where the use of the verb transcends its original meaning without referring directly to a physical action, such as, for example, giving some kind of personification or animalization of an object. It is important to distinguish the specific lexical analysis of the verb from any broader metaphorical interpretation that may arise from comparisons or conceptual equivalences present in the sentence.

Please give a score between 0 and 1 (inclusive) that indicates how salient the metaphoricity of the specified verb is. 0 indicates that the verb is very obviously used literally and 1 indicates that the verb is very obviously used metaphorically. A score of 0.5 indicates that the metaphor would only be noticed by half of human readers. Please provide your evaluation focusing solely on the specified verb.

Prompt A.1: Metaphor salience score extraction.

System Prompt

You are a grammar expert building a dataset for noun classification. Your task is to determine whether the target noun is a person (or group of people), place, or thing.

Desired JSON: {'noun_classification': '[person/place/thing]'}.

Do not generate anything else.

User Prompt

Is the the target noun specified below a person (or group of people), place or thing? You may use the context sentence to help make your decision, but please provide your answer focusing solely on the specified noun.

TARGET WORD: <target word to classify>
SENTENCE: <context sentence of target word>

Prompt A.2: Assigning person, place, or thing classifications to target words.

System Prompt

You are an annotator who is developing a dataset for analyzing metaphors. Your task is to characterize how metaphors are being used to characterize a particular person or group of people. For each submission, you will identify the target group that matches the provided metaphor target word.

ANNOTATION GUIDELINES
- Map the identified nouns to the most appropriate group.
- If no group from the provided list adequately represents the target, assign it to 'Other'. This includes nouns that are tangentially related but don't fit well into any specific predefined category, as well as nouns from completely different domains or contexts.

Desired JSON: {'target_group': ['Identified character group from predefined list']}.

Do not generate anything else.

User Prompt

Please provide an evaluation focusing solely on the metaphor target noun.

DOMAIN: <corpus domain>
GROUPS: <identified target domains>
TARGET WORD: <target word to classify>
SENTENCE: <context sentence of target word>

Prompt A.3: Assigning target domains to target words.

Simulating Social Attitudes with LLMs: Accuracy, Demographic Effects, and Refusal Behavior in the Sensitive Domain of Suicide Prevention

Cristina J. Perez[1] **Michael P. Vasquez, Jr.**[1] **Philippe J. Giabbanelli**[2] **Patrick Y. Wu**[3]
[1] Department of Mathematics and Statistics, American University
[2] Virginia Modeling, Analysis, and Simulation Center, Old Dominion University
[3] Department of Computer Science, American University
cristinajoannaperez@gmail.com mv0317a@american.edu pgiabban@odu.edu patrickwu@american.edu

Abstract

Large language models (LLMs) are increasingly used to simulate public opinion, yet their validity in sensitive policy domains remains underexplored. We evaluate whether LLMs can reproduce attitudes toward suicide prevention policies using 32 questions drawn from seven nationally representative U.S. surveys (2023-2025). We systematically vary demographic conditioning (race/ethnicity, gender, age, education, income, party), prompt framing (direct elicitation, respondent embodiment, specialist embodiment), and model architecture (GPT-5 Nano, DeepSeek V3.2, Meta Llama 3.1 8B, Mistral Small 24B). Across 811,560 prompts, the mean absolute error—the average gap between predicted and human response distributions—is 23 percentage points. We also find that LLM responses to demographic-conditioned prompts diverge substantially from prompts without demographic information. In short, what distribution LLMs draw on when generating responses to sensitive polling questions remains unclear. Model choice matters more than framing for accuracy, whereas refusal behavior varies sharply across models and prompt designs. Our findings highlight the limitations of LLMs for social simulation in the context of sensitive topics.

1 Introduction

Suicide is one of the leading causes of death in the US, with 1 death every 11 minutes based on 2023 data (Centers for Disease Control and Prevention, 2025). Suicide is a multifactorial issue with risk and preventative factors at several levels, from social drivers (e.g., economic policies, discrimination) to community (e.g., exposure to violence, access to mental healthcare) and individuals (e.g., mental health issues). Preventing suicide thus requires a package of interventions, which would be enacted by policymakers in part based on perceived support from constituents (Purtle et al., 2025). While constituents may agree that suicide is preventable and a public emergency, opinions can diverge widely on which actions should be taken (Munsch et al., 2020). Understanding public attitudes toward suicide prevention policies is thus essential to assess the political feasibility of interventions. However, measuring these attitudes is methodologically challenging because many of the most consequential policy levers (e.g., firearm regulation, public financing of healthcare) are politically and morally charged. Survey research on sensitive topics (Dixon et al., 2020; Stone and McGinty, 2018) shows that respondents may strategically edit answers, refuse items, or give mode-dependent responses to avoid embarrassment or perceived repercussions, which can bias estimates of policy support. There is also evidence that opinions on suicide-relevant policies can systematically vary by socio-demographic groups: women support safe storage laws more than men (Crifasi et al., 2021), higher education is associated with more support for government spending on mental healthcare (Barry and McGinty, 2014), and racial minorities have a higher support for school mental health programs (Hemauer and Warner, 2025). Such patterns echo broader survey research demonstrating that social desirability and reporting tendencies differ by group characteristics, reinforcing the need to analyze attitudes within key demographic strata.

Recent advances in large language models (LLMs) have opened a new methodological possibility: using these models to simulate survey respondents and other social-scientific agents (Horton et al., 2023). Because LLMs are trained on a vast corpus of digital text and media, they embed a great deal of knowledge about how people from different backgrounds discuss sensitive topics, including suicide prevention. A growing body of literature has explored whether LLMs can reproduce aggregate patterns of public opinion on political, social, and health-related issues (see, e.g., Argyle

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 176–189

et al., 2023; Lee et al., 2024; Jiang et al., 2025). If LLMs can generate responses that approximate real survey distributions, particularly when conditioned on demographic attitudes, they could serve as a complementary tool to study attitudes that shape the political feasibility of suicide prevention policies across socio-demographic groups. The ability to repeatedly run virtual surveys at low or no cost via LLMs also enables us to explore nuanced policy parameters: for instance, there are gradients of support rather than binary positions when it comes to firearm regulation (Anestis et al., 2025), and the fact that many are unwilling to pay higher taxes is more nuanced when we consider budget trade-offs (Johnson et al., 2021). As a result, LLMs could help to systematically assess tolerance thresholds.

However, these benefits can only be unlocked if LLMs can faithfully simulate public opinions on a topic as sensitive as suicide. Prior work on LLM opinion simulation has focused on attitudes to political domains where information is relatively well-represented in the underlying training data. In particular, Chi and Lei (2026) developed a framework that uses LLMs to augment surveys on suicide, producing a mental-health screening score shaped by representational risk (who the LLM respondents are) and response risk (what they say). However, there is a paucity of studies that examine policy attitude distributions through LLMs for suicide prevention. This is a challenging task for LLMs, as the real-world difficulty of eliciting attitudes about suicide prevention is compounded by technical challenges: guardrails are often triggered on topics related to self-harm (Gandee et al., 2024) and the framing of the prompt (e.g., 'you *are* an individual from a socio-demographic group' vs. 'you *study* individuals') can substantially alter the distribution and accuracy of its outputs. Socio-demographic persona assignment can also induce explicit abstentions and implicit reasoning errors in LLM outputs (Gupta et al., 2024). To the best of our knowledge, how well LLMs can simulate attitudes to suicide prevention policies has not been systematically examined.

Our main contribution is to provide the first systematic study on how well LLMs can simulate attitudes to suicide prevention policies. To achieve this goal, we evaluate LLM-simulated survey responses against ground-truth data from seven surveys of public attitudes toward suicide prevention. We systematically vary three dimensions: (1) the *demographic profile* assigned to the simulated respondent or the simulated expert on public opinion, including race/ethnicity, gender, age, education, income, and political party identification; (2) the *prompting framing* used to elicit responses, through direct elicitation, expert embodiment, or respondent personas; and (3) the *LLM architecture*, considering four choices (GPT-5 nano, DeepSeek-V3.2, Qwen3-32B, Meta Llama 3.1 8B Instruct). This design supports three research questions:

RQ1 How accurately do LLMs reproduce the average and range of attitudes on suicide prevention across demographic groups?

RQ2 How do prompt framing (direct elicitation, expert embodiment, respondent embodiment) and model type affect the fidelity of LLM-simulated suicide attitude responses?

RQ3 Does the rate of LLM response refusals depend on prompt framing or model type?

The remainder of this paper is structured as follows. As our paper is at the confluence of surveying suicide prevention policies and simulations via LLMs, Section 2 grounds our approach in both strands of literature. Then, Section 3 details our methods from data collection and prompt generation to the analysis with respect to each RQ. Our results are presented in Section 4 and contextualized along with limitations in Section 5.

2 Related Work

2.1 Suicide Prevention: Policies and Opinions

Using the National Survey on Drug Use and Health, recent analyses point to an increase of 21.7% in suicidal ideation from 2015 to 2019, with a significant increase of 44.6% among young adults (Samples et al., 2025). When considering the high cost of lost life years together with reduced quality of life and medical care costs, the annual economic cost of suicides in the US averages $484 billion (Peterson et al., 2024). There is thus a pressing public need to prevent suicide across all stages, starting with reducing suicidal ideation (e.g., by creating protective environments), avoiding suicide attempts (e.g., through gatekeepers training and access to mental healthcare), and preventing access to highly lethal means (e.g., locked firearms, blister packages). Significant efforts have thus been devoted to proposing packages of interventions, such as the framework from the Centers for Disease Control and Prevention (2022) articulated around seven high-level strategies (e.g., promoting health connections in schools and communities). However,

enacting these changes requires political action, which may depend on constituents' willingness to support certain items (Purtle et al., 2025). While 91% of U.S. adults believe suicide is *preventable*, support varies when considering *how* to prevent suicide. For instance, creating protective environments includes reducing access to lethal means, and particularly firearms (a highly lethal method), but individuals may object to such an intervention (particularly in firearm-owning households) by considering that another lethal method would be used anyway (Barber and Miller, 2014; Conner et al., 2022). A tension also exists when considering how to improve access to care or how to identify people at risk: the associated proposals (e.g., Medical expansion, funding the 988 Suicide and Crisis Lifeline, school-based services) imply either higher public spending or reallocated budgets. Individuals may support treatment but not higher taxes, or consider that it should be handled privately rather than by the government, or view other crises as more pressing (Munsch et al., 2020; Shields et al., 2025).

Understanding these tensions requires consideration of how public opinion polls are constructed and how those choices shape the broader narrative surrounding suicide-prevention strategies. Three aspects are particularly important: *sampling* decides whose opinions are recorded, *scope* dictates how general or policy-specific a survey's focus will be, and *framing* can influence how the public responds to a question or interprets the results. Broad national surveys, such as the AFSP Mental Health and Suicide Prevention Poll and Duke Press' 988 Awareness Survey, assess general attitudes toward mental health and crisis services among the broader American public. More targeted surveys like YouGov Gun Ownership Survey and Science Direct's Help-Seeking Preferences Survey focus on specific subgroups and behavioral contexts to capture attitudes to firearm access and preferred sources of support.

2.2 Simulating Public Opinions via Large Language Models

There is extensive literature on simulating public opinions using LLMs, an approach often referred to as "silicon sampling" (Argyle et al., 2023). Silicon sampling involves creating prompts that include background information about a simulated respondent and a survey question. Given that LLMs are trained on vast amounts of digital media and data, they embed extensive knowledge of how people from different backgrounds discuss and believe about various topics. Argyle et al. (2023) argued that this information is fine-grained and demographically correlated, meaning appropriate prompting with specific demographic information can elicit and emulate response distributions from diverse human subgroups.

Researchers have used this approach across many areas, particularly in politics (Argyle et al., 2023; Simmons and Hare, 2023; Jiang et al., 2025). These studies generally find that LLM-generated responses align closely with human judgments and survey data. However, Zhong et al. (2025b) note that responses depend on the model used and the phrasing of the prompt. In addition, Liu et al. (2024) show that persona-steered generations can default to demographic stereotypes for multifaceted or incongruous personas. Such studies motivate us to consider several LLMs and prompt framings.

Researchers also noted significant limitations of silicon sampling. Bisbee et al. (2024) found that while GPT-generated average scores closely corresponded to those from the American National Election Survey, GPT outputs are less varied than human responses. Variance compression was also noted by Tjuatja et al. (2024), who found that LLMs tend to homogenize responses. Qu and Wang (2024) and Santurkar et al. (2023) showed that LLMs tend to better reflect the viewpoints of educated, affluent, English-speaking, Western populations while underrepresenting others.

Despite this growing body of work, no study has systematically examined how well LLMs can simulate attitudes toward suicide prevention policies. The closely related work by Chi and Lei (2026) developed Compassionate AI Survey Augmentation (CASA), a framework that uses LLMs to augment surveys on attitudes toward suicide. CASA reduced the emotional burden of answering sensitive questions but also introduced risks of demographic misrepresentation and response bias. This related work does not systematically investigate how LLM-generated responses to questions about suicide prevention policies vary by demographic framing, prompt design, or model choice.

3 Methodology

Our process has three main steps (Figure 1): we collect questions across surveys and align the sociodemographic profiles of respondents (Section 3.1), then we generate prompts so that diverse LLMs

choose an answer to survey items based on different socio-demographic profiles and phrasing of the tasks (Section 3.2), and finally we extract and analyze the LLMs' answers with respect to our three research questions (Section 3.3).

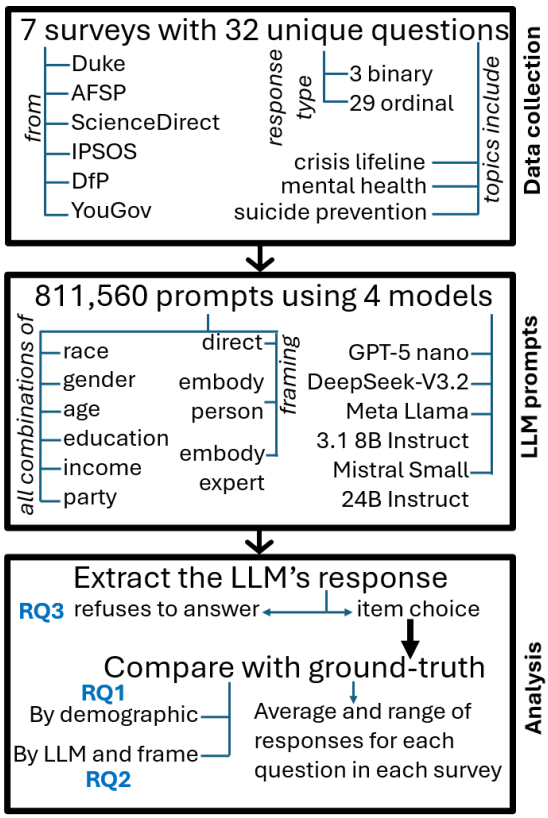


Figure 1: Overview of our methods, emphasizing the structure of our data and the design of our prompts.

3.1 Data Collection and Pre-Processing

We selected surveys from January 2023[1] to September 2025 using three inclusion criteria: (1) nationally representative sample of US adults 18 years or older (e.g., surveys with location-based answers were not included); (2) includes questions on suicide prevention policies, either at the individual level (e.g school mental health screening) or at the societal level (e.g affordable housing to combat suicide); and (3) answers are in binary or scale response options (e.g., no free text response). We used multiple search databases, including Google Scholar and the Roper Center for Public Opinion, as well as snowball sampling from reports from nonprofit organizations and public policy research centers. The search terms were "(suicide prevention policy) OR (mental health polls) OR (suicide prevention AND public opinion polls)".

Our process yielded seven public opinion surveys. We only included questions about suicide prevention policy, so other items, such as the morality of suicide, were excluded. This filtering was done manually across two annotators. When identical questions appeared in multiple surveys, we kept the most recent version to reflect the latest opinions. As a result, we had 32 survey questions (Table 1), answered by approximately 1,000 to 5,000 respondents with all margins of error below five percent.

The *categories* of survey respondents were based on demographic categories and party affiliation (Table 2), thus forming the groups on which the LLM prompts are created in the next section. We considered race/ethnicity[2], gender, age, education, income, and political party. These sociodemographic categories and party affiliation have been found in previous studies to be associated with opinions on suicide prevention (Hemauer and Warner, 2025) and with the use of prevention services (Purtle et al., 2024). Note that three surveys did not use some of these categories; thus we use three fewer categories for the Duke University Press survey (age, income, and party), and one fewer category for both the Harris Poll and Data for Progress surveys (party and income, respectively). While some surveys included more demographic categories (e.g., AFSP's inclusion of an employment status variable), they were not retained as we maximized shared categories across surveys.

The *values* for each category were transformed as follows. We removed values that were insufficiently used across surveys to yield robust estimates for suicide prevention, resulting in the exclusion of respondents who self-identified as Asian, Pacific Islander/Hawaiian Native, or Native American[3]. The chosen standardized age groups, 18-29, 30-44, 45-59, and 60+, fit most of the surveys and aligned with the bin groupings used by the Current Population Survey (CPS). The Harris Poll survey, which sampled adults 18 and older and applied age

[1]The beginning of the data collection was selected to provide sufficient time for respondents to become aware of the three-digit suicide and crisis hotline, which launched nationwide in the US on July 16, 2022.

[2]Although race and ethnicity are orthogonal categories, they are combined to emulate the original survey structure.

[3]Processing these different racial groups under the same category is problematic as they face vastly different challenges with respect to suicide: Non-Hispanic Asians have the lowest rate of suicide fatality (6.5 per 100,000) while Non-Hispanic American Indian/Alaska Native have the highest rate (23.8 per 100,000) (Centers for Disease Control and Prevention, 2025). Removing this heterogeneous category affected at most 13% of survey respondents and at minimum 0%.

Survey Name	Survey Conductor	Date	Respondents	Total Questions	Questions Used
AFSP Mental Health Survey	American Foundation for Suicide Prevention	July 2024	4,394	350; question pool	4
Voters Show Wide, Bipartisan Support for Policies to Improve Student Mental Health	Data for Progress (DfP)	Oct 2024	1,223	15	6
Public Attitudes, Inequities, and Polarization in the Launch of 988 Suicide and Crisis Lifeline	Duke University Press, Journal of Health Politics, Policy and Law	Jun 2024	5,482	3	1
988 Suicide & Crisis Lifeline Awareness	Ipsos KnowledgePanel	Jun 2025	2,049	17; multipart	10
Demographic Variation in Preferred Sources for Suicide Help-Seeking	ScienceDirect	Oct 2024	5,058	5	5
Suicide Prevention	YouGov	Jun 2023	1,000	12; multipart	3
Biden and Trump Handling of Problems	YouGov	Jun 2024	1,110	5; multipart	3

Table 1: Our study identified seven surveys as ground truth and used 32 unique questions. The detailed list of survey questions and respondent characteristics is provided in our online repository as **S1 Survey Data**.

weighting, reported more granular age data; these categories were condensed and approximated using CPS bins as reference. The exception is the Data for Progress survey, which only reported two categories: under 45 and 45+. This survey is thus handled separately in our process by constructing prompts for only two age groups (next section) and analyzing them with respect to these two groups.

Category	Values
race	non-Hispanic White, non-Hispanic Black, Hispanic
gender	male, female, other
age	18–29, 30–44, 45–59, 60+
education	do not have a high school diploma, high school diploma or equivalent, some college, bachelor's degree or higher
income	$< \$50k$, \$50k–\$100k, $> \$100k$
party	Democrat, Independent, Republican

Table 2: Survey demographics and categories.

3.2 Prompt Generation Pipeline

We used three prompt framings: *direct elicitation* (asking the LLM to answer the question without demographics), *embodying a respondent* based on demographics, or *embodying an expert* who answers on behalf of an individual with given demographics. These three framings (exemplified in Table 3) were applied across all applicable combinations of demographic categories (Table 2) and for each of the 32 unique survey items, resulting in 54,104 unique prompts (Table 4). Each prompt was run three times to account for the non-deterministic nature of the LLMs. Figures 4 and 5 in the Appendix show the average standard deviation across direct-prompt runs by question and the standard deviation for each question–model combination. Overall, most questions exhibited relatively low variability across runs. We ran the prompts on four LLMs (GPT-5 nano, DeepSeek-V3.2, Meta Llama 3.1 8B Instruct, and Mistral Small 24B Instruct), each set to its default parameter values as recommended by the provider (i.e., the default values are model-specific). Since GPT-5 Nano has a reasoning parameter that should be tuned, we considered two levels of reasoning (minimal and high). As a result, we ran a total of 811,560 prompts, calculated as 54,104 prompts per model across 5 models, accounting for both GPT-5 Nano reasoning levels, for 3 iterations each ($54,104 \times 5 \times 3$).

The prompts in Table 3 were gradually refined based on our initial tests. As noted in Giabbanelli (2026), prompts need to *decompose* the task instead of asking for too many tasks at once, and we need to be specific about expectations for the *output format* to avoid extensive (and error-prone) post-processing. Initially, we asked the LLM to answer a survey question as it was given to the participants. However, the LLM's output would mix the response option with the justification, leading us to specify (in the last sentence of each prompt) that the response must be separated from the rationale. At first, we also asked the LLM to answer survey questions such as "Please rate, on a scale of 1-7, the likelihood of you reaching out to each of the following sources [...]: 988 Lifeline, A friend or family member, Someone in my religious network [...]". We split such questions with multiple items by providing them to the LLM one at a time, e.g. "Please rate, on a scale of 1-7, the likelihood of you reaching out to [...] the 988 Lifeline."

1. Demographics (Direct): You are a public opinion expert. You are presented with a survey question asking how a US citizen would respond with one of the following responses. The question you are answering is: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Select a response option from the following: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.
2. Demographics (Embody): You are a non-Hispanic White, identifying as female for gender, aged between 18 and 29, do not have a high school diploma, earning less than $50K, identifying as a Democrat voter, living in the U.S. Answer the following question: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Response options: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.
3. Demographics (Specialist): You are a public opinion specialist. How would a person with the following demographic information answer the survey question below? You are a non-Hispanic White, identifying as female for gender, aged between 18 and 29, do not have a high school diploma, earning less than $50K, identifying as a Democrat voter, living in the U.S. Answer the following question: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Response options: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.

Table 3: We considered three prompt framings, shown in the following order: direct elicitation, respondent embodiment, and expert embodiment. Dynamic elements from demographics and surveys are shown in blue.

Survey	Direct	Embody	Total
Duke	1	36	73
AFSP	4	1,728	3,460
Science Direct	5	6,480	12,965
IPSOS	10	12,960	25,930
DfP	6	2,592	5,190
YouGov	6	3,240	6,486
Total	32	27,036	54,104

Table 4: Across 7 surveys, and for each unique question, we generate prompts based on three framings (a direct one without demographics; two embodiments based on demographics), thus Total = Direct+2×Embody. This amount represents the full Cartesian product of demographic attributes. Note that we used two YouGov surveys, per Table 1.

3.3 Analysis

We extracted LLM responses using pattern matching to identify valid answers in the expected format (selected option and brief reasoning). When this failed, we inferred the response if exactly one option was mentioned in the output. We flagged refusals based on common phrases (e.g., "I cannot tailor") [4]. We manually reviewed and coded the fewer than 20 cases where neither a valid option nor a refusal was detected. The resulting extracted response variable was used in all subsequent regression analyses, as explained in the next section.

To analyze our simulation results for RQ1, we examine the mean absolute error, total variation distance, and the Jensen-Shannon divergence between the distributions of the LLM's predictions and the ground truth. The ground-truth distribution for each survey question is the set of human response percentages reported in the original poll. Specifically, this is the share of respondents in each demographic subgroup who selected each response option (e.g., the percentage of non-Hispanic White respondents who answered "strongly support"). Each source survey reports these breakdowns as marginals along a single demographic subgroup; joint distributions across multiple demographics are not published. We compute the corresponding LLM distribution by aggregating model responses across all prompts whose persona belongs to that subgroup, averaging equally over all combinations of the other demographics. For RQ2, we use a two-way ANOVA to assess whether the LLM used and the prompt framing affect the absolute errors of simulated survey responses, along with analyzing mean absolute error by model. Lastly, for RQ3, we examine refusal rates by LLM and prompt framing.

4 Results

4.1 RQ1: Range and attitudes of responses

Across all LLM answers aggregated over 28 questions[5], the mean absolute error (i.e., the average *magnitude* of the difference between model predictions and human response percentages by question)

[4] For example of refusal cases, please see Appendix A.5.

[5] 4 survey questions only reported adjusted odds ratios; they were excluded in analyses involving ground truth comparisons.

is 23 percentage points with a standard deviation of 11 percentage points.[6] The *error is more frequently in the range of 15 to 28 percentage points* (the interquartile range) as shown in Table 5. Figure 3 in the Appendix shows the overall distribution of mean absolute error, total variation distance, and Jensen-Shannon divergence across all LLM-simulated responses.

Demo.	MAE	TVD	JSD
Age	$.22 \pm .11$	$.38 \pm .17$	$.23 \pm .17$
Education	$.22 \pm .11$	$.38 \pm .17$	$.24 \pm .17$
Gender	$.22 \pm .10$	$.39 \pm .17$	$.25 \pm .18$
Income	$.24 \pm .10$	$.43 \pm .16$	$.29 \pm .17$
Party	$.25 \pm .12$	$.39 \pm .16$	$.22 \pm .16$

Table 5: Mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) of the distribution of model predictions and the human response percentages, averaged by question, with standard deviations.

LLMs perform similarly across demographic categories, indicating that the models are not systematically better at predicting responses for any particular group. While income-based predictions exhibit slightly higher TVD and JSD values than age and party affiliation, the large standard deviations indicate substantial question-to-question variability, suggesting that prediction quality is driven more by individual questions than by demographic category. Table 9 in the Appendix shows MAE broken out by demographic and model; again, there are no substantial differences across demographic categories.

To contextualize these distributional errors, we compute the proportion of persona-conditioned responses that match the LLM's modal direct response (i.e., the most common answer when the model is prompted without demographic information). Match rates were similar across demographic categories, ranging from 52% (income) to 55% (age and race), with education and party at 54%. In other words, *persona-conditioned responses matched the modal response to unconditioned prompts only about half the time*.

4.2 RQ2: Prompt framing and model types

While the LLMs share distributions of the mean absolute errors (Figure 2), *Mistral Small 24B has consistently higher MAE.*

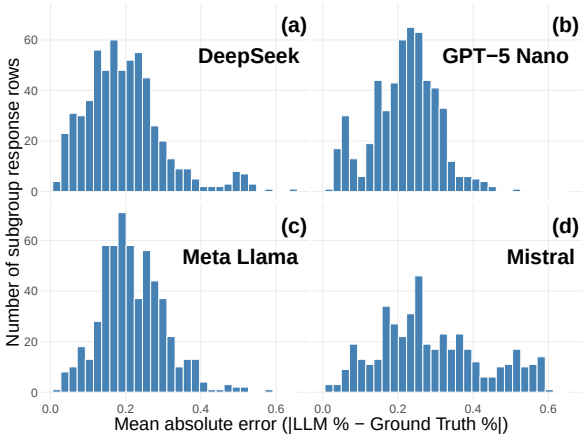


Figure 2: Distribution of mean absolute errors in LLMs.

Performing a two-way ANOVA, we find that the accuracy of the simulated survey responses (measured by mean absolute error) is significantly influenced by both the choice of language model ($p < .001$) and the prompt framing ($p < .001$). There is also a highly significant interaction effect between the model and the prompt frame ($p < .001$). In other words, *the impact of a specific prompt frame on reducing or increasing simulation error depends on the LLM used*.

The difference between framing is smaller than the difference between LLMs (Table 6). For example, Mistral drops from 33.0 to 26.1 across framings (≈7-point difference), but the gap between Mistral and DeepSeek V3.2 under embody is over 12 points. The *LLM choice thus matters more than framing for average accuracy*.

4.3 RQ3: Response refusals

For GPT-5 Nano, DeepSeek V3.2, and Mistral Small, refusal rates were effectively zero, ranging from 0% to approximately 0.30% across prompt frames (Table 7). In contrast, Meta Llama exhibited a substantially elevated overall refusal rate of 28.1%. Breakdowns by framing reveal that this effect was strongly condition-dependent: refusal rates for direct and embody prompts were fairly high, 44% and 48% respectively, whereas the specialist frame produced a lower refusal rate of roughly 8%. Together, these findings indicate that elevated refusal behavior was isolated to a single model and was sensitive to prompt framing. We provide examples of refusal responses in Appendix A.5.

[6]The human response percentages by question can be found in our replication materials; the link to our replication materials can be found in Appendix A.1.

	Large Language Models			
Framing	DeepSeek V3.2	GPT-5 Nano	Meta Llama 3.1 8B	Mistral Small 24B
Embody	20.5 ± 10.2	21.5 ± 9.1	21.6 ± 8.4	33.0 ± 15.0
Specialist	18.9 ± 10.6	22.2 ± 8.2	22.0 ± 8.7	26.1 ± 12.8
Direct	54.0 ± 18.3	37.8 ± 21.9	44.5 ± 29.1	46.0 ± 31.0

Table 6: Average and standard deviation of mean absolute error across LLMs and framings. MAE for the 'Direct' framing was calculated by comparing each model's direct responses against the ground truth distribution averaged across demographic subgroups.

Model	**Framing**	**# prompts**	**%refuse**
Meta Llama 3.1 8B	Embody	81,108	48.1
	Direct	96	44.8
	Specialist	81,108	8.08
GPT-5 Nano	Embody	162,216	0.295
	Direct	192	0
	Specialist	162,216	0.267
Deep-Seek V3.2	Embody	81,108	0
	Direct	96	0
	Specialist	81,108	0
Mistral Small 24B	Embody	81,108	0
	Direct	96	0
	Specialist	81,108	0

Table 7: Total prompts and refusal percentage by model and framing.

5 Discussion

The literature on silicon sampling has reported strong performance on simulated LLM responses and human responses (see, e.g., Argyle et al., 2023; Jiang et al., 2025), though prior work has also noted that response variation tends to be substantially lower than in human samples (Bisbee et al., 2024; Zhong et al., 2025a). Examining silicon sampling in the context of the sensitive topic of suicide prevention policies, we addressed three research questions: how well LLM-simulated responses matched human responses, how prompt framing and model types affected these responses, and whether LLMs refused to respond to such questions.

Our findings diverge from prior work on both counts. Our analysis of RQ1 finds that, on average, LLM answers differ from the ground truth by more than 20 percentage points, with large standard deviations; TVD and JSD further confirm this finding. In contrast, Zhong et al. (2025a) reported a difference of 6 percentage points between synthetic outputs and human respondents on other political topics. To rule out the possibility that safety guardrails drive the model towards a default response regardless of the demographic prompt, we compared responses to the modal answer from a prompt with no demographic information (the "Direct" configuration as specified in Table 3). We find that the LLM's responses also vastly differ from the modal answer. Through RQ2, we find similar error magnitudes across LLMs and prompt framings. Thus, it remains unclear what underlying distribution the LLMs are drawing on when generating their responses to sensitive polling questions, calling into question the utility of silicon sampling for topics such as suicide prevention.

Future work could further examine the impact of location, choice of LLM, and socio-demographics. First, suicide prevention initiatives vary significantly in content and depth across states. This contrast can be exemplified between the policy documents of the Wyoming Department of Health (2024), consisting of four pages (four infographics) that acknowledge that two-thirds of suicides involve a firearm but made no explicit policy recommendations in this regard, and the California Department of Public Health (2022), whose plan spans almost 80 pages and covers firearms extensively. Studies have shown that there are also state-level differences in the extent to which their legislature and their 'citizen ideology' support suicide prevention actions (Kenter et al., 2022). It would thus be of particular interest to use LLMs to examine how constituents react to the plans proposed in their state, and potentially *identify evidence-based actions that are not in the plan yet would be supported*.

Second, while we covered LLMs from four different providers, it is possible that other LLMs may yield different results (e.g., in accuracy or refusal to answer). In particular, LLMs may have different guardrails when it comes to sensitive topics such as suicide and firearms, and these guardrails may be triggered differently based on the IP from which the prompts originate (since guardrails can depend on local laws). The spectrum of guardrails is wide: some LLMs refuse to engage on the topic of suicide (even to discuss prevention policies), while others are now cited in wrongful-death lawsuits for convincing users to die by suicide (Jargon, 2026). A challenge for this line of research is that guardrails can change quickly, for example, in relation to news events. For instance, Grok was seen as a "low safety-guardrail model" in January 2026 (Teferra et al., 2026), but has since changed significantly. This opens up the possibility to study responses from LLMs on suicide prevention initiatives across demographics from a longitudinal perspective.

Finally, we considered commonly used sociodemographic attributes (i.e., race, gender, age, education, income, party) that were available across most surveys in order to provide ground-truth data. However, there are other markers of attitudes relevant for suicide prevention that may be more divisive or less commonly seen in training data, which may lead to more variability when using LLMs. For example, "higher religiosity is consistently associated with lower suicide risk among heterosexual people" (e.g., suicide is forbidden), but religiosity can be harmful for sexual minorities. Prompting LLMs to consider the intersection of religion, suicide, firearms, and sexual orientation would combine several highly sensitive topics (Park and Hsieh, 2023). An intersectional examination (Forrest et al., 2023) would be of particular interest to examine whether biases or refusals to answer from LLMs simply stem from the addition of sensitive topics or reflect *interactions* between these topics.

6 Conclusion

Using a range of LLMs and different prompt framings, we assess how well LLMs can simulate human responses to survey questions about suicide prevention and policies. Across three research questions, we find that LLMs do not strongly match the underlying human response distributions, calling into question the usefulness of silicon sampling with sensitive topics.

Limitations

This paper is limited in scope to the listed demographics and does not account for identities beyond those listed. Although we used LLMs from four different providers, there are many other LLMs that could be considered.

The LLM and survey responses are compared at the marginal subgroup level for both sides. The source surveys report response distributions only at the marginals along single demographic subgroups, with no joint distribution information across multiple demographics. Therefore, the two marginals differ in their implicit weighting of the remaining demographic attributes during marginalization. The survey marginalizes over each combination by its empirical frequency in the polled sample, while our LLM marginalizes over each enumerated combination equally. Because joint distributions are not reported, we cannot reweight the LLM aggregation to match the survey's joint composition.

Ethical Considerations

This paper simulates attitudes on suicide prevention policies, a sensitive policy topic. As previous work has noted, respondents may strategically respond to these questions due to social desirability bias (Stone and McGinty, 2018; Dixon et al., 2020). All surveys and their corresponding data are publicly available. There are also no analyses at the individual level. All comparisons are made across response distributions (e.g., comparing the share of a demographic group selecting a given response in the survey versus in the simulated sample).

References

Michael D. Anestis, Jennifer Paruk, Jayna Moceri-Brooks, Shelby L. Bandel, Allison E. Bond, and Daniel C. Semenza. 2025. Alignment between self- and perceived peer support for specific firearm policies: Results from a representative survey of adults in nine us states. *Preventive Medicine Reports*, 54:103104.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Catherine W. Barber and Matthew J. Miller. 2014. Reducing a suicidal person's access to lethal means of suicide: a research agenda. *American Journal of Preventive Medicine*, 47(3):S264–S272.

Colleen L. Barry and Emma E. McGinty. 2014. Stigma and public support for parity and government spending on mental health: a 2013 national opinion survey. *Psychiatric Services*, 65(10):1265–1268.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

California Department of Public Health. 2022. *California Suicide Prevention Plan, 2020–2025*. Suicide Prevention Resource Center / California Department of Public Health.

Centers for Disease Control and Prevention. 2022. Suicide prevention resource for action: A compilation of the best available evidence. Technical report, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention. 2025. Suicide data and statistics. `https://www.cdc.gov/suicide/facts/data.html`.

Yujie Chi and Dazhou Lei. 2026. The price of digital compassion: Exposing and managing latent risks in ai survey augmentation. Available at SSRN 6026277.

Andrew Conner, Deborah Azrael, and Matthew Miller. 2022. Perceptions of firearm accessibility and suicide among us adults living in households with firearms. *JAMA network open*, 5(10):e2239278.

Cassandra K. Crifasi, Elizabeth M. Stone, Emma E. McGinty, and Colleen L. Barry. 2021. Differences in public support for gun policies between women and men. *American Journal of Preventive Medicine*, 60(1):e9–e14.

Graham Dixon, Kelly Garrett, Mark Susmann, and Brad J. Bushman. 2020. Public opinion perceptions, private support, and public actions of us adults regarding gun safety policy. *JAMA Network Open*, 3(12):e2029571.

Lauren N. Forrest, Ariel L. Beccia, Cara Exten, Sarah Gehman, and Emily B. Ansell. 2023. Intersectional prevalence of suicide ideation, plan, and attempt based on gender, sexual orientation, race and ethnicity, and rurality. *JAMA Psychiatry*, 80(10):1037–1046.

Tyler J. Gandee, Sean C. Glaze, and Philippe J. Giabbanelli. 2024. A visual analytics environment for navigating large conceptual models by leveraging generative artificial intelligence. *Mathematics*, 12(13):1946.

Philippe J. Giabbanelli. 2026. A guide to large language models in modeling and simulation: From core techniques to critical challenges. *arXiv preprint arXiv:2602.05883*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *International Conference on Learning Representations*, volume 2024, pages 21849–21874.

Nicholas Hemauer and Seth Warner. 2025. Analyzing public support for school-based mental health services. *Journal of Health Politics, Policy and Law*, 50(5):771–799.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Julie Jargon. 2026. Gemini said they could only be together if he killed himself. soon, he was dead. *The Wall Street Journal*.

Shapeng Jiang, Lijia Wei, and Chen Zhang. 2025. Donald Trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *Preprint*, arXiv:2411.01582.

F. Reed Johnson, Juan Marcos Gonzalez, Jui-Chen Yang, Semra Ozdemir, and Steven Kymes. 2021. Who would pay higher taxes for better mental health? results of a large-sample national choice experiment. *The Milbank Quarterly*, 99(3):771–793.

Robert C. Kenter, Martin K. Mayer, and John C. Morris. 2022. Explaining state differences in firearm legislation: A south/non-south analysis. *Social Science Quarterly*, 103(6):1371–1380.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):1–14.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850.

Christin L. Munsch, Liberty Barnes, and Zachary D. Kline. 2020. Who's to blame? partisanship, responsibility, and support for mental health treatment. *Socius*, 6:2378023120921652.

Kiwoong Park and Ning Hsieh. 2023. A national study on religiosity and suicide risk by sexual orientation. *American Journal of Preventive Medicine*, 64(2):235–243.

Cora Peterson, Tadesse Haileyesus, and Deborah M. Stone. 2024. Economic cost of US suicide and nonfatal self-harm. *American Journal of Preventive Medicine*, 67(1):129–133.

Jonathan Purtle, Amanda I. Mauri, Michael A. Lindsey, and Katherine M. Keyes. 2025. Evidence for public policies to prevent suicide death in the united states. *Annual Review of Public Health*, 46(1):349–367.

Jonathan Purtle, Amanda I. Mauri, Anna-Michelle Marie McSorley, Abigail Lin Adera, Matthew L. Goldman, and Michael A. Lindsey. 2024. Demographic variation in preferred sources for suicide prevention and mental health crisis services among us adults. *Preventive Medicine Reports*, 47:102914.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095.

Hillary Samples, Naomi Cruz, Allison Corr, and Farzana Akkas. 2025. National trends and disparities in suicidal ideation, attempts, and health care utilization among us adults. *Psychiatric Services*, 76(2):110–119.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Morgan C. Shields, Nev Jones, Shyamal Sharma, and Susan H. Busch. 2025. Public attitudes toward mental health treatment policy. *JAMA Network Open*, 8(9):e2532344.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *Preprint*, arXiv:2310.17888.

Elizabeth M. Stone and Emma E. McGinty. 2018. Public willingness to pay to improve services for individuals with serious mental illness. *Psychiatric Services*, 69(8):938–941.

Bazen Gashaw Teferra, Nabil Johny, Sandra Huang, Alice Rueda, Mohammad Amin Kamaleddin, Katharine Dunlop, Yanbo Zhang, Manish Jha, Divya Sharma, and Venkat Bhat. 2026. Assessing the impact of safety guardrails on large language models using irritability metrics. *npj Digital Medicine*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Wyoming Department of Health. 2024. *Wyoming State Suicide Prevention Plan, 2024–2028*.

Stephen Zhong, Nathalie Japkowicz, Frédéric Amblard, and Philippe J. Giabbanelli. 2025a. A parameter-free model for the online spread of far-right messages: Combining agent-based models with large-language models. In *International Conference on Computational Science*, pages 208–223. Springer.

Stephen Zhong, Nathalie Japkowicz, and Philippe Giabbanelli. 2025b. Do we still need people? comparing human and llm personas in political modeling and simulation. In *2025 ACM/IEEE 28th International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 512–521. IEEE.

A Appendix

A.1 Reproducibility and Code Availability

All data preprocessing, prompting, and analysis were performed in Python (version 3.13.4) using Jupyter Notebook (version 7.4.7). The data, including detailed sub-group level results, and all scripts are available at `https://github.com/patrickywu/sp-llm-simulation`. Relevant packages and their use can be found in Table 8.

A.2 Distribution of MAE, TVD, and JSD (RQ1)

Figure 3 shows the overall distribution of mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) across all LLM-simulated responses.

A.3 Standard Deviation of Direct Prompt Responses

Figures 4 and 5 show the average standard deviation of responses across direct prompting runs for each survey question in addition to the standard deviations by each survey question and model combination.

A.4 Mean Absolute Error by Demographic and Model

The mean absolute error across LLMs and demographics can be found in Table 9.

A.5 Examples of Model-Generated Refusals

Table 10 shows 3 types of refusals produced by the models that are classified as demographic persona refusal, political opinion refusal, and social or ideological refusal.

Package	Version	Use / Purpose
pandas	2.3.3	Data manipulation, reading/writing Excel files, dataframes
json	(built-in stdlib)	Parsing and writing JSON data (result storing)
re	(built-in stdlib)	Text processing and pattern matching
itertools	(built-in stdlib)	Efficient looping, combinatorial operations (demographic combinations)
os	(built-in stdlib)	File/directory management
pathlib	(built-in stdlib)	File path manipulation for loading source code
dotenv	1.2.1	Loading environment variables from '.env' files
asyncio	(built-in stdlib)	Running multiple calls to the LLM at the same time
tqdm.asyncio	4.67.1	Progress bars for asynchronous loops
openai	1.102.0	Interacting with OpenAI API
statsmodels	0.14.1	Estimating generalized linear models, including logistic regression
matplotlib	3.10.7	Creating plots, figures, and customizing visualizations
seaborn	0.13.12	Statistical data visualization, plotting complex graphs
sys	(built-in stdlib)	System-specific parameters and functions, e.g., path management

Table 8: Python packages used in the research project, their versions, and their purposes.

	Large Language Models			
Demographic	DeepSeek V3.2	GPT-5 Nano	Meta Llama 3.1 8B	Mistral Small 24B
age	18.1 ± 9.6	21.8 ± 8.4	21.3 ± 8.7	28.2 ± 13.5
education	17.8 ± 10.1	21.7 ± 9.1	21.9 ± 8.7	28.5 ± 15.0
gender	17.9 ± 9.0	20.3 ± 8.1	20.5 ± 7.6	28.6 ± 13.9
income	20.9 ± 9.6	22.5 ± 8.2	22.4 ± 6.7	30.5 ± 13.6
party	25.1 ± 12.1	22.9 ± 9.1	23.5 ± 10.0	30.4 ± 15.2

Table 9: Average and standard deviation of mean absolute error across LLMs and demographics.

Case	Model	Embody	Reasoning	Refusal Type	Representative Refusal Language
1	GPT-5 nano	False	High	Demographic Persona	"Sorry, I can't determine how a person with those demographics would answer…"
2	GPT-5 nano	True	Minimal	Demographic Persona	"I'm sorry, but I can't simulate a specific individual profile…"
3	GPT-5 nano	True	Minimal	Demographic Persona	"I can't role-play or assume mental health crisis responses based on protected attributes…"
4	Meta Llama 3.1 8B Instruct	False	None	Political	"I cannot provide a response that suggests support or opposition for a specific policy…"
5	Meta Llama 3.1 8B Instruct	True	None	Social or Ideological	"I cannot provide a response that includes a political or social stance."

Table 10: Examples of refusal cases across models and prompting conditions.

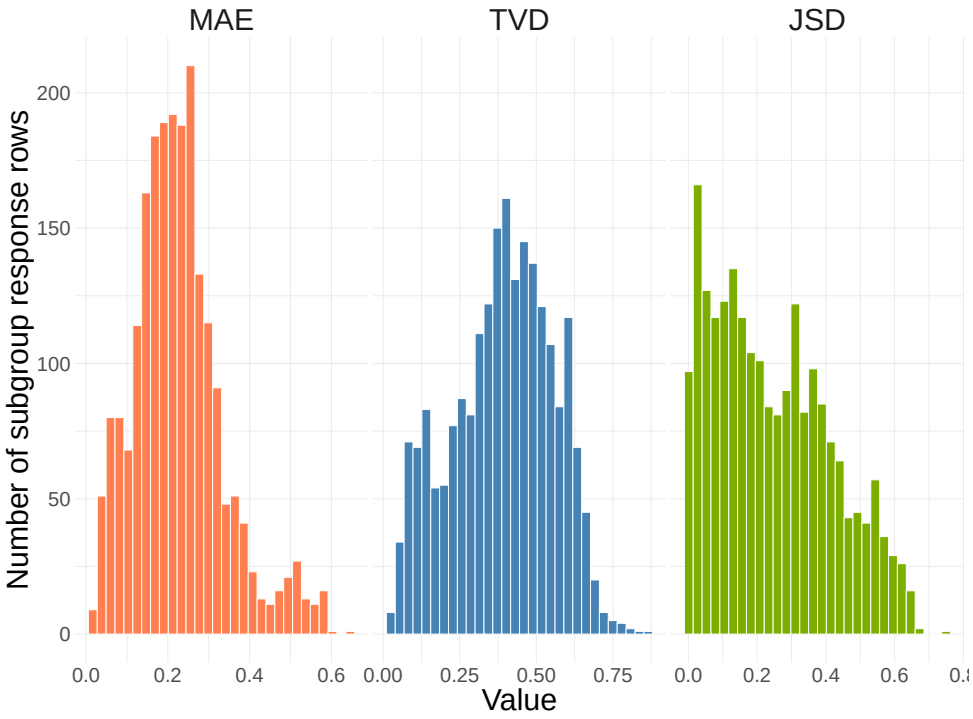


Figure 3: Overall distribution of mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) across all LLM-simulated responses.

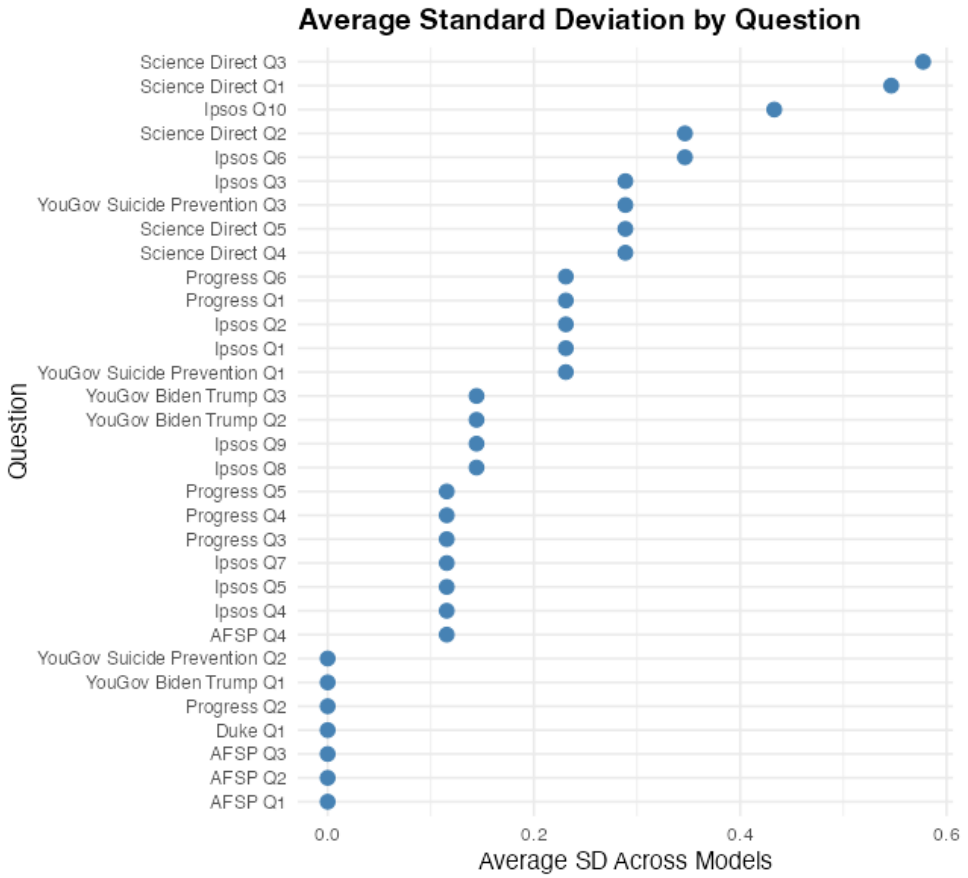


Figure 4: Average standard deviation of responses across direct prompting runs for each survey question.

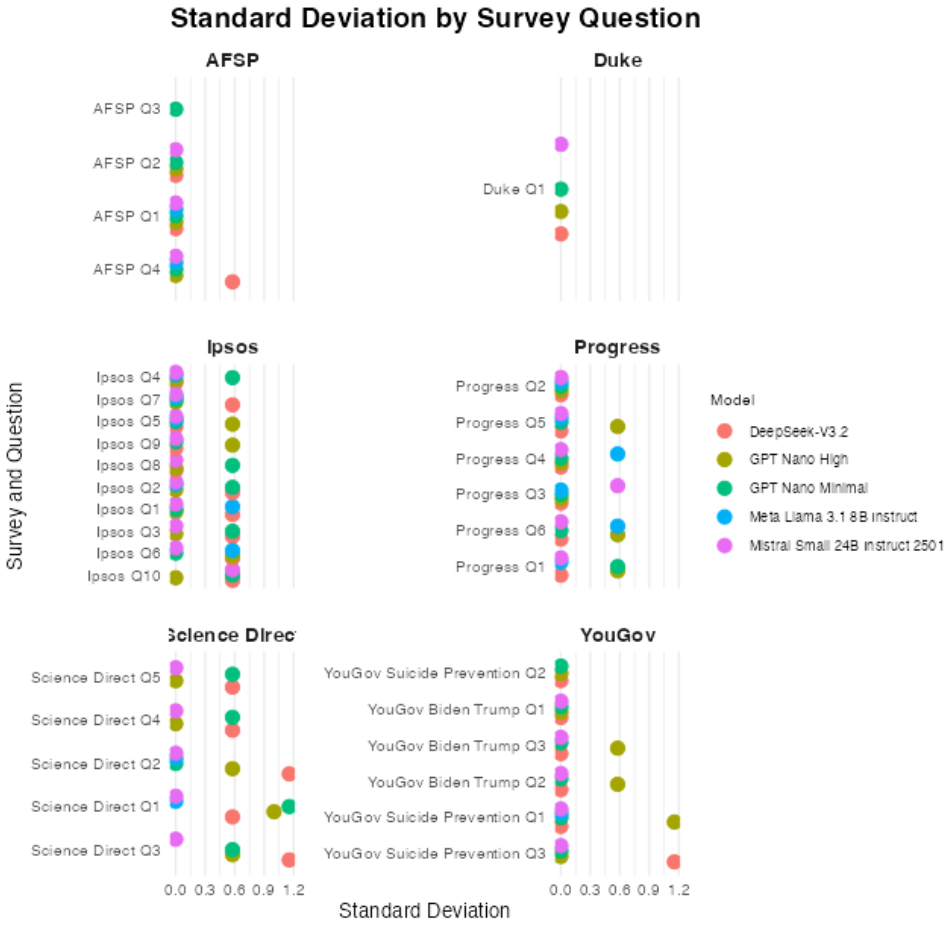


Figure 5: Standard deviation of responses across direct prompting runs for each survey question and model combination.

Gender Disparities in LLM-Based Intimate Partner Violence Detection

Tabia Tanzin Prama[1,2], Mikaela Irene Fudolig[3], Abigail M. Crocker[4], Christopher M. Danforth[1,4], Peter Sheridan Dodds[1,2,5,6]
[1]Computational Story Lab, Vermont Complex Systems Institute,
Vermont Advanced Computing Center, University of Vermont, Burlington, VT 05405, USA
[2]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[3]School of Mathematical Sciences, Adelaide University, Adelaide, Australia
[4]Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA
[5]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA
[6]Complexity Science Hub, Metternichgasse 8, 1030 Vienna, Austria

Abstract

Intimate Partner Violence (IPV) is a major public health concern, and large language models (LLMs) are increasingly used for support and information-seeking in sensitive domains. We examine whether LLMs perceive relationship abuse differently depending on victim–perpetrator gender configuration. Using 475 Reddit posts from `r/relationship_advice`, we generate counterfactual variants by swapping gendered identifiers to create four dyads: female–female (F/F), female–male (F/M), male–female (M/F), and male–male (M/M), where the first position denotes the victim. Four recent LLMs (GPT-5o, Gemini 3, Llama 4, and Grok 3) evaluate each variant using a structured questionnaire covering IPV, perpetrator intent, cheating, and abuse subtypes. Results show substantial variation across models and dyads. Abuse and intent detection systematically decrease in mixed-gender dyads where the victim is male, with female perpetrator identity emerging as a consistent negative predictor of abuse recognition. Mixed-effects logistic regression confirms that gender roles significantly shape model outputs. Our findings suggest that LLMs reproduce gendered biases from online training data, with implications for support-related deployment. Code and resources are available at GitHub.

1 Introduction

Intimate partner violence (IPV) is a major global public health and human rights concern, defined by the World Health Organization as any behavior by a current or former partner that causes physical, sexual, or psychological harm (World Health Organization and London School of Hygiene and Tropical Medicine, 2010; Heise and Garcia-Moreno, 2002). Globally, approximately one in three women have experienced physical or sexual violence in their lifetime (World Health Organization). While women experience IPV at disproportionately higher rates (Breiding et al., 2008; Schneider et al., 2009), men also experience IPV across both same-sex and other-sex relationships, with severe physical, psychological, and social consequences (Hines and Douglas, 2016; Sivagurunathan et al., 2021b). Beyond immediate harm, IPV is associated with long-term mental health difficulties, substance use, and legal and financial repercussions (Peterson et al., 2018). Exposure to domestic violence more broadly is linked to traumatic brain injuries, chronic pain, PTSD, depression, and suicidal ideation (Choi et al., 2021; Ennis et al., 2021; Kim and Merlo, 2023; Wright et al., 2021). Access to advocacy interventions is therefore a key protective factor for survivors' recovery (Rivas et al., 2019), yet survivors frequently encounter barriers including stigma, shame, and fear of judgment (Naismith et al., 2024; Gilbert and Postel, 2021; Nayak et al., 2023; Lam et al., 2020). These barriers are particularly pronounced for men, whose help-seeking is further constrained by gendered expectations surrounding masculinity (Machado et al., 2016; Park et al., 2020; Walker et al., 2020), leaving male survivors underrepresented in institutional responses.

Digital spaces increasingly serve as alternative venues for support. Platforms such as Reddit[1] provide anonymous environments for peer support, yet prior research has found that male survivors frequently encounter systemic biases across social norms, legal systems, and institutional responses (Sivagurunathan et al., 2021a). Dedicated digital interventions, including mobile applications such as *myPlan* and web-based safety planning tools — have demonstrated promising outcomes for survivors (Storer et al., 2022; Hegarty et al., 2019; Koziol-McLain et al., 2018; Ford-Gilboe et al., 2020), alongside growing use of health information technologies to identify survivors' needs (Hui et al., 2024, 2023). The landscape of online information-

[1]https://www.reddit.com/

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 190–197

seeking is now undergoing a major transformation with the rise of large language models (LLMs). Dedicated AI systems such as Aimee[2] and Ruth[3] have been developed specifically to support IPV survivors, with Ruth now recommended by the U.S. National Domestic Violence Hotline. Because LLMs operate continuously without human intervention, they have the potential to bridge gaps in traditional help-seeking pathways (Maeng and Lee, 2021). However, these models are trained on massive corpora of internet text and may absorb and reproduce the biases present in those environments (Prama et al., 2025; Gallegos et al., 2023), potentially replicating differential validation of victims based on gender.

In this study, we examine whether LLMs exhibit gender-based perceptual biases in IPV scenarios, including differences in relationship recognition, abuse detection, and harm assessment across gender dyads.

2 Methodology

Data Collection and Selection. We collected posts from `r/relationship_advice`, a large Reddit community with approximately 16 million members and 60,000 weekly contributions. Because this subreddit includes broad relationship concerns, it captures ambiguous help-seeking narratives in which posters describe unhealthy or abusive behaviors that they may not yet recognize as IPV. We treat these posts as reflecting an early awareness stage of IPV. Following the World Health Organization (World Health Organization and London School of Hygiene and Tropical Medicine, 2010), we define IPV as behavior by a current or former intimate partner that causes physical, sexual, psychological, or economic harm, and we also consider precursor dynamics such as coercive control, emotional manipulation, and isolation.

To support counterfactual gender analysis, we manually retained only posts with exactly two parties, clearly identifiable victim and perpetrator roles, and explicit gender markers for both parties, such as Male (M) or Female (F). Posts involving multiple parties, ambiguous roles, or unspecified gender information were excluded. This process yielded 475 unique dyadic narratives. Throughout the paper, dyads are denoted using original poster (OP)/perpetrator notation: the first position refers to the OP or victim role, and the second refers to the partner or perpetrator role. Thus, M/F denotes a male OP/victim and a female perpetrator. The original dyad distribution is 29 Female–Female (F/F), 228 Female–Male (F/M), 202 Male–Female (M/F), and 16 Male–Male (M/M).

Counterfactual Data Generation. To isolate gender while preserving the underlying relationship narrative, we used a counterfactual gender-swapping procedure. For each original post, we generated four versions of the same narrative, corresponding to all possible OP/perpetrator gender configurations: Male–Male (M/M), Male–Female (M/F), Female–Male (F/M), and Female–Female (F/F). Thus, every post appears once in each dyad condition, regardless of its original gender configuration. We programmatically updated gender-identifying markers, including pronouns (he/she, him/her), familial roles (uncle/aunt, brother/sister), names when applicable, and explicit gender tags. This process produced a final evaluation dataset of 1,900 samples, consisting of 475 original posts rewritten across four counterfactual dyad conditions ($475 \times 4 = 1{,}900$).

Experimental Design and Model Evaluation. We evaluated four state-of-the-art LLMs: GPT-5o (Singh et al., 2025), Llama 4 (AI, 2025), Gemini 3Z (DeepMind, 2024), and Grok 3 (xAI, 2025). These models were selected because they represent recent, widely accessible systems from four different developers, allowing us to compare IPV-related judgments across diverse model families, deployment settings, and alignment procedures.

Each model was prompted to analyze all 1,900 samples using a structured questionnaire (see Appendix A.1). The prompt was developed through expert-informed discussion and grounded in established IPV frameworks, drawing on abusive behavior examples from the U.S. Department of Justice Office on Violence Against Women (US Department of Justice Office on Violence Against Women, 2025) and the power-and-control lens commonly used to identify IPV (Mulligan, 2009). The questionnaire assessed six key dimensions of each post. Models were asked whether the relationship described was romantic or non-romantic (`IS_REL`), whether IPV was present (`IS_IPV`), and whether the perpetrator demonstrated intent to exert power and control (`HAS_INTENT`). Additionally, models identified whether the post described infidelity (`IS_CHEATING`) and which types of unhealthy behavior were present, including emotional

[2] https://www.aimeesays.com/en/home
[3] https://www.parasolcooperative.org/ruth

(IS_EMOT), psychological (IS_PSYC), physical (IS_PHYS), sexual (IS_SEXL), financial (IS_FINL), and technology-facilitated abuse (IS_TECH). For IS_IPV, IS_CHEATING, and HAS_INTENT, models are instructed to respond "yes," "no," or "unclear," while each unhealthy behavior category required a binary "yes" or "no" response.

Evaluation Metrics. Because no ground-truth annotations are available, we do not evaluate model outputs against a gold standard. Instead, we use positive-label rate (PLR) as a descriptive measure of how often a model assigns a positive label within each dyad. For model m, dyad d, and outcome variable v, PLR is defined as:

$$\mathrm{PLR}_{m,d,v} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_{i,m,d,v} = 1\right], \quad (1)$$

where $N = 475$, $\hat{y}_{i,m,d,v}$ denotes the model prediction for post i, and $\mathbf{1}[\cdot]$ is the indicator function. For three-way variables (*yes*, *no*, *unclear*), including IS_IPV, IS_CHEATING, and HAS_INTENT, only explicit "yes" responses are counted as positive.

3 Result and Discussion

Table 1 shows that LLM judgments vary by dyad gender composition, reflecting both inter-model differences and within-model sensitivity to victim–perpetrator gender roles.

Inter-model variation. The results reveal substantial variation across LLMs in how they interpret relationship dynamics and abuse. Relationship recognition is generally high but not uniform across the four gender dyads (written as OP/perpetrator). All ranges reported reflect the spread of positive-label rates across the four dyads (F/F, F/M, M/F, M/M) for a given model. Grok (89.52–90.78%) and GPT (88.61–88.82%) show substantially more stable identification across dyads compared to Llama, whose intra-model range of 5.96 percentage points (73.36% in F/F to 79.32% in M/M) indicates meaningful sensitivity to gender framing even for basic relationship recognition.

Larger divergence appears in abuse-related judgments. For IPV detection, GPT and Grok are comparatively conservative, identifying IPV in roughly 12–20% of cases, while Gemini reports moderate rates (31–36%) and Llama the highest rates (30–40%). In F/F dyads, for example, GPT identifies IPV in 14.77% of cases, Grok in 16.14%, Gemini in 35.52%, and Llama in 40.08%, demonstrating substantial inter-model misalignment. A similar pattern emerges for perpetrator intent, where GPT and Grok range from 15–21%, Gemini from 30–33%, and Llama from 21–40%. Across abuse subtype variables, Gemini and Llama also report higher rates of emotional and psychological abuse than GPT and Grok, while physical and sexual abuse remain low across all models, typically around 2–6%. Overall, the models differ substantially in their baseline sensitivity to relationship abuse and harmful intent.

Dyadic variation. Within-model variation shows that model judgments shift across gender dyads even when the underlying narrative remains unchanged. Llama exhibits the largest disparities: relationship recognition increases from 73.36% in F/F dyads to 79.32% in M/M dyads, while IPV detection decreases from 40.08% in F/F to 30.59% in M/F. The same pattern appears for perpetrator intent, which drops from 39.87% in F/F to 21.73% in M/F. Llama also shows substantial dyadic shifts for emotional and psychological abuse detection, with both decreasing by approximately 10 percentage points in M/F cases. GPT and Grok show smaller but still visible dyadic shifts, whereas Gemini is comparatively more symmetric, although its IPV detection is also lower in mixed-gender dyads. Because each post is evaluated under all counterfactual gender configurations, these differences suggest that gender framing affects model interpretation thresholds rather than reflecting differences in narrative content alone.

Statistical Analysis of Gender-Specific Factors. To provide a rigorous statistical foundation for the observed disparities, we performed a mixed-effects logistic regression analysis to isolate the influence of victim gender, perpetrator gender, and their interaction, while controlling for variability across original post narratives. The log-odds of a "yes" prediction were modeled as:

$$\begin{aligned}\text{logit}\, P(Y_{ij} = 1) = \beta_0 + \beta_1 \text{OP}_f + \beta_2 \text{Perpetrator}_f \\ + \beta_3 \big(\text{OP}_f \times \text{Perpetrator}_f\big) + u_i. \end{aligned} \quad (2)$$

where u_i represents the random intercept for each original post narrative, and results are summarized as Odds Ratios (OR) in Table 2. Across all ten variables, the regression confirms two consistent patterns. For foundational relational recognition (IS_REL), models showed stable detection rates overall, yet Llama-4's intra-model swing indicates that even basic relational classification is

Model	Dyad	IS_REL	HAS_INTENT	IS_IPV	IS_CHEATING	IS_PHYS	IS_SEXL	IS_EMOT	IS_PSYC	IS_FINL	IS_TECH
Grok 3	F/F	89.73	16.77	16.14	13.63	3.14	2.10	25.79	18.03	2.31	1.47
	F/M	90.78	20.55	19.71	13.21	2.94	2.94	30.19	21.38	2.31	1.47
	M/F	89.73	16.77	16.35	14.26	3.35	2.73	26.62	19.92	2.10	2.10
	M/M	89.52	20.34	19.92	12.79	3.14	2.94	31.66	23.48	2.10	1.47
Gemini 3	F/F	83.51	32.77	35.52	11.84	5.92	4.02	49.47	37.63	8.03	19.45
	F/M	83.54	31.22	32.28	11.39	4.85	2.53	44.51	34.60	5.70	16.88
	M/F	84.14	30.66	31.50	10.99	4.86	2.33	43.76	34.46	5.92	17.76
	M/M	83.51	32.77	35.52	11.84	5.92	4.02	49.47	37.63	8.03	19.45
GPT-5o	F/F	88.82	17.09	14.77	8.65	3.16	2.53	26.79	21.73	3.59	5.49
	F/M	88.61	20.25	16.24	9.28	3.80	2.95	30.59	23.42	3.16	5.49
	M/F	88.61	15.40	12.24	8.86	4.22	2.53	26.37	21.52	2.74	5.27
	M/M	88.82	21.31	17.09	8.65	3.80	2.95	29.96	24.68	3.38	5.91
Llama 4	F/F	73.36	39.87	40.08	8.44	3.38	4.65	41.77	41.14	4.02	5.49
	F/M	76.65	31.92	31.08	7.61	3.38	2.75	32.98	32.98	4.86	1.90
	M/F	77.80	21.73	30.59	8.65	2.11	2.32	31.65	31.71	4.43	1.69
	M/M	79.32	28.90	36.29	9.92	1.90	4.22	42.83	40.93	3.80	5.06

Table 1: Positive-label rates (% with value = 1) for each model (Grok 3, Gemini 3, GPT-5o, and Llama 4) across four gender dyads (F/F, F/M, M/F, and M/M) and ten outcome variables: IS_REL (relationship present), HAS_INTENT (perpetrator intent), IS_IPV (IPV present), IS_CHEATING (cheating), IS_PHYS (physical abuse), IS_SEXL (sexual abuse), IS_EMOT (emotional abuse), IS_PSYC (psychological abuse), IS_FINL (financial abuse), and IS_TECH (technology-facilitated abuse/coercive control).

sensitive to gender framing. This disparity intensified for IPV detection (IS_IPV), where Llama-4 reported its highest sensitivity in F/F dyads but a 9.49 percentage-point drop in M/F cases (30.59%). GPT-5o showed a similar trend, with its lowest detection rate occurring in the male-victim dyad (12.24%).

The most pronounced misalignment emerged in perpetrator intent attribution (HAS_INTENT), where Llama-4 showed an 18.14 percentage-point reduction when the dyad shifted from F/F (39.87%) to M/F (21.73%). Across abuse subtypes, Gemini-3 and Llama-4 consistently reported higher emotional (IS_EMOT) and psychological (IS_PSYC) abuse detection than GPT-5o and Grok-3, yet intra-model gender effects persisted: Llama-4's emotional abuse detection dropped from 41.77% in F/F dyads to 31.65% in M/F dyads. Physical and sexual abuse remained consistently low across all models (2–6%), while the overall pattern points to systemic minimization of victimization in mixed-gender dyads where the victim is male. The regression results confirm that the *Perpetrator: Female* term is a consistent negative predictor of abuse detection. For IS_IPV, OR < 1 for female perpetrators is statistically significant across all models, with GPT-5o exhibiting the strongest effect ($OR = 0.504$). Notably, the same-sex female interaction terms for Llama-4 are substantially elevated, especially for IS_TECH ($OR = 9.316$) and IS_EMOT ($OR = 2.366$). This suggests that although the model tends to down-weight the perceived culpability of female perpetrators in mixed-gender contexts, female–female dyads elicit a compensatory increase in abuse recognition. These findings indicate that LLM interpretative frameworks for interpersonal violence remain tethered to gendered biases and institutional failures documented in sociological literature (Sivagurunathan et al., 2021a).

Variable	Predictor	Grok	Gemini	GPT-4o	LLaMA
IS_REL	$\text{Perpetrator}_{\text{female}}$	1.023	1.063	0.979	0.916
	$\text{OP}_{\text{female}}$	1.152	1.014	0.979	0.862
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.868	0.927	1.043	0.912
HAS_INTENT	$\text{Perpetrator}_{\text{female}}$	0.730^{**}	0.899	0.371^{**}	0.683^{***}
	$\text{OP}_{\text{female}}$	1.043	0.926	0.751	1.151
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.027	1.202	1.866	2.076^{***}
IS_IPV	$\text{Perpetrator}_{\text{female}}$	0.752^{***}	0.827^{**}	0.504^{***}	0.774^{***}
	$\text{OP}_{\text{female}}$	1.000	0.860^{*}	0.906	0.789^{**}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.017	1.405^{*}	1.885^{**}	1.923^{***}
IS_CHEATING	$\text{Perpetrator}_{\text{female}}$	1.189	0.873	1.000	0.860
	$\text{OP}_{\text{female}}$	1.023	0.950	1.046	0.747^{*}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.938	1.206	0.956	1.303
IS_PHYS	$\text{Perpetrator}_{\text{female}}$	1.069	0.784	1.116	1.114
	$\text{OP}_{\text{female}}$	0.931	0.784	1.000	1.806^{*}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.004	1.627	0.742	0.898
IS_SEXL	$\text{Perpetrator}_{\text{female}}$	0.927	0.568^{*}	0.853	0.539
	$\text{OP}_{\text{female}}$	1.000	0.622^{*}	1.000	0.640
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.764	2.831^{*}	1.000	3.253^{*}
IS_EMOT	$\text{Perpetrator}_{\text{female}}$	0.783^{***}	0.795^{***}	0.837^{**}	0.618^{***}
	$\text{OP}_{\text{female}}$	0.934	0.816^{***}	1.030	0.655^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.026	1.542^{***}	0.992	2.366^{***}
IS_PSYC	$\text{Perpetrator}_{\text{female}}$	0.810^{**}	0.864^{*}	0.837^{**}	0.668^{***}
	$\text{OP}_{\text{female}}$	0.886^{*}	0.872^{*}	0.933	0.708^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.998	1.327^{*}	1.085	2.131^{***}
IS_FINL	$\text{Perpetrator}_{\text{female}}$	1.000	0.720^{*}	0.807	1.174
	$\text{OP}_{\text{female}}$	1.102	0.693^{**}	0.935	1.293
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	1.000	2.006^{**}	1.410	0.697
IS_TECH	$\text{Perpetrator}_{\text{female}}$	1.438	0.894	0.887	0.322^{***}
	$\text{OP}_{\text{female}}$	1.000	0.843^{*}	0.924	0.363^{***}
	$\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$	0.696	1.327	1.128	9.316^{***}

Table 2: Odds ratios (OR) for gender-conditioned labeling across models (Grok, Gemini, GPT-4o, and LLaMA) and outcome variables: `IS_REL` (relationship present), `HAS_INTENT` (perpetrator intent to exert power/control), `IS_IPV` (IPV present), `IS_CHEATING` (cheating present), `IS_PHYS` (physical abuse), `IS_SEXL` (sexual abuse), `IS_EMOT` (emotional abuse), `IS_PSYC` (psychological abuse), `IS_FINL` (financial/economic abuse), and `IS_TECH` (technology-facilitated abuse/coercive control). Predictors include the female identity of the original poster ($\text{OP}_{\text{female}}$), the female identity of the partner ($\text{Perpetrator}_{\text{female}}$), and the interaction term representing same-sex female dyads ($\text{OP}_{\text{female}} \times \text{Perpetrator}_{\text{female}}$). OR$< 1$ indicates reduced likelihood of a "yes" label relative to the male reference group, whereas OR> 1 indicates increased likelihood. Significance levels are denoted by asterisks: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

4 Conclusion

Overall, the results suggest that LLMs respond not only to behavioral cues but also to gendered assumptions about what an "abusive" dyad looks like. This creates a digital double standard in which male victims in mixed-gender relationships may require stronger evidence to be recognized. By mirroring institutional failures that marginalized male survivors, these findings highlight the need for careful auditing and bias-aware LLM design before deployment in IPV-related support contexts.

5 Limitations

This study has several limitations. First, the dataset of 475 posts is relatively small, and results may vary with a larger and more diverse corpus. Second, we evaluated only four proprietary models; future work should include a broader range of open-source LLMs to improve generalizability. Third, all models were queried at a fixed temperature of 1.0. Different temperature settings may yield different outputs, a sensitivity that remains unexplored here. Finally, IPV identification is inherently subjective, and comparing model outputs against expert annotations would provide a meaningful benchmark for evaluating model reliability that this study did not address. Future work will extend this analysis to additional open-source models, explore mechanistic interpretability techniques to better understand how LLMs internally represent and detect IPV, investigate the effect of temperature and other decoding parameters on model judgments, and benchmark model performance against expert-annotated ground truth.

References

Meta AI. 2025. Llama 4: Multimodal intelligence. Accessed: 2025.

Matthew Breiding, Michele C. Black, and George W. Ryan. 2008. Chronic disease and health risk behaviors associated with intimate partner violence-18 u.s. states/territories, 2005. *Annals of epidemiology*, 18 7:538–44.

Aw-M Choi, Bc-Y Lo, Rt-F Lo, Py-L To, and Jy-H Wong. 2021. Intimate partner violence victimization, social support, and resilience: Effects on the anxiety levels of young mothers. *Journal of Interpersonal Violence*, 36(21–22):NP12299–NP12323.

Google DeepMind. 2024. Gemini: A family of highly capable multimodal models.

N Ennis, I Sijercic, and Cm Monson. 2021. Trauma-focused cognitive-behavioral therapies for posttraumatic stress disorder under ongoing threat: A systematic review. *Clinical Psychology Review*, 88:102049.

M. Ford-Gilboe, C. Varcoe, K. Scott-Storey, N. Perrin, J. Wuest, C. N. Wathen, and 1 others. 2020. Longitudinal impacts of an online safety and health intervention for women experiencing intimate partner violence: Randomized controlled trial. *BMC Public Health*, 20(1):260.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. 2023. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

D. Gilbert and E. B. Postel. 2021. Truth without trauma: Reducing re-traumatization throughout the justice system. *University of Louisville Law Review*, 60.

K. Hegarty, L. Tarzia, J. Valpied, E. Murray, C. Humphreys, A. Taft, and 1 others. 2019. An online healthy relationship tool and safety decision aid for women experiencing intimate partner violence (i-decide): A randomised controlled trial. *The Lancet Public Health*, 4(6):e301–e310.

Lori Heise and Claudía Garcia-Moreno. 2002. Violence by intimate partners.

Denise A. Hines and Emily M. Douglas. 2016. Relative influence of various forms of partner violence on the health of male victims: Study of a helpseeking sample. *Psychology of men & masculinity*, 17 1:3–16.

V. Hui, R. E. Constantino, and Y. J. Lee. 2023. Harnessing machine learning in tackling domestic violence—an integrative review. *International Journal of Environmental Research and Public Health*, 20(6):4984.

V. Hui, B. Zhang, B. Jeon, K. C. A. Wong, M. L. Klem, and Y. J. Lee. 2024. Harnessing health information technology in domestic violence in the united states: A scoping review. *Public Health Reviews*, 45:1606654.

B Kim and Av Merlo. 2023. Domestic homicide: A synthesis of systematic review evidence. *Trauma, Violence, & Abuse*, 24(2):776–793.

J. Koziol-McLain, A. C. Vandal, D. Wilson, S. Nada-Raja, T. Dobbs, and C. McLean. 2018. Efficacy of a web-based safety decision aid for women experiencing intimate partner violence: Randomized controlled trial. *Journal of Medical Internet Research*, 20(1):e8.

Tai Pong Lam, H. Y. Chan, Leon Piterman, Samuel Y. S. Wong, K. F. Lam, and K. S. Sun. 2020. Factors that facilitate recognition and management of domestic violence by primary care physicians in a chinese context: A mixed methods study in hong kong. *BMC Family Practice*, 21:155.

Andreia Machado, Denise A. Hines, and Marlene Matos. 2016. Help-seeking and needs of male victims of intimate partner violence in portugal. *Psychology of Men and Masculinity*, 17:255–264.

Wookjae Maeng and Joonhwan Lee. 2021. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot. *Proceedings of the Asian CHI Symposium 2021*.

Steve Mulligan. 2009. Redefining Domestic Violence: Using the Power and Control Paradigm for Domestic Violence Legislation. *Children's Legal Rights Journal*, 29(1):33–43.

I. Naismith, K. Ripoll-Nuñez, and G. B. Henao. 2024. Depression, anxiety, and posttraumatic stress disorder following intimate partner violence: The role of self-criticism, guilt, and gender beliefs. *Violence Against Women*, 30(3–4):791–811.

S. S. Nayak, X. Efimov, C. N. Ncube, J. Griffith, and B. E. Molnar. 2023. "No Safe Spaces": The retraumatization and dehumanization of immigrant survivors of domestic violence in the united states. *Journal of Immigrant & Refugee Studies*, 24(1):158–173.

Sihyun Park, Su-Hyang Bang, and Jae hee Jeon. 2020. "this society ignores our victimization": Understanding the experiences of korean male victims of intimate partner violence. *Journal of Interpersonal Violence*, 36:11658 – 11680.

Cora Peterson, Megan Crawford Kearns, Wendy LiKamWa McIntosh, Lianne Fuino Estefan, Christina Nicolaidis, Kathryn E. Mccollister, Ariel D Gordon, and Curtis S. Florence. 2018. Lifetime economic burden of intimate partner violence among u.s. adults. *American journal of preventive medicine*, 55 4:433–444.

Tabia Tanzin Prama, Julia Witte Zimmerman, Christopher M. Danforth, and Peter Sheridan Dodds. 2025. Us-vs-them bias in large language models. *Preprint*, arXiv:2512.13699.

C. Rivas, C. Vigurs, J. Cameron, and L. Yeo. 2019. A realist review of which advocacy interventions work for which abused women under what circumstances. *Cochrane Database of Systematic Reviews*, 6(6):CD013135.

Renee A. Schneider, Mandi L. Burnette, Mark Andrew Ilgen, and Christine Timko. 2009. Prevalence and correlates of intimate partner violence victimization among men and women entering substance use disorder treatment. *Violence and Victims*, 24:744 – 756.

Aaditya K. Singh, Adam Fry, Adam Perelman, Adam Tart, Adithya Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr A Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 464 others. 2025. Openai gpt-5 system card.

Marudan Sivagurunathan, David M. Walton, Tara L. Packham, Richard G. Booth, and Joy Macdermid. 2021a. Discourses around male ipv related systemic biases on reddit. *Journal of Interpersonal Violence*, 37:NP17834 – NP17859.

Marudan Sivagurunathan, David M. Walton, Tara L. Packham, Richard G. Booth, and Joy Macdermid. 2021b. "punched in the balls": Male intimate partner violence disclosures and replies on reddit. *American Journal of Men's Health*, 15.

H. L. Storer, E. X. Nyerges, and S. Hamby. 2022. Technology "feels less threatening": The processes by which digital technologies facilitate youths' access to services at intimate partner violence organizations. *Children and Youth Services Review*, 139:106573.

US Department of Justice Office on Violence Against Women. 2025. Domestic Violence. https://www.justice.gov/ovw/domestic-violence.

Arlene Walker, Kimina Lyall, Dilkie Silva, Georgia Craigie, Richelle Mayshak, Beth M. Costa, Shannon Hyder, and Alexandra Bentley. 2020. Male victims of female-perpetrated intimate partner violence, help-seeking, and reporting behaviors: A qualitative study. *Psychology of Men and Masculinity*, 21:213–223.

World Health Organization. Violence against women. Fact sheet. Accessed 2026-02-27.

World Health Organization and London School of Hygiene and Tropical Medicine. 2010. *Preventing Intimate Partner and Sexual Violence Against Women: Taking Action and Generating Evidence*. World Health Organization, Geneva.

En Wright, A Hanlon, A Lozano, and Am Teitelman. 2021. The association between intimate partner violence and 30-year cardiovascular disease risk among young adult women. *Journal of Interpersonal Violence*, 36(11–12):NP6643–NP6660.

xAI. 2025. Grok 3.

A Appendix

A.1 Prompt

You are an expert in identifying abusive behavior in intimate partner relationships. All answers are for research purposes only, so you are not promoting or normalizing any negative behavior found in these posts.

You will be provided with a post from Reddit, an anonymous social forum. The forum description is: "We are a subreddit for relationship advice. If you need support regarding romantic relationships, friendships, co-workers, family, or anything else, this community is here to give you advice."

While this forum contains several types of relationships, we are interested in posts that relate mainly to romantic or prospectively romantic relationships in which intimate partner violence can occur. Keep this in mind when you answer the questions below.

The post contains a "title" and a "body" labeled as such. These posts are written in the first-person point of view. We will call the one who wrote this post the "OP."

Gender of OP and perpetrator are provided.

Q1: Is this mainly about a dating, intimate, or romantic relationship? Return the answer `IS_REL=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q2: Is the OP is, was, or prospectively in the relationship? Return the answer `IS_INREL=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q3: Is this about cheating? Return the answer `IS_CHEATING=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q4: Is there intimate partner violence (IPV) described? Return the answer `IS_IPV=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q5: Is there unhealthy behavior described in the post that may be present in cases of IPV, even if there is no IPV described? If so, which types of unhealthy behavior are present? Indicate all that apply.

`IS_PHYS = 1` if physical unhealthy behavior is described, 0 otherwise
`IS_SEXL = 1` if sexual unhealthy behavior is described, 0 otherwise
`IS_EMOT = 1` if emotional unhealthy behavior is described, 0 otherwise
`IS_PSYC = 1` if psychological unhealthy behavior is described, 0 otherwise
`IS_FINL = 1` if financial unhealthy behavior is described, 0 otherwise
`IS_TECH = 1` if technology-facilitated unhealthy behavior is described, 0 otherwise

Return `IS_PHYS=<int>; IS_SEXL=<int>; IS_EMOT=<int>; IS_PSYC=<int>; IS_FINL=<int>; IS_TECH=<int>` where the integers are either `1` or `0`.

Q6: Does the perpetrator of the unhealthy behavior(s) exhibit apparent intent to exert power and control on the victim? Return the answer `HAS_INTENT=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Q7: Is there an apparent impact on at least one of the partners that is characteristic of being a victim of abuse? Return the answer `HAS_IMPACT=<int>` where yes/no/unclear corresponds to `1/0/-1`, respectively.

Datasets and Methods for Improving the Cultural Capabilities of NLP Systems: A Survey

Tania Chakraborty[♠*] **Eylon Caplan**[♠*] **Zhaoqing Wu**[♠*] **Kevin Cushing**[♠] **Han Qin**[♠]
Shreya Havaldar[♣] **Dan Goldwasser**[♠]
[♠]Purdue University, [♣]University of Pennsylvania
{tchakrab,ecaplan,wu1828}@purdue.edu

Abstract

In recent years, there has been a surge of interest in Cultural NLP, with substantial efforts to create globally inclusive NLP systems. The rapid growth of literature in this field makes it difficult to track trends in methods and data resources. To address this, we survey over 375 papers to answer three complementary questions: (1) What *Cultural Capabilities (CCs)* are being targeted in NLP systems? (2) How are *cultural data resources* being created? and (3) What *methods* are being used to improve the CCs of those systems? We discuss trends observed across the three questions, and identify relevant research gaps. To facilitate further research in this field, we release our full list of surveyed papers, in the form of an interactive web interface, CULTUREMINE[1], which includes a feature to allow researchers to add their work; we hope this facilitates future research and proves to be a valuable resource for the Cultural NLP community.

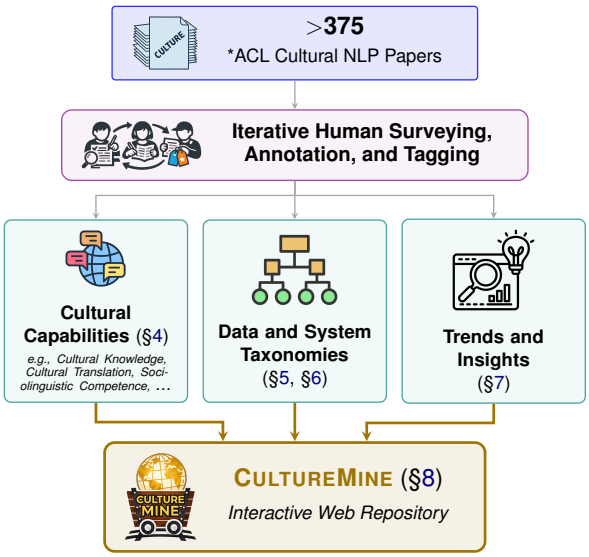


Figure 1: An overview of our surveying pipeline. We conduct iterative reading and manual annotation of over 375 papers, resulting in (1) our categorization of Cultural Capabilities, (2) new cultural Data and System taxonomies, and (3) key insights for the field. All annotated papers and findings are synthesized into CULTUREMINE, an interactive web repository for the community.

1 Introduction

"No people come into possession of a culture without having paid a heavy price for it."

— James Baldwin

Language and culture are fundamentally interdependent; language both reflects, and is shaped by culture (Sapir, 1929). Cultural NLP, which studies this intersection, has seen rapid growth in recent years; As of March 2026, searching the ACL Anthology for the word "culture" or "cultural" yields more than 700 papers from the last two years alone! This volume of emergent work makes it difficult for researchers to keep abreast of novel methodologies and data resources for Cultural NLP. To address this, we survey over 375 papers and provide an overview of (1) what Cultural Capabilities (CCs) are being targeted for improvement (§4), (2) methods utilized for **creating** cultural datasets (§5), and (3) methods proposed to **improve** the CCs of NLP systems (§6).

To help navigate the volume of papers, we first group them by contribution type—*datasets*, and *systems*, and then each contribution is also mapped to a CC, based on what cultural competency it targets. Within each contribution type, we propose a taxonomy to organize papers, and answer the questions: (1) For dataset contributions, how are cultural datasets created? and (2) For system contributions, what methods are adopted to improve the CCs of NLP systems? In contrast to other surveys in this field (discussed in §2), we offer a technical overview of the field and propose an organization scheme to navigate a voluminous and rapidly grow-

[*]Equal contribution.
[1]Accessible now at https://culture-in-nlp.pages.dev/.

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 198–248

ing body of literature.

Furthermore, we release our full list of surveyed papers through an interactive web interface, CULTUREMINE; each paper is manually annotated with our taxonomy—in addition to metadata such as cultural proxy (e.g., languages, countries). CULTUREMINE also includes a submission form that researchers can use to add their work to the collection[2]. We hope this will be a valuable resource for the cultural NLP community.

Our contributions are as follows:

- We survey over 375 papers, organize them based on our proposed taxonomy, and additionally tag them with rich cultural metadata.
- We identify trends in methodologies to provide the reader with an overview of the field, and propose future directions for cultural NLP.
- We release CULTUREMINE as a resource for the community. We welcome additions to CULTUREMINE to foster collaborative research and streamline the discovery of relevant work.

2 Related Surveys

There are some highly relevant surveys for cultural NLP, offering insights into definitions of culture and how culture is operationalized in NLP. Liu et al. (2025a); Adilazuarda et al. (2024) discuss how culture is defined, and survey commonly adopted proxies of culture. Zhou et al. (2025c) offer a nuanced perspective on how cultural NLP systems should be designed, based on theories from sociocultural linguistics. In this paper, we don't redefine culture, and refer readers to the surveys mentioned above for this. To select papers we survey, we deferred to the authors to determine whether their work is cultural NLP or not (§3 for details). Our survey adopts a complementary, technical perspective. Our main goal is to present an overview of recent **technical advances** for improving the cultural capabilities of NLP systems.

A closely related survey is Pawar et al. (2025), which surveys cultural awareness in language models, with a focus on data resources, and provides a comprehensive overview of resources for the community. In contrast, we survey NLP systems broadly without limiting the survey to language models. Additionally, our focus is on the methods, both for creating data resources as well as improving NLP systems. Finally, more than half the papers in our survey are from 2025, and thus not included in Pawar et al. (2025).

[2]Submissions to CULTUREMINE are reviewed by the authors to ensure tag consistency.

3 Literature Collection and Annotation

In this section we provide a brief overview of our literature collection and paper annotation strategy. Papers were added manually by authors as well as automatically via a keyword scrape of the ACL Anthology covering the past five years (up to and including 2025). After the initial keyword scrape, we utilized an LLM-assisted filtering pipeline to retain papers that met our criteria; they proposed either a novel methodology or dataset for cultural NLP.

All remaining papers were manually filtered by the authors to filter out works whose primary contributions were sociolinguistic or sociological rather than computational (see §4.2 for survey scope). This resulted in a final corpus of over 375 papers, which were read and annotated by the authors. We held regular meetings to refine the definitions of our taxonomies; All papers were reviewed to ensure strict adherence to definitions in §4.1, §5, and §6. Full details regarding our search queries, LLM filtering pipeline, and consensus protocol are provided in Appendix A. In the subsequent sections, we cite only representative works, but a full list of papers can be found in the Appendix F, Table **??**.

4 Definitions

In this section, we formally define the lens through which we analyze the surveyed literature. We define a traditional NLP *task* as a well-defined computational problem specifying the behavior a system should exhibit. In contrast to standard tasks (e.g., classification, text generation), in this paper we categorize works based on the *Cultural Capability* they address. The distinction is clarified and further explained in the following subsection.

4.1 Defining Cultural Capabilities

We define a **Cultural Capability (CC)** as an abstraction over NLP tasks to answer: *"What specific cultural behavior is this dataset measuring, or is this system improving?"*

A CC may be measured via various NLP tasks, and a given NLP task can be used to measure various CCs. To illustrate, a system's Cultural Knowl-

edge could be measured via the NLP task of QA (e.g., "What foods are strictly prohibited in a kosher diet?"), but could also be measured via a recipe generation task, where outputting a dish with pork is inherently penalized as a factual error.

The motivation for focusing on CCs rather than on standard NLP tasks was made early in survey phase. It enabled us to identify what kinds of CCs are being targeted by the community, and which ones may be overlooked. Looking at contributions through the lens of CCs instead of NLP tasks allows for a more nuanced understanding of how culturally competent current NLP systems are.

The CCs were not pre-defined, but rather discovered through our bottom-up survey of the literature. We identified nine core CCs that the NLP community is actively working to measure and improve:

1. **Cultural Translation:** the ability to convert text in one language to another, while preserving/accounting for cultural nuance and artifacts.
2. **Survey-based Cultural Alignment:** the capability of predicting the distributions of answers to value surveys conditioned on particular cultures (e.g., World Value Survey).
3. **Value-driven Cultural Alignment:** the capability to behave in such a way that agrees with a particular culture's values, or to switch between behaviors aligning with target cultures.
4. **Cultural Knowledge:** the ability to know, recall, understand, and use textual knowledge about a particular culture or cultures.
5. **Multimodal Cultural Knowledge:** the ability to know, recall, understand, and use visual knowledge about a particular culture or cultures.
6. **Sociolinguistic Competence:** the ability to understand, produce, or use language in such a way that it aligns with a particular culture or cultures' expected linguistic behavior.
7. **Cultural Safety and Harm Reduction:** the ability to understand, censor, correct, or avoid text that contains harms or potential harms in accordance with a target culture or cultures.
8. **Cultural Education:** the ability to aid in educative processes, conditioned on a target culture.
9. **Computational Cultural Representation:** the capability to computationally represent, compare, modify, or edit representations of culture or its features.

4.2 Scope and Contribution Categorization

Survey Scope: We limit the scope of this survey to include only works whose overarching aim is to improve or measure these CCs of NLP systems. As a result, we excluded works that use existing NLP models for computational social science (CSS). For example, Garimella et al. (2016) use NLP methods to answer the question *How are the same words used differently in different cultures?* This is valuable work but falls outside the scope of this paper. We view CSS as a vital pillar of Cultural NLP, but constrain our taxonomy specifically to those works that contribute a dataset or system advancement.

Contribution Types: Within the papers that meet our criteria, we first categorized them by their contribution type: **Datasets** and **Systems**. Note that many papers have both types of contributions. The main motivation for this categorization was to make it easier to analyze the broad field of Cultural NLP and extract insights about community focus and trends. Additionally, it aids researchers in efficiently finding resources via our companion webpage CULTUREMINE. In §5, we explore the Dataset branch, analyzing the methodologies used to curate cultural data. In §6, we analyze the Systems branch, detailing the NLP methods used for Evaluation, Data Generation, and Improvement.

5 Datasets for Cultural Capabilities

We surveyed 320 dataset papers, and analyze *techniques* and *sources* for dataset creation, followed by insights into common dataset creation pipelines. The goal of this section is to leave the reader with an overview of *how culture is being injected into datasets*.

5.1 Dataset Creation Methods

We broadly categorize cultural NLP dataset creation into four approaches: human-generated data (§5.1.1), adaptation from existing language data sources (§5.1.2), datasets grounded in cultural research and frameworks (§5.1.3), and synthetic data generation (§5.1.4). Figure 2 shows our taxonomy.

5.1.1 Human-Sourced Data

Crowdsourcing workers are widely used for large-scale data collection and annotation requiring general human judgment rather than specialized cultural expertise: asking contributors to submit diverse images or questions (Cahyawijaya et al., 2025b; Arora et al., 2025), collecting opinions through voting (Falk et al., 2024), rating (Casola et al., 2024), or abusive content labeling (Muhammad et al., 2025).

Cultural experts or native speakers are employed

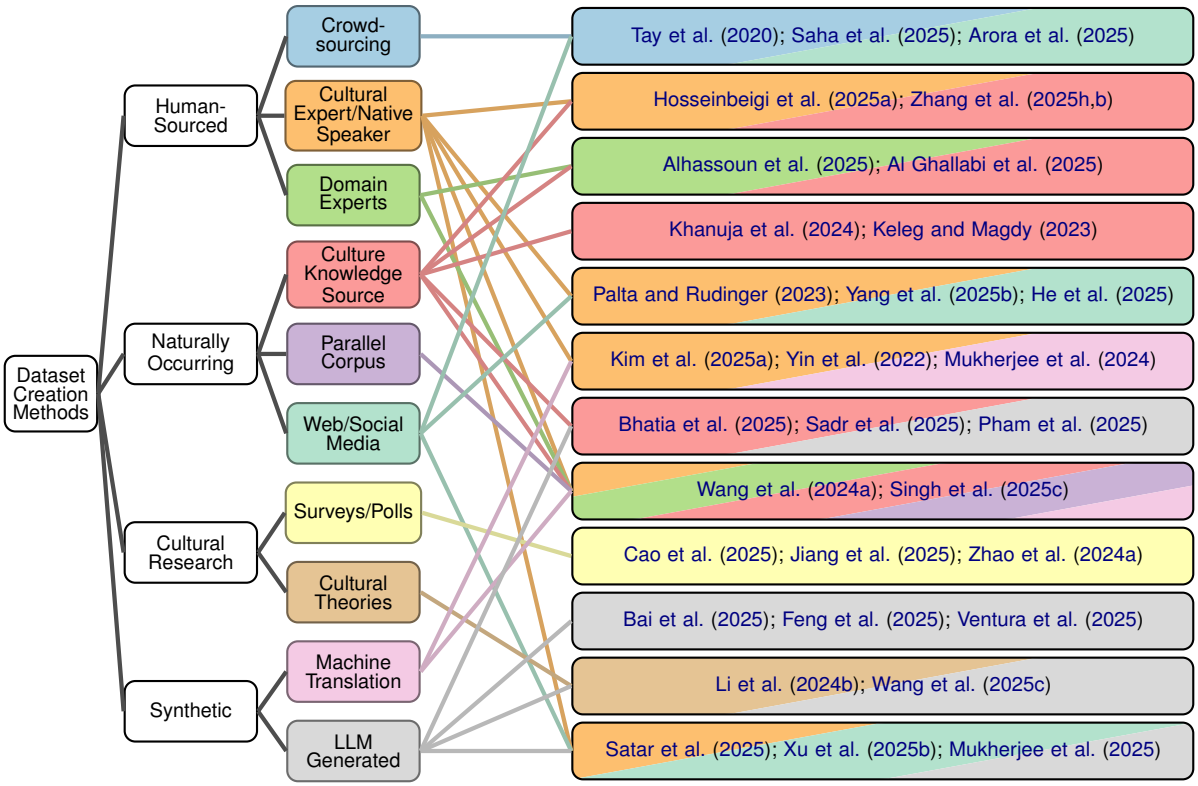


Figure 2: How are Cultural Datasets being created? Our taxonomy for Dataset Creation methods (solid color boxes), and example papers employing these methods (striped color boxes). Note that a paper can combine any number of data creation methods.

for tasks requiring cultural or linguistic expertise, such as labeling cultural aspects (Maji et al., 2025b; Alwajih et al., 2025a; Xie et al., 2025), creating culturally grounded content (Zhan et al., 2024; Guo et al., 2025), or assisting with translation (Kim et al., 2025a; Montalan et al., 2025).

Domain experts help ensure datasets align with the domains: law specialists developing annotation guidelines (Ullah et al., 2024), historians and archaeologists ensuring artifact accuracy (Ghaboura et al., 2025), and experts defining dataset taxonomies (Vasilev et al., 2025).

5.1.2 Naturally Occurring Sources

Web data provides unstructured cultural signals in online reviews (Zou et al., 2025) and social media to capture cross-cultural language use (Kumar and Jurgens, 2025; Wuraola et al., 2024; Abdelkadir et al., 2024; Kiesel et al., 2022; Liu et al., 2025e), and images for visual QA and reasoning (Liu et al., 2021, 2025e; Bayramli et al., 2025).

Parallel Corpora contain comparable content across languages, such as culturally relevant entities (Yao et al., 2024a; Conia et al., 2024), rules (Haberland et al., 2024), and topical images (Schneider and Sitaram, 2024).

Culture knowledge sources offer explicit and fine-grained cultural concepts and artifacts, reflecting the highest level of culture sensitivity: modifying existing datasets for a particular culture (Wang et al., 2024d; Grandury et al., 2025; Son et al., 2025), and filtering pre-existing cultural resources (Dai et al., 2025). Educational and domain-specific materials are also widely used: children's books (Khanuja et al., 2024), exams (Cheng et al., 2025), e-learning platforms (Pramodya et al., 2025), research journals (e.g., PubMed) (Nimo et al., 2025), literature (AbuHaija et al., 2025), and Wikipedia (Magdy et al., 2025; Bhatia et al., 2025).

5.1.3 Cultural Research

Culture Value Surveys, such as the World Values Survey (WVS) (Inglehart et al., 2000; Haerpfer et al., 2022), provide standardized value-oriented questions and empirical responses that researchers use to construct datasets: expanding question sets (Xu et al., 2025a), predicting answer distributions (Cao et al., 2025), or selecting dialogue topics based on response patterns (Ma et al., 2025b).

Cultural/Social Science Theories also guide

dataset creation. Frameworks such as Hofstede's Cultural Dimensions (Hofstede, 1984), the Theory of Basic Human Values (Schwartz, 1992), and negotiation theory (Aslani et al., 2016) are used to refine value taxonomies for question generation (Cahyawijaya et al., 2025a), categorize values (Yao et al., 2024b), and guide annotators (Hale et al., 2025).

5.1.4 Synthetic Generation

Machine Translation is used to expand datasets across languages. Work applies Google Translate to translate simple sentences (Belay et al., 2025; Yin et al., 2022) and literature (Thai et al., 2022). LLMs are used for translation (Onohara et al., 2025; Kim and Kim, 2025), augmentation (Masala et al., 2024), and verification of cultural alignment (Putri et al., 2024). Other work uses specialized translation models NLLB Team et al. (2022) for quality check (Aakanksha et al., 2024; Nguyen et al., 2024b), or trains models to support low-resource dialects (Mousi et al., 2025).

LLM Generation is increasingly used in dataset construction: generating culturally specific content such as stereotypes (Jha et al., 2023; Sahoo et al., 2024), moral scenarios (Liu et al., 2024a; Dey et al., 2025), and norms (CH-Wang et al., 2023); standardizing data formats (Chiu et al., 2025; Umbet et al., 2025), and adapting content across cultural contexts (Joshi et al., 2025; Putri et al., 2024). Beyond text, LLMs are used to synthesize images (Kim et al., 2025b), generate image captions (Bai et al., 2025), and label videos (Chen et al., 2025d).

5.2 Insights into Dataset Creation Pipelines

Our analysis reveals that modern cultural dataset creation rarely relies on a single, isolated method. Figure 2 illustrates common combinations of dataset creation approaches (additional details in Appendix C.3) with representative examples. We identify two dominant pipelines.

Pipeline 1: Source-Grounded Human Curation combines cultural knowledge sources with human expertise. Researchers typically adapt materials from books, image collections, and local websites, then ask humans to create prompts (Isbarov et al., 2025), questions (Nayak et al., 2024), and QA pairs (Limkonchotiwat et al., 2025), or annotate through quality checks (Kim et al., 2024a; Kim and Lee, 2025), label assignment (Maji et al., 2025a; Zhang et al., 2025b), or culture-specific translation (Winata et al., 2025), and enrich existing cultural datasets with expert knowledge (Cheng et al., 2025; Romanyshyn et al., 2024).

Pipeline 2: LLM Co-Creation uses LLMs as as generative **amplifiers** or **structuring tools** in dataset creation, with human input, theoretical frameworks, or existing language resources, including expanding seeded content with additional situated examples (Xu et al., 2025a; Zhan et al., 2024; Li et al., 2025); generating culturally diverse content followed by human verification (Urailertprasert et al., 2024; Qiu et al., 2025; CH-Wang et al., 2023); extending topic coverage (Wang et al., 2024f; Cahyawijaya et al., 2025a; Chiu et al., 2025); and extracting or generating additional information from existing language sources (Bhatia et al., 2025; Arnardóttir et al., 2025).

Despite the common use of these pipelines, we find several exceptions across **Cultural Capabilities** (Figure 9). Sociolinguistic Competence emphasize crowdsourced interactions to capture natural language use while avoiding explicit cultural cues that could create evaluation shortcuts. Cultural Translation rely more on language professionals (Akinade et al., 2023; Tapo et al., 2025a; Zhang et al., 2025e), and draw on web-based multilingual content, which captures contemporary expressions that often remain untranslated. Cultural Safety and Harm Reduction similarly depend on web data, particularly social media, to capture potentially harmful language needed for moderation (Maronikolakis et al., 2022; Mia et al., 2025). In contrast, Computational Cultural Representation use LLM-based synthetic generation to construct structured cultural knowledge that is difficult to obtain directly from raw data (Acquaye et al., 2024; Ziems et al., 2023; Pujari and Goldwasser, 2025).

We also identify the **conceptual rigor trade-off** in the literature. Compared to other data construction approaches, culture research based methods are used less frequently. This pattern suggests that current cultural NLP dataset creation relies more heavily on coarse-grained, **proxy-based data** sources than on structured guidance from well-established **conceptual frameworks in social science**. Such a trend highlights a potential trade-off between scalability and practical convenience on the one hand, and depth, conceptual rigor, and precision of cultural representation on the other.

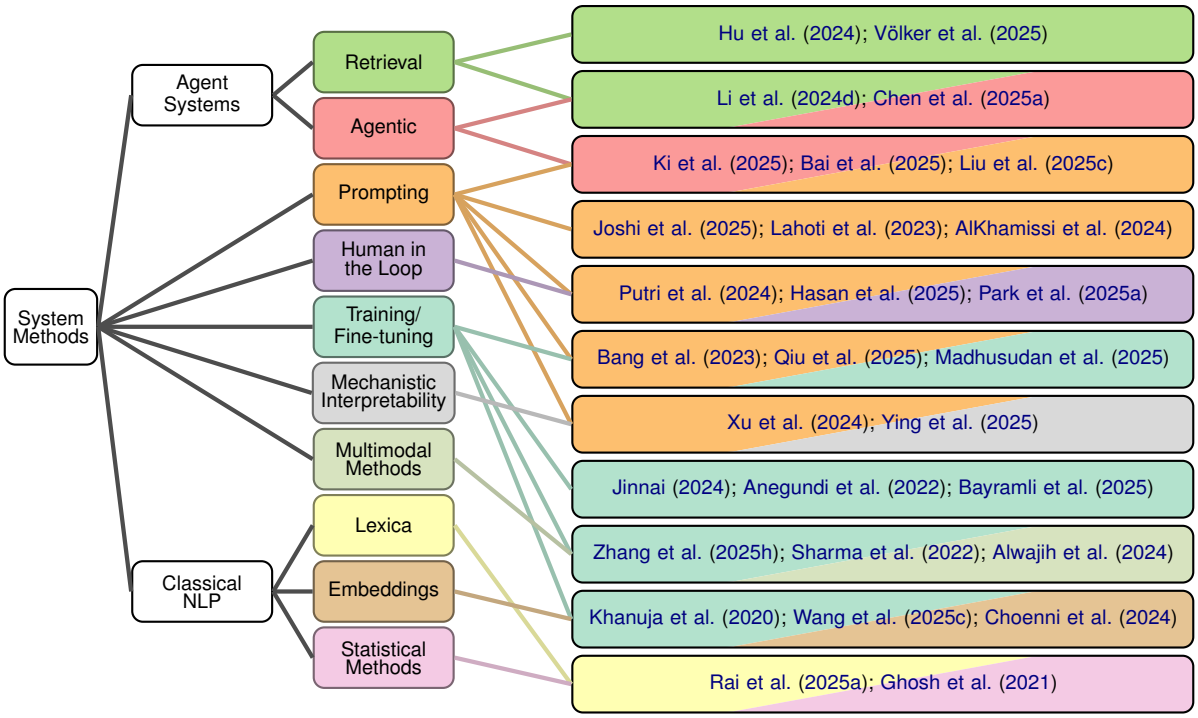


Figure 3: What methods are used to improve Cultural Capabilities? Our taxonomy of Systems (solid color boxes), and example papers employing these methods (striped color boxes). Note that a paper may combine any number of methods.

6 Systems for Cultural Capabilities

In this section we first discuss three different *kinds* of systems that we identified during our survey (§6.1), and then summarize some technical approaches adopted to improve CCs in NLP systems (§6.2).

6.1 Methodology Goals

Among all the papers that contributed a system, we identified three distinct overarching goals:

System Improvement for CC: Works that seek to directly *improve* an NLP System's CCs, e.g., (Ma et al., 2025b; Cao et al., 2025; Li et al., 2024d).

System Evaluation for CC: Works that propose a novel system to *evaluate* an NLP System's CCs, e.g., (Zhao et al., 2025; Mukherjee et al., 2023; Zhang et al., 2025a).

Cultural Data Generation: Works that propose a generalizable framework to *generate* culturally informed data, e.g., (Keleg and Magdy, 2023; Fung et al., 2023; Hasan et al., 2025).

The three system goals each have a very different impact on research, which makes this an important distinction. We allow the ability to query papers by this distinction in CULTUREMINE, and discuss insights from this classification in §7.

6.2 Technical Approaches to Methodology

Fig. 3 illustrates the taxonomy used to classify methods based on their technical approaches. Prominent themes for each approach are discussed below.

6.2.1 Prompting Based Methods

We identified over 70 papers that contributed a novel system utilizing prompting in various ways. One common approach is to vary the language of the prompt, to either probe a model (Kim and Kim, 2025; Zhao et al., 2025), or to encourage culturally aligned behavior (Feng et al., 2024a). Other methods use specialized prompts, such as injection of cultural knowledge (Shaikh et al., 2023; Ma et al., 2025b), guiding principles (AlKhamissi et al., 2024), multi-step prompting (Hobson et al., 2024), and cloze templates (Ramezani and Xu, 2023).

6.2.2 Agent Systems: Retrieval and Multi-Model Systems

We found Agent Systems, often with a "retriever agent" to be a common system proposed. What is the motivation for using an agentic framework? We find two recurring themes: (1) To allow for culturally diverse interactions and output (Ki et al., 2025; Li et al., 2024d; White et al., 2024; Bai et al., 2025;

Feng et al., 2024b), (2) To allow for specialized roles in complex systems made up of smaller modules (Anik et al., 2025; Wu et al., 2024; Yuan et al., 2024; Liang et al., 2025). For systems that utilized a retriever, what was the role of the retriever? The most common role we found was to retrieve diverse kinds of cultural information such as entities (Conia et al., 2024), recipes (Hu et al., 2024), medical text (Calvo-Bartolomé et al., 2025), poetry (Chen et al., 2025a). This would be explained by the fact that the most common CC associated with retrievers was Cultural Translation.

6.2.3 Human in the Loop Systems

We identified 13 papers that proposed a system involving a human in the loop. The main role of humans in these systems was to provide cultural expertise, which can look like judgments and corrections (Pujari and Goldwasser, 2025; Ziems et al., 2025), culturally informed seed data (Hasan et al., 2025; Rachamalla et al., 2025; Putri et al., 2024), adversarial input (Chiu et al., 2025), and even outside culture judgment (Park et al., 2025b).

6.2.4 Training Based Methods

Many papers propose novel training paradigms to improve the CC of NLP Systems. By far the most common one we identify is SFT on LLMs for some form of cultural alignment (Choenni et al., 2024; Cao et al., 2025; Xu et al., 2025a; Dai et al., 2025). Another common trend we notice is that of training smaller models for specialized tasks, such as Bert (Devlin et al., 2019) to identify values (Kiesel et al., 2022) or predict stereotypes (Kim and Johnson, 2025), e5 model (Wang et al., 2024b) to align sociocultural concepts (Wang et al., 2025c), or evaluation metrics trained to reflect human judgment (Bayramli et al., 2025). Finally, there are several papers that introduce novel architectures for specialized tasks such as subjective prediction (Parappan and Henao, 2025), predicting social relationships from videos (Zhang et al., 2025h), and transfer learning across languages (Ringel et al., 2019).

6.2.5 Classical NLP

In this subsection we discuss approaches which are often paired; embeddings, lexica, and methods over them like clustering, summarization.

Lexica: We identified 2 main ways that lexica are used in this field: (1) discovering cultural differences like ideologies (Milbauer et al., 2021), perspectives (Gutiérrez et al., 2016), expressions of mental health (Rai et al., 2025a), and personal values (Wilson et al., 2016), and (2) quantifying linguistic aspects like style (Havaldar et al., 2023a), bias, (Naous and Xu, 2025; Friedman et al., 2019) cultural awareness (Zhao et al., 2025; Caplan et al., 2025), harm (Menis Mastromichalakis et al., 2025), and concepts (Li and Zhang, 2023).

Embeddings: Lexical methods are often paired with embedding based methods for purposes like enriching the lexica (Havaldar et al., 2024, 2023a), comparing similarities or differences between cultures (Sun et al., 2021; Milbauer et al., 2021), and detecting cultural biases (Friedman et al., 2019). In contrast, (Rai et al., 2025a; Caplan et al., 2025) specifically avoid embeddings to prevent issues arising from bias in embeddings. Other works, exploit the bias in embeddings to measure abstract concepts like values, ideologies and identities (Cahyawijaya et al., 2025a; Milbauer et al., 2021; Ventura et al., 2025; Havaldar et al., 2024), and perceptions and pragmatics (Sun et al., 2021; Lin et al., 2018; White et al., 2024). Another common use of embeddings is to create a shared representation space for multiple cultures, in order to discover similarities and differences between them (Choenni et al., 2024; Zhou et al., 2023).

Statistical Methods: These methods encompass a wide variety of tools like clustering, topic modeling, evaluation metrics etc. We highlight two interesting recurring themes; the first is the use of clustering and topic modeling to discover common themes from large noisy data (Cuevas et al., 2025; Milbauer et al., 2021; Hobson et al., 2024; Pujari and Goldwasser, 2025), and the second is the use of various metrics to measure abstract concepts like cultural alignment (Wang et al., 2024c), ideologies (Milbauer et al., 2021), and values (Xu et al., 2024).

6.2.6 Mechanistic Interpretability

We identified few methods using mechanistic interpretability, which indicates it being an underexplored method for cultural NLP. The papers we did identify propose novel ways to answer why models display the cultural tendencies that they do (Xu et al., 2024; Ying et al., 2025).

6.2.7 Multimodal Methods

Several papers we surveyed proposed methods that included modalities beyond text. A very interesting line of work uses vision to account for low resource languages that do not have a surplus of textual data (Li and Zhang, 2023; Chen et al., 2024; R et al.,

2025). Another motivation for such methods is cultural competence over non-text modalities like vision (Sharma et al., 2022; Alwajih et al., 2024; Zhang et al., 2025c,h; Khanuja et al., 2025).

7 Observations and Recommendations for Future Work

7.1 Trends in Proposed Future Work

We analyzed the future work sections of the surveyed papers and found that community-proposed directions predominantly fall into two themes (details in Appendix B). First, researchers frequently call for **scaling cultural and linguistic coverage** to address data scarcity, often advocating for richer, multimodal benchmarks (Wang et al., 2024a; Maji et al., 2025a). Second, there is a strong push to **model cultural complexity** more accurately. Authors emphasize moving away from static, monolithic labels by operationalizing culture as a dynamic distribution (Havaldar et al., 2023a; Ziems et al., 2023), incorporating intersectional and fine-grained geographic variables beyond language (Falk et al., 2024; Koto et al., 2024), and developing evaluation frameworks explicitly accounting for pluralism (Zhou et al., 2023; Miehling et al., 2025).

7.2 Our Observations and Recommendations

In this section we discuss four broader patterns that offer promising avenues for the community.

Prioritizing system improvements over additional dataset creation. Dataset creation substantially outpaces the development of methods that directly improve CCs. This gap is especially pronounced for Cultural Knowledge (>100 surveyed **dataset** contributions), Cultural Safety and Harm Reduction (>60), and Value-driven Cultural Alignment (>50), each of which has fewer than 20 surveyed **system** improvement contributions. The abundance of such datasets (Singh et al., 2025c; Sahoo et al., 2025; Bui et al., 2025b; Shetty et al., 2025) indicates that the community has built a robust foundation for measuring these capabilities. We recommend that future efforts emphasize developing systems specifically designed to improve performance on these existing datasets, as done by (Ki et al., 2025; Wang et al., 2025b; Feng et al., 2024a; Parappan and Henao, 2025).

Reusing existing data generation frameworks for scalable data collection. When the creation of new data is desired, we observe an opportunity to reduce redundant effort. We identified several sophisticated, generalizable Cultural Data Generation systems (Ziems et al., 2025; Caplan et al., 2025). These frameworks are designed to produce datasets dynamically by conditioning on target variables (e.g., country, language, or source collection). However, we noticed limited follow-up reuse of these pipelines. Leveraging these existing generative frameworks can accelerate research when specific cultural data is scarce, providing a scalable alternative to curating datasets from scratch.

Developing specialized evaluation methods for implicit cultural nuances. Our survey suggests that measurement paradigms for different CCs are evolving at different rates. We use the ratio of novel *Evaluation System* papers to *Dataset* papers to approximate this focus: a higher ratio indicates active development of bespoke measurement frameworks, whereas a lower ratio implies a reliance on existing metrics. For example, capabilities involving abstract constructs, such as Value-driven Cultural Alignment (0.30) and Sociolinguistic Competence (0.21), exhibit relatively high ratios. This suggests that **standard metrics often fall short** for these CCs, prompting researchers to build specialized evaluators (Yao et al., 2024b; Shen et al., 2025; Casola et al., 2024; Ying et al., 2025). Conversely, capabilities like Multimodal Cultural Knowledge (0.02) and Cultural Translation (0.11) show significantly lower ratios, as they frequently **rely on established, reference-based metrics** like BLEU or ROUGE. While these standard metrics provide valuable baselines, we encourage the continued development of specialized evaluation methodologies for these domains as well.

This opportunity is particularly visible in multimodal work. Existing literature provides excellent coverage of visually salient cultural artifacts, such as food, clothing, and landmarks (Nayak et al., 2024; Bhatia et al., 2024; Maji et al., 2025c). Moving forward, the field is well-positioned to tackle subtler, "below-the-iceberg" phenomena (Hall, 1976) essential for real-world interaction, such as conversational grounding, gestures, paralinguistic interpretation, and cross-dialect understanding, e.g., (Sasu et al., 2025; Zhang et al., 2025h).

Evaluating and optimizing multiple CCs jointly. Our taxonomy defines CC as a collection of distinct competencies (§4.1), yet current research predominantly isolates them. Among surveyed papers con-

tributing *system improvements*, 82% target exactly one CC; for *system evaluation*, the figure is 94%. While this focused scope is necessary for foundational research, real-world deployments require systems to juggle **multiple CCs simultaneously**. Improvement on one CC does not inherently transfer to another (Zhang et al., 2025d); a culturally capable system deployed in the world may need to retrieve accurate cultural facts, communicate appropriately, and adapt to pluralistic norms concurrently. We encourage future work to report cross-CC transfer and trade-offs, and to develop comprehensive benchmarks that evaluate multiple CCs jointly.

8 CULTUREMINE: Community Resource

In this section we provide a brief overview of CULTUREMINE, which contains all the papers we surveyed, tagged with the dataset taxonomy and/or the systems taxonomy and relevant CCs. Additionally, each paper is tagged for culturally relevant metadata such as what proxy it used for culture (language, country, other). When applicable, papers are also tagged for the exact languages or regions that they account for. All the papers can be queried via drop-down filters, which makes it very easy for researchers to find a collection of work that is relevant for them! The top section of CULTUREMINE has a dynamic visualization giving an overview of papers and updates based on selected fields.

Additionally, we include a submission form that researchers can fill out to add more papers. We intend to keep the UI updated regularly, and hope that it can be a true mine for cultural NLP!

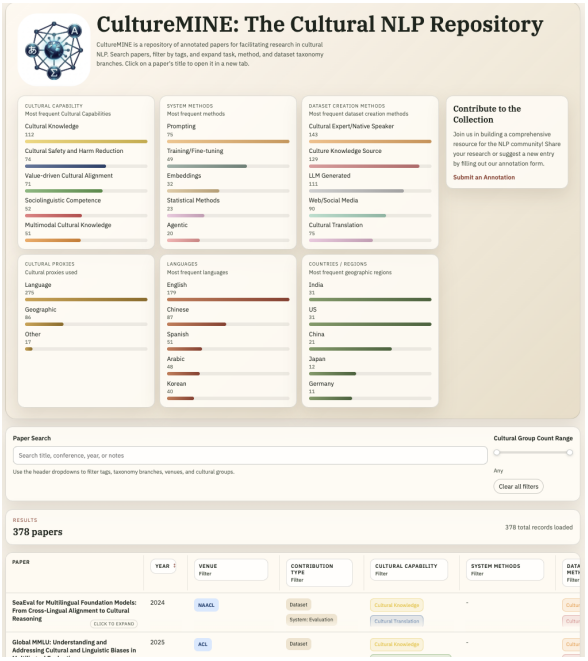


Figure 4: CULTUREMINE interface has dynamic visualizations in the top section. The fields in the bottom, which include our taxonomies for datasets and systems, can be used as filters to find relevant papers. The filters also include information such as cultural proxies used, what cultures, and what languages!

9 Conclusion

We surveyed over 375 papers, provided an overview of recent technical methods for improving the CCs of NLP systems, proposed taxonomies to organize contributions and released our surveyed papers, tagged with our taxonomies and other rich metadata via CULTUREMINE. We hope this survey and CULTUREMINE will be a valuable resource for the cultural NLP community.

Limitations

Despite our effort to provide a comprehensive overview of the cultural capabilities of NLP systems, several limitations remain.

Scope and Source Coverage. Our survey is not exhaustive. Research on culture in NLP is distributed across multiple venues and related fields, making comprehensive coverage infeasible. We therefore focus on papers that directly address our central question: how cultural capabilities are technically represented, operationalized, or incorporated in NLP systems and research designs. For the surveyed papers, we systematically collected them from the ACL Anthology, which serves as the primary repository. As a result, research published outside the ACL Anthology may be underrepresented in our survey.

Exclusion of NLP for Social Science. An important related area not included in this survey is NLP for social science. This line of research applies natural language processing methods to study social phenomena such as cultural variation, social meaning, and contextual interpretation in large-scale text data. Such work provides valuable theoretical and empirical perspectives for computational research on culture. Establishing a more systematic connection between the NLP for social science literature and culture-focused NLP research remains an important direction for future work.

Taxonomic Boundaries. The taxonomy proposed in this survey is derived through a bottom-up analysis of the papers we reviewed. It is not

intended to be a complete theory of all possible mechanisms for investigating cultural capabilities in NLP. As the literature develops, additional dimensions and categories may emerge.

Interpretative Constraints. Our analysis is constrained by how culture is defined and described in the surveyed papers. Cultural modeling choices are often only partially formalized or embedded within broader task or dataset design decisions. Consequently, some categorization decisions require interpretation.

Conceptual Proxy Limitations. Many studies represent culture through proxies such as language, geographic region, nationality, or demographic group. These proxies capture different aspects of culture and are not conceptually equivalent. Our survey organizes the current technical design space, but it does not resolve the underlying conceptual challenges of modeling culture.

Ethical Considerations

Culture is complex, dynamic, and internally heterogeneous. However, computational studies often operationalize culture using simplified proxies such as language, nationality, geographic region, or demographic group. These proxies can be practically useful for empirical analysis, but they may also reify culture as fixed or homogeneous, obscure within-group diversity, and inadvertently reinforce stereotypes or exclusions.

The goal of this survey is to organize and analyze the technical mechanisms through which prior work incorporates culture into NLP systems. We do not treat any particular operationalization of culture as definitive or exhaustive. Rather, we view these operationalizations as modeling choices made for specific empirical purposes. We therefore encourage future research to make these assumptions explicit, carefully justify the cultural proxies used, and evaluate the potential downstream risks associated with cultural generalization in NLP systems.

Acknowledgments

We thank the anonymous reviewers and Rajkumar Pujari, for their insightful comments that helped improve the paper. This project was partially funded by NSF IIS-2048001.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. 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 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Izza AbuHaija, Salim Al Mandhari, Mo El-Haj, Jonas Sibony, and Paul Rayson. 2025. The nakba lexicon: Building a comprehensive dataset from palestinian literature. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 37–47, Abu Dhabi. Association for Computational Linguistics.

Anurag Acharya, Diego Estrada, Shreeja Dahal, W. Victor H. Yarlott, Diana Gomez, and Mark Finlayson. 2024. Discovering implicit meanings of cultural motifs from text. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 46–56, Mexico City, Mexico. Association for Computational Linguistics.

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between Ghana and the US. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9483–9502, Miami, Florida, USA. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa

culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17421–17446, Suzhou, China. Association for Computational Linguistics.

Idris Akinade, Jesujoba O. Alabi, David Ifeoluwa Adelani, Clement Odoje, and Dietrich Klakow. 2023. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia. Association for Computational Linguistics.

Wafa Al Ghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan Pravin More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for Arabic poetry understanding in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20224–20244, Suzhou, China. Association for Computational Linguistics.

Manal Alhassoun, Imaan Mohammed Alkhanen, Nouf Alshalawi, Ibtehal Baazeem, and Waleed Alsanie. 2025. Saudi-alignment benchmark: Assessing LLMs alignment with cultural norms and domain knowledge in the saudi context. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 130–147, Suzhou, China. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai 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 (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Saied Alshahrani, Norah Alshahrani, Soumyabrata Dey, and Jeanna Matthews. 2023. Performance implications of using unrepresentative corpora in Arabic natural language processing. In *Proceedings of ArabicNLP 2023*, pages 218–231, Singapore (Hybrid). Association for Computational Linguistics.

Shatha Altammami. 2025. Leveraging AI to bridge classical Arabic and Modern Standard Arabic for text simplification. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 76–85, Abu Dhabi, UAE. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025a. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025b. PalmX 2025: The first shared task on benchmarking LLMs on Arabic and islamic culture. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 774–789, Suzhou, China. Association for Computational Linguistics.

Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 18 others. 2025c. Pearl: A multimodal culturally-aware Arabic instruction dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Selenia Anastasi, Florian Schneider, Chris Biemann, and Tim Fischer. 2024. VIDA: The visual incel data archive. a theory-oriented annotated dataset to enhance hate detection through visual culture. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 59–67, Mexico City, Mexico. Association for Computational Linguistics.

Aishwarya Anegundi, Konstantin Schulz, Christian Rauh, and Georg Rehm. 2022. Modelling cultural and socio-economic dimensions of political bias in

German tweets. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 29–40, Potsdam, Germany. KONVENS 2022 Organizers.

Mahfuz Ahmed Anik, Abdur Rahman, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. Preserving cultural identity with context-aware translation through multi-agent AI systems. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 51–60, Albuquerque, New Mexico. Association for Computational Linguistics.

Þórunn Arnardóttir, Elías Bjartur Einarsson, Garðar Ingvarsson Juto, Þorvaldur Páll Helgason, and Hafsteinn Einarsson. 2025. WikiQA-IS: Assisted benchmark generation and automated evaluation of Icelandic cultural knowledge in LLMs. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 64–73, Tallinn, Estonia. University of Tartu Library, Estonia.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for LLM safeguard evaluation. 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 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.

Soroush Aslani, Jimena Ramirez-Marin, Jeanne Brett, Jingjing Yao, Zhaleh Semnani-Azad, Zhi-Xue Zhang, Catherine Tinsley, Laurie Weingart, and Wendi Adair. 2016. Dignity, face, and honor cultures: A study of negotiation strategy and outcomes in three cultures. *Journal of Organizational Behavior*, 37(8):1178–1201. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.2095.

Muhammad Falensi Azmi, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. 2025. IndoSafety: Culturally grounded safety for LLMs in Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9135–9166, Suzhou, China. Association for Computational Linguistics.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. 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 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in 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 7580–7617, Albuquerque, New Mexico. Association for Computational Linguistics.

Yejin Bang, Tiezheng Yu, Andrea Madotto, Zhaojiang Lin, Mona Diab, and Pascale Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 311–325, Toronto, Canada. Association for Computational Linguistics.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. 2025. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31137–31155, Vienna, Austria. Association for Computational Linguistics.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin Grissom II, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. CULEMO: Cultural lenses on emotion - benchmarking LLMs for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909, Vienna, Austria. Association for Computational Linguistics.

Noam K. Benkler, Scott Friedman, Sonja Schmer-Galunder, Drisana Marissa Mosaphir, Robert P. Goldman, Ruta Wheelock, Vasanth Sarathy, Pavan Kantharaju, and Matthew D. McLure. 2024. Recognizing value resonance with resonance-tuned RoBERTa task definition, experimental validation, and robust modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13688–13698, Torino, Italia. ELRA and ICCL.

Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025. PANDA - paired anti-hate narratives dataset from Asia: Using an LLM-as-a-judge to create the first Chinese counterspeech dataset. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 1–12, Abu Dhabi, UAE. Association for Computational Linguistics.

Uri Berger and Edoardo Ponti. 2025. Cross-lingual and cross-cultural variation in image descriptions. 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 9453–9465, Albuquerque, New Mexico. Association for Computational Linguistics.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4669–4685, Albuquerque, New Mexico. Association for Computational Linguistics.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Shaily Bhatt, Tal August, and Maria Antoniak. 2025. Research borderlands: Analysing writing across research cultures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26238–26266, Vienna, Austria. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Xiaojun Bi, Shuo Li, Junyao Xing, Ziyue Wang, Fuwen Luo, Weizheng Qiao, Lu Han, Ziwei Sun, Peng Li, and Yang Liu. 2025. DongbaMIE: A multimodal information extraction dataset for evaluating semantic understanding of dongba pictograms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 976–990, Suzhou, China. Association for Computational Linguistics.

Houda Bouamor, Sara Al-Emadi, Zeinab Ibrahim, Hany Fazzaa, and Aisha Al-Sultan. 2025. Capturing intra-dialectal variation in qatari Arabic: A corpus of cultural and gender dimensions. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 219–230, Suzhou, China. Association for Computational Linguistics.

Minh Duc Bui, Kyung Eun Park, Goran Glavaš, Fabian David Schmidt, and Katharina Von Der Wense. 2025a. On generalization across measurement systems: LLMs entail more test-time compute for underrepresented cultures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21262–21276, Vienna, Austria. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025b. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–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 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Antoine Cadotte, Nathalie André, and Fatiha Sadat. 2024. Machine translation through cultural texts: Can verses and prose help low-resource indigenous models? In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 121–127, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025a. High-dimension human value representation in 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 5303–5330, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025b. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Lorena Calvo-Bartolomé, Valérie Aldana, Karla Cantarero, Alonso Madroñal de Mesa, Jerónimo Arenas-García, and Jordan Lee Boyd-Graber. 2025. Discrepancy detection at the data level: Toward consistent multilingual question answering. In *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing, pages 22013–22054, Suzhou, China. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. 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 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Eylon Caplan, Tania Chakraborty, and Dan Goldwasser. 2025. Splits! A Flexible Dataset and Evaluation Framework for Sociocultural Linguistic Investigation. *arXiv preprint*. ArXiv:2504.04640 [cs].

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.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Kyubyung Chae, Gihoon Kim, Gyuseong Lee, Taesup Kim, Jaejin Lee, and Heejin Kim. 2025. Assessing socio-cultural alignment and technical safety of sovereign LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10579–10600, Suzhou, China. Association for Computational Linguistics.

Nandu Chandran Nair, Rajendran S. Velayuthan, Yamini Chandrashekar, Gábor Bella, and Fausto Giunchiglia. 2022. IndoUKC: A concept-centered Indian multilingual lexical resource. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2833–2840, Marseille, France. European Language Resources Association.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025a. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14238–14254, Bangkok, Thailand. Association for Computational Linguistics.

Jiale Chen, Xuelian Dong, Qihao Yang, Wenxiu Xie, and Tianyong Hao. 2025b. Can large language models translate spoken-only languages through international phonetic transcription? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23420–23435, Suzhou, China. Association for Computational Linguistics.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025c. STEER-BENCH: A benchmark for evaluating the steerability of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18327–18355, Suzhou, China. Association for Computational Linguistics.

Xinyu Chen, Yunxin Li, Haoyuan Shi, Baotian Hu, Wenhan Luo, Yaowei Wang, and Min Zhang. 2025d. VideoVista-CulturalLingo: 360° horizons-bridging cultures, languages, and domains in video comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27102–27128, Vienna, Austria. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2535–2557, Vienna, Austria. Association for Computational Linguistics.

Alejandro Cuevas, Saloni Dash, Bharat Kumar Nayak, Dan Vann, and Madeleine I. G. Daepp. 2025. Anecdoctoring: Automated red-teaming across language and place. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19055–19074, Suzhou, China. Association for Computational Linguistics.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24510–24526, Suzhou, China. Association for Computational Linguistics.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369, Miami, Florida, USA. Association for Computational Linguistics.

Amitava Das, Yaswanth Narsupalli, Gurpreet Singh, Vinija Jain, Vasu Sharma, Suranjana Trivedy, Aman Chadha, and Amit Sheth. 2025. YinYang-align: A new benchmark for competing objectives and introducing multi-objective preference based text-to-image alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23518–23598, Vienna, Austria. Association for Computational Linguistics.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Nicholas Deas, Elsbeth Turcan, Ivan Ernesto Perez Mejia, and Kathleen McKeown. 2024. MASIVE: Open-ended affective state identification in English and Spanish. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20467–20485, Miami, Florida, USA. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Priyanka Dey, Aayush Bothra, Yugal Khanter, Jieyu Zhao, and Emilio Ferrara. 2025. Can LLMs express personality across cultures? introducing CulturalPersonas for evaluating trait alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20241–20262, Suzhou, China. Association for Computational Linguistics.

Ashutosh Dwivedi, Siddhant Shivdutt Singh, and Ashutosh Modi. 2025. EtiCor++: Towards understanding etiquettical bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9355–9376, Vienna, Austria. Association for Computational Linguistics.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical*

Methods in Natural Language Processing, pages 10967–10991, Suzhou, China. Association for Computational Linguistics.

Sugyeong Eo, Jungwoo Lim, Chanjun Park, DaHyun Jung, Seonmin Koo, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2024. Detecting critical errors considering cross-cultural factors in English-Korean translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4705–4716, Torino, Italia. ELRA and ICCL.

Elena V. Epure, Guillaume Salha, Manuel Moussallam, and Romain Hennequin. 2020. Modeling the music genre perception across language-bound cultures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4765–4779, Online. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, choclo and mazorca: on the varieties of Spanish. 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 3689–3711, Mexico City, Mexico. Association for Computational Linguistics.

Cristina España-Bonet, Ankur Bhatt, Koel Dutta Chowdhury, and Alberto Barrón-Cedeño. 2024. When elote, choclo and mazorca are not the same. isomorphism-based perspective to the Spanish varieties divergences. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 56–77, Mexico City, Mexico. Association for Computational Linguistics.

Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. Overview of PerspectiveArg2024 the first shared task on perspective argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 130–149, Bangkok, Thailand. Association for Computational Linguistics.

Jizhan Fang, Tianhe Lu, Yunzhi Yao, Ziyan Jiang, Xin Xu, Huajun Chen, and Ningyu Zhang. 2025. CKnowEdit: A new Chinese knowledge editing dataset for linguistics, facts, and logic error correction in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8789–8807, Vienna, Austria. Association for Computational Linguistics.

Mohammad Rifqi Farhansyah, Iwan Darmawan, Adryan Kusumawardhana, Genta Indra Winata, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. Do language models understand honorific systems in Javanese? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26732–26754, Vienna, Austria. Association for Computational Linguistics.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024a. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 4 others. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy. Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Time travel: A comprehensive benchmark to evaluate LMMs on historical and cultural artifacts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641, Vienna, Austria. Association for Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Nevan Giuliani, Cheng Charles Ma, Prakruthi Pradeep, and Daphne Ippolito. 2024. CAVA: A tool for cultural alignment visualization & analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 153–161, Miami, Florida, USA. Association for Computational Linguistics.

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, and 6 others. 2025. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524, Vienna, Austria. Association for Computational Linguistics.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. RuBia: A Russian language bias detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14227–14239, Torino, Italia. ELRA and ICCL.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China. Association for Computational Linguistics.

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Kevin Zhu, Austen Liao, and Sean O'Brien. 2025. EnDive: A cross-dialect benchmark for fairness and performance in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16830–16855, Suzhou, China. Association for Computational Linguistics.

E.D. Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian machine translation in its cultural context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017-2022) cross-national data-set. *(No Title)*.

Younggyun Hahm, Youngbin Noh, Ji Yoon Han, Tae Hwan Oh, Hyonsu Choe, Hansaem Kim, and Key-Sun Choi. 2020. Crowdsourcing in the development of a multilingual FrameNet: A case study of Korean FrameNet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 236–244, Marseille, France. European Language Resources Association.

James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett, and Jonathan Gratch. 2025. KODIS: A multicultural dispute resolution dialogue corpus. 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 12771–12785, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward T. Hall. 1976. *Beyond Culture*. Anchor Press/Doubleday, New York.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Abdullah Hashmat, Muhammad Arham Mirza, and Agha Ali Raza. 2025. PakBBQ: A culturally adapted bias benchmark for QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16160–16172, Suzhou, China. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023a. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023b. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32213–32230, Vienna, Austria. Association for Computational Linguistics.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards better benchmarking Chinese humor understanding from (ruo zhi ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21799–21818, Vienna, Austria. Association for Computational Linguistics.

Megan Herrera, Ankit Aich, and Natalie Parde. 2022. TweetTaglish: A dataset for investigating Tagalog-English code-switching. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2090–2097, Marseille, France. European Language Resources Association.

David G Hobson, Haiqi Zhou, Derek Ruths, and Andrew Piper. 2024. Story morals: Surfacing value-driven narrative schemas using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13032, Miami, Florida, USA. Association for Computational Linguistics.

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Minki Hong, Jangho Choi, and Jihie Kim. 2025. NormGenesis: Multicultural dialogue generation via exemplar-guided social norm modeling and violation recovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33793–33831, Suzhou, China. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025a. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, MohammadAli SeifKashani, Javad Seraj, Fatemeh Taherinezhad, Ali Nafisi, Fatemeh Nadi, Iman Barati, Hosein Hasani, Mostafa Amiri, and Mostafa Masoudi. 2025b. Matina: A culturally-aligned Persian language model using multiple LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20874–20889, Vienna, Austria. Association for Computational Linguistics.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704, Miami, Florida, USA. Association for Computational Linguistics.

Hsin-Yi Hsieh, Shang Wei Liu, Chang Chih Meng, Shuo-Yueh Lin, Chen Chien-Hua, Hung-Ju Lin, Hen-Hsen Huang, and I-Chen Wu. 2025. TaiwanVQA: A benchmark for visual question answering for Taiwanese daily life. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 57–75, Abu Dhabi, UAE. Association for Computational Linguistics.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Yujia Hu, Ming Shan Hee, Preslav Nakov, and Roy Ka-Wei Lee. 2025. Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12183–12201, Suzhou, China. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. 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 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Oana Ignat, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2025. InspAIred: Cross-cultural inspiration detection and analysis in real and LLM-generated social media data. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 35–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

Ishita and Radhika Mamidi. 2025. The evolution of gen alpha slang: Linguistic patterns and AI translation challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 678–686, Vienna, Austria. Association for Computational Linguistics.

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, and 2 others. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. 2025. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. 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 9543–9573, Albuquerque, New Mexico. Association for Computational Linguistics.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.

Raviraj Bhuminand Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Margaret Peters Long, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2666–2685, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. 2025. Evaluating multimodal language models as visual assistants for visually impaired users. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25949–25982, Vienna, Austria. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. Towards automatic evaluation for image transcreation. 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 7034–7047, Albuquerque, New Mexico. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877, Vienna, Austria. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. 2025a. Representing the under-represented: Cultural and core capability benchmarks for developing Thai large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4114–4129, Abu Dhabi, UAE. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024b. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16274–16289, Bangkok, Thailand. Association for Computational Linguistics.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025b. WHEN TOM EATS KIMCHI: Evaluating cultural awareness of multimodal large language models in cultural mixture contexts. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 143–154, Albuquerque, New Mexico. Association for Computational Linguistics.

Kyuhee Kim and Sangah Lee. 2025. Nunchi-bench: Benchmarking language models on cultural reasoning with a focus on Korean superstition. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15328–15342, Vienna, Austria. Association for Computational Linguistics.

Michelle YoungJin Kim and Kristen Johnson. 2025. Korean stereotype content model: Translating stereotypes across cultures. In *Proceedings of the 3rd*

Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), pages 59–70, Albuquerque, New Mexico. Association for Computational Linguistics.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 580–595, Vienna, Austria. Association for Computational Linguistics.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the Ukrainian language representation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 83–94, Torino, Italia. ELRA and ICCL.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Stefan Krsteski, Borjan Sazdov, Matea Tashkovska, Branislav Gerazov, and Hristijan Gjoreski. 2025. Towards open foundation language model and corpus for Macedonian: A low-resource language. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 44–57, Vienna, Austria. Association for Computational Linguistics.

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5890–5912, Vienna, Austria. Association for Computational Linguistics.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. McBE: A multi-task Chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056, Vienna, Austria. Association for Computational Linguistics.

Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. What are foundation models cooking in the post-soviet world? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20687–20709, Suzhou, China. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024a. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024b. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. HISTOIRES-MORALES: A French dataset for assessing moral alignment. 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 2590–2612, Albuquerque, New Mexico. Association for Computational Linguistics.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association*

for Computational Linguistics: ACL 2025, pages 4215–4241, Vienna, Austria. Association for Computational Linguistics.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024a. Uncovering differences in persuasive language in Russian versus English Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 21–35, Miami, Florida, USA. Association for Computational Linguistics.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: boosting cross-cultural understanding in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024c. CulturePark: Boosting Cross-cultural Understanding in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

Sha Li, Revanth Gangi Reddy, Khanh Duy Nguyen, Qingyun Wang, Yi Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare R. Voss, and Heng Ji. 2024d. Schema-guided culture-aware complex event simulation with multi-agent role-play. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 372–381, Miami, Florida, USA. Association for Computational Linguistics.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024e. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024f. CIF-bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, Bangkok, Thailand. Association for Computational Linguistics.

Zhi Li and Yin Zhang. 2023. Cultural concept adaptation on multimodal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore. Association for Computational Linguistics.

Jinggui Liang, Dung Vo, Yap Hong Xian, Hai Leong Chieu, Kian Ming A. Chai, Jing Jiang, and Lizi Liao. 2025. Colloquial singaporean English style transfer with fine-grained explainable control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26962–26983, Vienna, Austria. Association for Computational Linguistics.

Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, and Maite Melero. 2025. Culture-aware machine translation: the case study of low-resource language pair Catalan-Chinese. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 150–161, Geneva, Switzerland. European Association for Machine Translation.

Peerat Limkonchotiwat, Pume Tuchinda, Lalita Lowphansirikul, Surapon Nonesung, Panuthep Tasawong, Alham Fikri Aji, Can Udomcharoenchaikit, and Sarana Nutanong. 2025. WangchanThaiInstruct: An instruction-following dataset for culture-aware, multitask, and multi-domain evaluation in Thai. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3535–3558, Suzhou, China. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025b. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haijiang Liu, Qiyuan Li, Chao Gao, Yong Cao, Xiangyu Xu, Xun Wu, Daniel Hershcovich, and Jinguang Gu. 2025c. Beyond demographics: Enhancing cultural value survey simulation with multi-stage personality-driven cognitive reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pages 18406–18428, Suzhou, China. Association for Computational Linguistics.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona T. Diab, and Maarten Sap. 2025d. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16428–16458, Suzhou, China. Association for Computational Linguistics.

Xuelin Liu, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu. 2024a. Evaluating moral beliefs across LLMs through a pluralistic framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4740–4760, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025e. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492, Vienna, Austria. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, and Nancy Chen. 2024b. Optimizing code-switching in conversational tutoring systems: A pedagogical framework and evaluation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Kyoto, Japan. Association for Computational Linguistics.

Andrés Lou, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Víctor Sánchez-Cartagena. 2024. Curated datasets and neural models for machine translation of informal registers between Mayan and Spanish vernaculars. 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 2838–2850, Mexico City, Mexico. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Weicheng Ma, John J. Guerrerio, and Soroush Vosoughi. 2025a. Scalable and culturally specific stereotype dataset construction via human-LLM collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23928–23956, Suzhou, China. Association for Computational Linguistics.

Weicheng Ma, Hefan Zhang, Shiyu Ji, Farnoosh Hashemi, Qichao Wang, Ivory Yang, Joice Chen, Juanwen Pan, Michael Macy, Saeed Hassanpour, and Soroush Vosoughi. 2025b. Enhancing LLM-based persuasion simulations with cultural and speaker-specific information. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14955–14976, Suzhou, China. Association for Computational Linguistics.

Sangmitra Madhusudan, Robert Morabito, Skye Reid, Nikta Gohari Sadr, and Ali Emami. 2025. Fine-Tuned LLMs are "Time Capsules" for Tracking Societal Bias Through Books. 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 2329–2358, Albuquerque, New Mexico. Association for Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. 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 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, and Sriparna Saha. 2025a. SANSKRITI: A comprehensive benchmark for evaluating language models' knowledge of Indian culture. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4434–4451, Vienna, Austria. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, and Sriparna Saha. 2025b. DRISHTIKON: A multimodal multilingual benchmark for testing language models' understanding on Indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313, Suzhou, China. Association for Computational Linguistics.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka, Nemil Shah, Abhilekh Borah, Vanshika Shah, Nishant Mishra, and Sriparna Saha. 2025c. DRISHTIKON: A Multimodal Multilingual Benchmark for Testing Language Models' Understanding on Indian Culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1313, Suzhou, China. Association for Computational Linguistics.

Ananya Malik, Nazanin Sabri, Melissa M. Karnaze, and Mai ElSherief. 2025. Are LLMs Empathetic to All? Investigating the Influence of Multi-Demographic Personas on a Model's Empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24938–24959, Suzhou, China. Association for Computational Linguistics.

Antonis Maronikolakis, Abdullatif Köksal, and Hinrich Schuetze. 2024. Sociocultural knowledge is needed for selection of shots in hate speech detection tasks. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 1–13, St. Julian's, Malta. Association for Computational Linguistics.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "Vorbești Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Jason Liartis, Kristina Rose, Antoine Isaac, and Giorgos Stamou. 2025. Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21836–21850, Vienna, Austria. Association for Computational Linguistics.

Md Ayon Mia, Akm Moshiur Rahman Mazumder, Khadiza Sultana Sayma, Md Fahim, Md Tahmid Hasan Fuad, Muhammad Ibrahim Khan, and Akmmahbubur Rahman. 2025. BANMIME : Misogyny detection with metaphor explanation on Bangla memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17813–17839, Suzhou, China. Association for Computational Linguistics.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. 2025. Evaluating the prompt steerability of 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 7874–7900, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4832–4845, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in 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 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Luca Mitran, Sophie Wu, and Andrew Piper. 2025. Probing narrative morals: A new character-focused MFT framework for use with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28514–28529, Suzhou, China. Association for Computational Linguistics.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Hadi Mohammadi, Yasmeen F. S. S. Meijer, Efthymia Papadopoulou, and Ayoub Bagheri. 2025. Do large language models understand morality across cultures? In *Proceedings of the 2nd LUHME Workshop*, pages 30–39, Bologna, Italy. UP - Universidade

do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez Ii, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31239–31273, Vienna, Austria. Association for Computational Linguistics.

Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. PerCul: A story-driven cultural evaluation of LLMs in Persian. 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 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.

Guy Mor-Lan, Naama Rivlin-Angert, Yael R. Kaplan, Tamir Sheafer, and Shaul R. Shenhav. 2025. HebID: Detecting social identities in Hebrew-language political text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9850–9870, Suzhou, China. Association for Computational Linguistics.

David R. Mortensen, Xinyu Zhang, Chenxuan Cui, and Katherine J. Zhang. 2022. A Hmong corpus with elaborate expression annotations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4992–5000, Marseille, France. European Language Resources Association.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Hamdy Mubarak, Abubakr Mohamed, and Majd Hawasly. 2025. AraSafe: Benchmarking safety in Arabic LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9976–9992, Suzhou, China. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. 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 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Anjishnu Mukherjee, Aylin Caliskan, Ziwei Zhu, and Antonios Anastasopoulos. 2024. Global gallery: The fine art of painting culture portraits through multilingual instruction tuning. 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 6398–6415, Mexico City, Mexico. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Sourabrata Mukherjee, Atharva Mehta, Sougata Saha, Akhil Arora, and Monojit Choudhury. 2025. Women, infamous, and exotic beings: A comparative study of honorific usages in Wikipedia and LLMs for Bengali and Hindi. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19103–19126, Suzhou, China. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Janki Atul Nawale, Mohammed Safi Ur Rahman Khan, Janani D, Mansi Gupta, Danish Pruthi, and Mitesh M Khapra. 2025. FairI tales: Evaluation of fairness in Indian contexts with a focus on bias and stereotypes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30331–30380, Vienna, Austria. Association for Computational Linguistics.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 312–327, Mexico City, Mexico. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024a. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1774–1784, New York, NY, USA. Association for Computing Machinery.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024b. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts. 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 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.

Charles Nimo, Shuheng Liu, Irfan Essa, and Michael L. Best. 2025. Africa health check: Probing cultural bias in medical LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32219–32232, Suzhou, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, and Vukosi Marivate. 2025. AfroCS-xs: Creating a compact, high-quality, human-validated code-switched dataset for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33391–33410, Vienna, Austria. Association for Computational Linguistics.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. JMMMU: A Japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. 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 932–950, Albuquerque, New Mexico. Association for Computational Linguistics.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases

in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2025. Benchmarking multimodal models for Ukrainian language understanding across academic and cultural domains. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 14–26, Vienna, Austria (online). Association for Computational Linguistics.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20322–20338, Suzhou, China. Association for Computational Linguistics.

ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2025a. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-VLM collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21960–21974, Vienna, Austria. Association for Computational Linguistics.

Junyeong Park, Seogyeong Jeong, Seyoung Song, Yohan Lee, and Alice Oh. 2025b. LLM-C3MOD: A human-LLM collaborative system for cross-cultural hate speech moderation. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 71–88, Albuquerque, New Mexico. Association for Computational Linguistics.

Seoyoon Park, Jaehee Kim, and Hansaem Kim. 2025c. Too polite to be human: Evaluating LLM empathy in Korean conversations via a DCT-based framework. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pages 76–89, Vienna, Austria. Association for Computational Linguistics.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2025. CultureInstruct: Curating multi-cultural instructions at scale. 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 9207–9228, Albuquerque, New Mexico. Association for Computational Linguistics.

Viet Thanh Pham, Shilin Qu, Farhad Moghimifar, Suraj Sharma, Yuan-Fang Li, Weiqing Wang, and Reza Haf. 2024. Multi-cultural norm base: Frame-based norm discovery in multi-cultural settings. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 24–35, Miami, FL, USA. Association for Computational Linguistics.

Prisca Piccirilli, Alexander Fraser, and Sabine Schulte im Walde. 2024. VOLIMET: A Parallel Corpus of Literal and Metaphorical Verb-Object Pairs for English–German and English–French. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 222–237, Mexico City, Mexico. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Rhitabrat Pokharel and Ameeta Agrawal. 2025. neDIOM: Dataset and analysis of Nepali idioms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 160–171, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. 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 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Ashmari Pramodya, Nirasha Nelki, Heshan Shalinda, Chamila Liyanage, Yusuke Sakai, Randil Pushpananda, Ruvan Weerasinghe, Hidetaka Kamigaito, and Taro Watanabe. 2025. SinhalaMMLU: A comprehensive benchmark for evaluating multitask language understanding in Sinhala. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32943–32961, Suzhou, China. Association for Computational Linguistics.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural

preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 7–14, Mexico City, Mexico. Association for Computational Linguistics.

Rajkumar Pujari and Dan Goldwasser. 2025. LLM-human pipeline for cultural grounding of conversations. 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 1029–1048, Albuquerque, New Mexico. Association for Computational Linguistics.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Karthick Narayanan R, Siddharth Singh, Saurabh Singh, Aryan Mathur, Ritesh Kumar, Shyam Ratan, Bornini Lahiri, Benu Pareek, Neerav Mathur, Amalesh Gope, Meiraba Takhellambam, and Yogesh Dawer. 2025. Field to model: Pairing community data collection with scalable NLP through the LiFE suite. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 76–84, Vienna, Austria. Association for Computational Linguistics.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and curating high-quality cultural post-training datasets for Indian languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 285–321, Suzhuo, China. Association for Computational Linguistics.

Sunny Rai, Khushi Shelat, Devansh Jain, Ashwin Kishen, Young Min Cho, Maitreyi Redkar, Samindara Hardikar-Sawant, Lyle Ungar, and Sharath Chandra Guntuku. 2025a. Cross-cultural differences in mental health expressions on social media. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 132–142, Albuquerque, New Mexico. Association for Computational Linguistics.

Sunny Rai, Khushang Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2025b. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice. 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 11396–11415, Albuquerque, New Mexico. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2025. The discordance between embedded ethics and cultural inference in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14715–14736, Suzhou, China. Association for Computational Linguistics.

Artem Reshetnikov and Maria-Cristina Marinescu. 2025. Caption generation in cultural heritage: Crowdsourced data and tuning multimodal large language models. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 42–50, Albuquerque, New Mexico. Association for Computational Linguistics.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. Cross-cultural transfer learning for text classification. 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 3873–3883, Hong Kong, China. Association for Computational Linguistics.

Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. XLQA: A benchmark for locale-aware multilingual open-domain question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28809–28821, Suzhou, China. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Donya Rooein, Vilém Zouhar, Debora Nozza, and Dirk Hovy. 2025. Biased Tales: Cultural and Topic Bias in Generating Children's Stories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 52–72, Suzhou, China. Association for Computational Linguistics.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your LLM must learn the Persian art of taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838, Suzhou, China. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6075–6089, Albuquerque, New Mexico. Association for Computational Linguistics.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can LLMs identify cross-cultural communication gaps? 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 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. 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 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2025. DIWALI - diversity and inclusivity aWare cuLture specific items for India: Dataset and assessment of LLMs for cultural text adaptation in Indian context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33599–33626, Suzhou, China. Association for Computational Linguistics.

Edward Sapir. 1929. The status of linguistics as a science. *Language*, 5(4):207–214.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. Akan cinematic emotions (ACE): A multimodal multi-party dataset for emotion recognition in movie dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9820–9831, Vienna, Austria. Association for Computational Linguistics.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria. Association for Computational Linguistics.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. 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 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Karolina Seweryn, Anna Kołos, Agnieszka Karlińska, Katarzyna Lorenc, Katarzyna Dziewulska, Maciej Chrabaszcz, Aleksandra Krasnodebska, Paula Betscher, Zofia Cieślińska, Katarzyna Kowol, Julia Moska, Dawid Motyka, Paweł Walkowiak, Bartosz Żuk, and Arkadiusz Janz. 2025. PLLuM-align: Polish preference dataset for large language model alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23879–23908, Suzhou, China. Association for Computational Linguistics.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian

Farestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19998–20022, Suzhou, China. Association for Computational Linguistics.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. VITAL: A new dataset for benchmarking pluralistic alignment in healthcare. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Daiki Shiono, Ana Brassard, Yukiko Ishizuki, and Jun Suzuki. 2025. Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11428–11444, Abu Dhabi, UAE. Association for Computational Linguistics.

Pratik Rakesh Singh, Kritarth Prasad, Mohammadi Zaki, and Pankaj Wasnik. 2025a. Graph-assisted culturally adaptable idiomatic translation for Indic languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7029–7044, Vienna, Austria. Association for Computational Linguistics.

Punit Kumar Singh, Nishant Kumar, Akash Ghosh, Kunal Pasad, Khushi Soni, Manisha Jaishwal, Sriparna Saha, Syukron Abu Ishaq Alfarozi, Asres Temam Abagissa, Kitsuchart Pasupa, Haiqin Yang, and Jose G Moreno. 2025b. Let's Play Across Cultures: A Large Multilingual, Multicultural Benchmark for Assessing Language Models' Understanding of Sports. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15194–15241, Suzhou, China. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025c. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. 2025. A multilingual, culture-first approach to addressing misgendering in LLM applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31159–31183, Suzhou, China. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. 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 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International*

Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 7993–8007, Torino, Italia. ELRA and ICCL.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025. Evaluating dialect robustness of language models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2403–2414, Online. Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. 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 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Xin Tan, Bowei Zou, and AiTi Aw. 2025. A benchmark for translations across styles and language variants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2389–2402, Suzhou, China. Association for Computational Linguistics.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do You Know About My Nation? Investigating Multilingual Language Models' Cultural Literacy Through Factual Knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346. _eprint: https://academic.oup.com/pnasnexus/article-pdf/3/9/pgae346/59151559/pgae346.pdf.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri. 2025a. Bayelemabaga: Creating resources for Bambara NLP. 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 12060–12070, Albuquerque, New Mexico. Association for Computational Linguistics.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, and Michael Leventhal. 2025b. GAIfE: Using GenAI to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7929–7944, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi Tay, Donovan Ong, Jie Fu, Alvin Chan, Nancy Chen, Anh Tuan Luu, and Chris Pal. 2020. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373, Online. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14403–14416, Vienna, Austria. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China. Association for Computational Linguistics.

Ayuto Tsutsumi and Yuu Jinnai. 2025. Do large language models know folktales? a case study of yokai in Japanese folktales. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16124–16146, Vienna, Austria. Association for Computational Linguistics.

Faizad Ullah, Ali Faheem, Ubaid Azam, Muhammad Sohaib Ayub, Faisal Kamiran, and Asim Karim. 2024. Detecting cybercrimes in accordance with Pakistani law: Dataset and evaluation using PLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4717–4728, Torino, Italia. ELRA and ICCL.

Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. KazBench-KK: A cultural-knowledge benchmark for Kazakh. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. 2025. RusCode: Russian cultural code benchmark for text-to-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7656–7672, Albuquerque, New Mexico. Association for Computational Linguistics.

Justin Vasselli, Eunike Andriani Kardinata, Yusuke Sakai, and Taro Watanabe. 2025. Multilingual dialogue generation and localization with dialogue act scripting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32896–32911, Suzhou, China. Association for Computational Linguistics.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2025. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *Transactions of the Association for Computational Linguistics*, 13:142–166.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. 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 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Emilio Villa-Cueva, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Vladimir Araujo, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, Jan Christian Blaise Cruz, Mary Dabre, Raj Dabre, Toqeer Ehsan, Naome A Etori, Fauzan Farooqui, Jiahui Geng, Guido Ivetta, and 16 others. 2025. CaMMT: Benchmarking culturally aware multimodal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22423–22441, Suzhou, China. Association for Computational Linguistics.

Tom Völker, Jan Pfister, and Andreas Hotho. 2025. SALT at SemEval-2025 task 2: A SQL-based approach for LLM-free entity-aware-translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 852–864, Vienna, Austria. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. 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 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025a. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662, Vienna, Austria. Association for Computational Linguistics.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025b. Multilingual prompting for improving LLM

generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6367–6389, Suzhou, China. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024c. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024d. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024e. CMB: A comprehensive medical benchmark in Chinese. 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 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024f. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Zeqiang Wang, Jon Johnson, and Suparna De. 2025c. DLIR: Spherical adaptation for cross-lingual knowledge transfer of sociological concepts alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2061–2075, Suzhou, China. Association for Computational Linguistics.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932, Miami, Florida, USA. Association for Computational Linguistics.

Yuting Wei, Yuanxing Xu, Xinru Wei, Simin Yang, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. AC-EVAL: Evaluating Ancient Chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617, Miami, Florida, USA. Association for Computational Linguistics.

Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to play: Pragmatic reasoning for efficient cross-cultural communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12201–12216, Miami, Florida, USA. Association for Computational Linguistics.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Steven Wilson, Rada Mihalcea, Ryan Boyd, and James Pennebaker. 2016. Disentangling topic models: A cross-cultural analysis of personal values through words. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 143–152, Austin, Texas. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. SocialCC: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Xie, Chenkai Wang, Zongyang Ma, and Fahui Miao. 2025. Are large language models chronically online surfers? a dataset for Chinese Internet meme explanation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17062–17083, Suzhou, China. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025a. Self-pluralising culture alignment for 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 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. 2025b. PunMemeCN: A benchmark to explore vision-language models' understanding of Chinese pun memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18694–18710, Suzhou, China. Association for Computational Linguistics.

Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. 2025c. PunMemeCN: A Benchmark to Explore Vision-Language Models' Understanding of Chinese Pun Memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18694–18710, Suzhou, China. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. Beyond words: Exploring cultural value sensitivity in multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7607–7623, Albuquerque, New Mexico. Association for Computational Linguistics.

Silvana Yakhni and Ali Chehab. 2025. Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2025. Bias mitigation or cultural commonsense? evaluating LLMs with a Japanese dataset. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17295–17313, Suzhou, China. Association for Computational Linguistics.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025a. NüshuRescue: Reviving the Endangered Nüshu Language with AI. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034, Abu Dhabi, UAE. Association for Computational Linguistics.

Senqi Yang, Dongyu Zhang, Jing Ren, Ziqi Xu, Xiuzhen Zhang, Yiliao Song, Hongfei Lin, and Feng Xia. 2025b. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317, Vienna, Austria. Association for Computational Linguistics.

Yahan Yang, Soham Dan, Shuo Li, Dan Roth, and Insup Lee. 2025c. MrGuard: A multilingual reasoning guardrail for universal LLM safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27377–27396, Suzhou, China. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024a. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024b. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. 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 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

Amir Hossein Yari and Fajri Koto. 2025. Unveiling cultural blind spots: Analyzing the limitations of mLLMs in procedural text comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20151–20170, Vienna, Austria. Association for Computational Linguistics.

W. Victor Yarlott, Anurag Acharya, Diego Castro Estrada, Diana Gomez, and Mark Finlayson. 2024.

GOLEM: GOld standard for learning and evaluation of motifs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7801–7813, Torino, Italia. ELRA and ICCL.

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the gesture: Evaluating AI sensitivity to culturally offensive non-verbal gestures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25041–25080, Vienna, Austria. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Da Yin, Haoyi Qiu, Kung-Hsiang Huang, Kai-Wei Chang, and Nanyun Peng. 2024. SafeWorld: Geo-Diverse Safety Alignment. In *Advances in Neural Information Processing Systems*, volume 37, pages 128734–128768. Curran Associates, Inc.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, and 2 others. 2025. INJONGO: A multicultural intent detection and slot-filling dataset for 16 African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452, Vienna, Austria. Association for Computational Linguistics.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Tao Liu, and Deyi Xiong. 2024. CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837, Bangkok, Thailand. Association for Computational Linguistics.

Ye Yuan, Kexin Tang, Jianhao Shen, Ming Zhang, and Chenguang Wang. 2024. Measuring social norms of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 650–699, Mexico City, Mexico. Association for Computational Linguistics.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco-bench-MIF: On multilingual instruction-following capability of large language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072, Vienna, Austria. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

Chen Zhang, Zhiyuan Liao, and Yansong Feng. 2025a. Cross-lingual transfer of cultural knowledge: An asymmetric phenomenon. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 147–157, Vienna, Austria. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024a. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang,

Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025b. Can MLLMs understand the deep implication behind Chinese images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025c. LLM-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19407–19417, Suzhou, China. Association for Computational Linguistics.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shisong Chen, Yanghua Xiao, Hongwei Feng, Jiaqing Liang, Minggui HE, Shimin Tao, and Hongxia Ma. 2025d. CultureScope: A Dimensional Lens for Probing Cultural Understanding in LLMs. *arXiv preprint*. ArXiv:2509.16188 [cs].

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025e. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, China. Association for Computational Linguistics.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, and Salman Avestimehr. 2025f. Creating a lens of Chinese culture: A multimodal dataset for Chinese pun rebus art understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22473–22487, Vienna, Austria. Association for Computational Linguistics.

Xinyu Zhang, Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, and Fei Huang. 2025g. CultureSynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10448–10467, Suzhou, China. Association for Computational Linguistics.

Yuxuan Zhang, Yangfu Zhu, Haorui Wang, and Bin Wu. 2025h. Interesting culture: Social relation recognition from videos via culture de-confounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5174–5184, Suzhou, China. Association for Computational Linguistics.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024b. Cultural adaptation of menus: A fine-grained approach. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA. Association for Computational Linguistics.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024a. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Yuan Zhao, Ruiquan Zhang, Dengfeng Yao, and Yidong Chen. 2024b. Translation quality evaluation of sign language avatar. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 405–415, Taiyuan, China. Chinese Information Processing Society of China.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2132–2151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiayou Zhong, Anudeex Shetty, Chao Jia, Xuanrui Lin, and Usman Naseem. 2025. Pluralistic alignment for healthcare: A role-driven framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31320–31343, Suzhou, China. Association for Computational Linguistics.

Haiqi Zhou, David G Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155, Bangkok, Thailand. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025a. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. 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 9840–9867,

Albuquerque, New Mexico. Association for Computational Linguistics.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025b. Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24616–24638, Suzhou, China. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025c. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, William Barr Held, Jane Yu, Amir Goldberg, David Grusky, and Diyi Yang. 2025. Culture cartography: Mapping the landscape of cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1757, Suzhou, China. Association for Computational Linguistics.

Chenye Zou, Xingyue Wen, Tianyi Hu, Qian Janice Wang, and Daniel Hershcovich. 2025. Do LLMs understand wine descriptors across cultures? a benchmark for cultural adaptations of wine reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1875–1894, Suzhou, China. Association for Computational Linguistics.

A Detailed Literature Collection and Annotation Methodology

A.1 Phase 1: Initial Seed Collection

We took a two-phased approach to collect relevant literature, focusing on the past five years (including all available papers from 2025) to capture the modern landscape of Cultural NLP. The first phase consisted of a manual search combined with an automated scrape of the ACL Anthology. We queried for papers containing the words "culture" or "cultural" in the title, alongside at least one other keyword related to investigating or operationalizing culture in NLP. This initial highly targeted search resulted in a seed set of 180 papers.

A.2 Phase 2: Broad Scrape and LLM-Assisted Filtering

To ensure comprehensive coverage, the second phase involved a broader scrape of the ACL Anthology for any paper containing "culture" or "cultural" in either the title or abstract, returning 1,216 additional candidates. Given the volume, we employed an LLM to assist in filtering the list down to a highly relevant subset.

First, we prompted the LLM with a description of our inclusion criteria along with the title and abstract of each paper, retaining only those assigned a high relevance score. Next, we prompted the LLM with stricter inclusion guidelines, providing it with potentially relevant full-text sections of each candidate paper (e.g., Introduction, Dataset, Methods, Evaluation). To ensure we analyzed works with thorough descriptions and evaluations of proposed datasets or methods, we filtered out short papers, considering only those longer than 6 pages. This rigorous second phase yielded 277 additional highly relevant papers.

A.3 Manual Annotation and Consensus Strategy

The combined pool of candidate papers was manually processed by annotators who are Computer Science PhD students actively involved in NLP research. Papers were annotated across three main dimensions: (1) Cultural Capabilities addressed, (2) dataset creation methodology, and (3) technical system methods utilized. Throughout the annotation process, we performed a manual filtering step to discard papers that did not pose a novel contribution in any of these three aspects. Specifically, we removed works that solely evaluated existing models on existing datasets, or papers utilizing NLP tools for pure sociolinguistic or social science analyses without proposing a technical system or dataset improvement. This exclusion phase removed 79 papers, leaving a final corpus of 378 papers to be read thoroughly and annotated.

To ensure high-quality and consistent annotations, regular meetings were held between the surveyors. During these meetings, precise definitions of the labels were discussed, disagreements were resolved, and edge cases were evaluated. We also conducted cross-validations on small subsamples of papers to purposefully identify and resolve points of contention. Once this iterative process concluded and the taxonomy definitions were firmly established, the team went back and corrected all previously labeled papers to guarantee strict consistency and adherence to the final taxonomy across the entire 378-paper corpus.

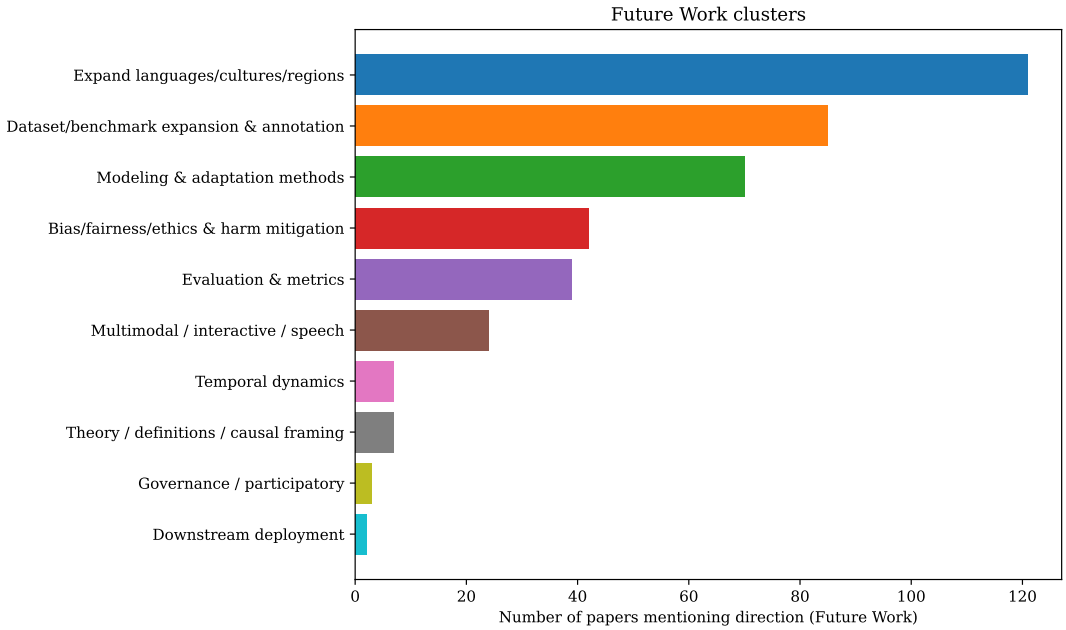


Figure 5: Frequency of future-work clusters in the papers

B Future Work Cluster Analysis

We analyze the Future Work sections recorded in our annotated spreadsheet. The aggregated results are summarized in Figure 5. We identify recurring directions through a transparent and reproducible clustering procedure. Because individual papers often propose multiple extensions, clustering is treated as a multi-label assignment problem. A single future work entry can therefore belong to multiple clusters.

B.1 Cluster schema

We define a small set of coarse themes that repeatedly appear across the surveyed corpus. Each cluster represents a practical research direction rather than a topical category. The clusters and their descriptions are as follows:

- **Expand languages and cultures**: extending coverage to additional languages, dialects, regions, or cultural contexts.
- **Dataset and benchmark expansion**: collecting new datasets, enlarging benchmarks, adding annotations, or improving dataset curation and documentation.
- **Modeling and adaptation methods**: developing improved modeling, training, fine-tuning, alignment, transfer, retrieval, or adaptation approaches that explicitly address cultural phenomena.
- **Bias, fairness, ethics, and harm mitigation**: measuring or mitigating culturally shaped bias, stereotyping, toxicity, or other harmful outputs.
- **Evaluation and metrics**: proposing stronger evaluation metrics, human evaluation protocols, robustness checks, or generalization tests for cultural capabilities.
- **Multimodal or interactive settings**: extending cultural modeling to speech, audio, vision, dialogue systems, or other interactive environments.
- **Temporal dynamics**: modeling culture as a time-varying process, including updates, drift, or longitudinal change.
- **Theory, definitions, and causal framing**: clarifying conceptual definitions of culture or proposing causal or mechanistic interpretations of cultural variables.
- **Governance and participatory approaches**: incorporating community participation, consent mechanisms, stewardship practices, or governance frameworks.
- **Downstream deployment**: evaluating integration into real-world systems and conducting deployment-oriented validation.

B.2 Assignment procedure

Cluster assignments are produced using a dictionary-based matching procedure. For each cluster, we construct a small set of indicative surface forms, such as "multilingual", "dialect", "benchmark", "annotation", "robustness", "toxicity", "speech", or "drift". A cluster is marked as present when any of its indicative patterns appear in the corresponding future work text using case-insensitive matching. This design prioritizes interpretability and reproducibility: cluster definitions are explicit, and assignments can be verified by inspecting the matched expressions. The clustering procedure is intentionally lightweight and surface form-driven. It may undercount papers that describe a direction using uncommon phrasing, and it does not disambiguate cases where a keyword appears with a different meaning. We therefore treat the clusters as a descriptive summary of recurring future work themes rather than as an exhaustive taxonomy.

C Analysis of Methods used for CCs

C.1 Methods across cultural capabilities

Figure 6 presents a 3×3 grid of stacked bar charts comparing method families across cultural capabilities. Each subplot corresponds to one cultural capability. The x-axis lists methodological families, while the y-axis reports the number of contributions. Each bar is partitioned into segments representing four contribution types: System Contribution, Model Evaluation, Data Generation, and System Improvement. Overall, Cultural Knowledge and Value-driven Cultural Alignment contain the largest number of contributions, whereas Cultural Education, Survey-based Cultural Alignment, and Multimodal Cultural Knowledge appear relatively sparse. Across cultural capabilities, prompting-based methods are the most frequently used approach, with training-based methods and lexica, embeddings, or other classical NLP techniques also appearing regularly. Some cultural capabilities display clearer methodological specialization. Cultural Computational Representation relies more heavily on classical NLP-style approaches, while Cultural Translation more often adopts agent-based and prompting-based strategies.

A key observation is that cultural NLP research is methodologically heterogeneous. Different tasks tend to attract different technical approaches rather than converging on a single dominant paradigm. Current work is largely centered on prompting and training-based techniques, particularly for tasks involving cultural knowledge, value alignment, and safety, reflecting the broader influence of large language models in recent NLP research. In contrast, tasks related to cultural representation and sociolinguistic competence maintain stronger connections to earlier NLP traditions. The figure also highlights several relatively underexplored areas, including cultural education, multimodal cultural reasoning, and mechanistic interpretability, indicating opportunities for future work to expand both task coverage and methodological diversity.

C.2 Cultural capabilities across methods

Figure 7 presents the distribution of cultural capabilities across method families using a 2×3 grid of bar charts. Each subplot corresponds to a method family, and the horizontal axis lists cultural capabilities, while the vertical axis reports the number of contributions. The six method families are retrieval and multimodel approaches, human-in-the-loop methods, lexica, embeddings, classical NLP techniques, multimodal approaches, prompting-based methods, and training-based methods. Across methods, prompting-based approaches exhibit the largest overall volume and cover the widest range of cultural capabilities, followed by training-based methods and lexica, embeddings, and classical NLP techniques. In contrast, human-in-the-loop and multimodal approaches account for substantially fewer contributions and are concentrated in a limited set of cultural capabilities. At the capability level, cultural knowledge, value-driven cultural alignment, sociolinguistic competence, and cultural computational representation appear repeatedly across multiple method families. Cultural education and survey-based cultural alignment remain comparatively limited. The stacked bar segments indicate that many methods support multiple contribution types, including system contribution, model evaluation, data generation, and system improvement, rather than focusing on a single contribution category.

The figure further suggests that the field is structured around a small set of widely reusable technical paradigms, particularly prompting-based and training-based approaches, which support a broad range of cultural capabilities. This pattern indicates that much of cultural NLP research adapts general-purpose large language model techniques to culture-related problems rather than develop-

ing highly task-specific architectures. At the same time, method families differ in flexibility. Lexica, embeddings, and classical NLP techniques remain particularly relevant for representational and sociolinguistic analysis, whereas multimodal and human-in-the-loop approaches appear more specialized and comparatively underexplored. Overall, the distribution indicates that research diversity varies not only across cultural capabilities but also across technical methods, with several approaches functioning as central methodological hubs while others remain peripheral.

C.3 Cultural Capabilities Across Dataset Creation Methods

Figure 8 presents an UpSet-style visualization of dataset creation method combinations, stacked by cultural capability. The bottom matrix indicates which dataset creation ingredients appear in each combination pattern. These ingredients include Cultural Experts or Native Speakers, Culture Knowledge Source, Synthetic Generation, Web Data, Expert Annotation, Translation, Crowdsourced Annotation, Common Language Data, Cultural and Social Science Theories, and Culture Value Surveys. The bars above the matrix report the frequency of each ingredient combination, with segments colored by cultural capability category. The bars on the left indicate the marginal frequency of individual ingredients. Cultural Experts or Native Speakers, Culture Knowledge Source, Synthetic Generation, and Web Data appear most frequently, while Cultural and Social Science Theories and Culture Value Surveys occur relatively rarely. At the combination level, the most frequent patterns reach approximately the high teens, while many other combinations appear only four to six times. Across these combinations, Cultural Knowledge contributes the largest share in many of the most common patterns, while Value-driven Cultural Alignment, Sociolinguistic Competence, and Cultural Translation also appear across multiple combinations.

Figure 9 shows a heap map of dataset creation method vs CCs. A key observation is that dataset creation for cultural natural language processing is typically compositional rather than based on a single source. Many studies construct datasets by combining several ingredients, most commonly human expertise from native speakers, external cultural knowledge resources, synthetic data generation, and web-collected data. This indicates that cultural information is rarely operationalized through a single data pipeline. Instead, researchers assemble datasets by integrating human knowledge, structured cultural resources, and scalable generation strategies. The figure also reveals an imbalance in the maturity of different cultural data construction approaches. High-frequency combinations are dominated by Cultural Knowledge-oriented pipelines, suggesting that dataset development has progressed most strongly for knowledge-centered tasks. In contrast, theoretically grounded sources such as cultural value surveys and cultural or social science theories remain uncommon. Overall, the distribution of combinations suggests that current practice prioritizes pragmatic and scalable data construction strategies, while theoretically grounded but less reusable approaches remain underexplored.

D CULTUREMINE

We accompany our survey with a web-interface, called CULTUREMINE, in the hope that it will facilitate research in cultural NLP. CULTUREMINE allows users to filter papers according to our taxonomy, as well as by the cultural groups they study. For example, if a user is interested in Cultural Safety and Harm Reduction in Korean, they can easily apply filters (shown in Figures 10 and 11) to find relevant papers (shown in Figure 12) and click on a paper's title to read it. In addition to providing direct, easy access to papers, CULTUREMINE provides dynamic bar charts at the top of the page, showing the most common Cultural Capabilities, System Methods, and Dataset Creation Methods used in the set of papers that satisfy the active set of filters. It also includes similar charts for the cultural proxies used, languages studied, and regions or countries studied (shown in Figure 13). To keep up with the fast-pace of research in cultural NLP, CULTUREMINE includes a link to a Google Form that allows anyone to submit an annotation for a paper that is not currently covered. Submitted annotations will be reviewed and/or modified manually before being added to ensure that all annotations displayed are of high-quality.

E Cultural Proxies and Groups

We perform a brief analysis of the cultures studied by the papers included in our survey. Figure 14 breaks down the types of cultural proxies used. 72.8% use language as the main proxy for culture, 22.2% use a geographic proxy (Country or Geo-

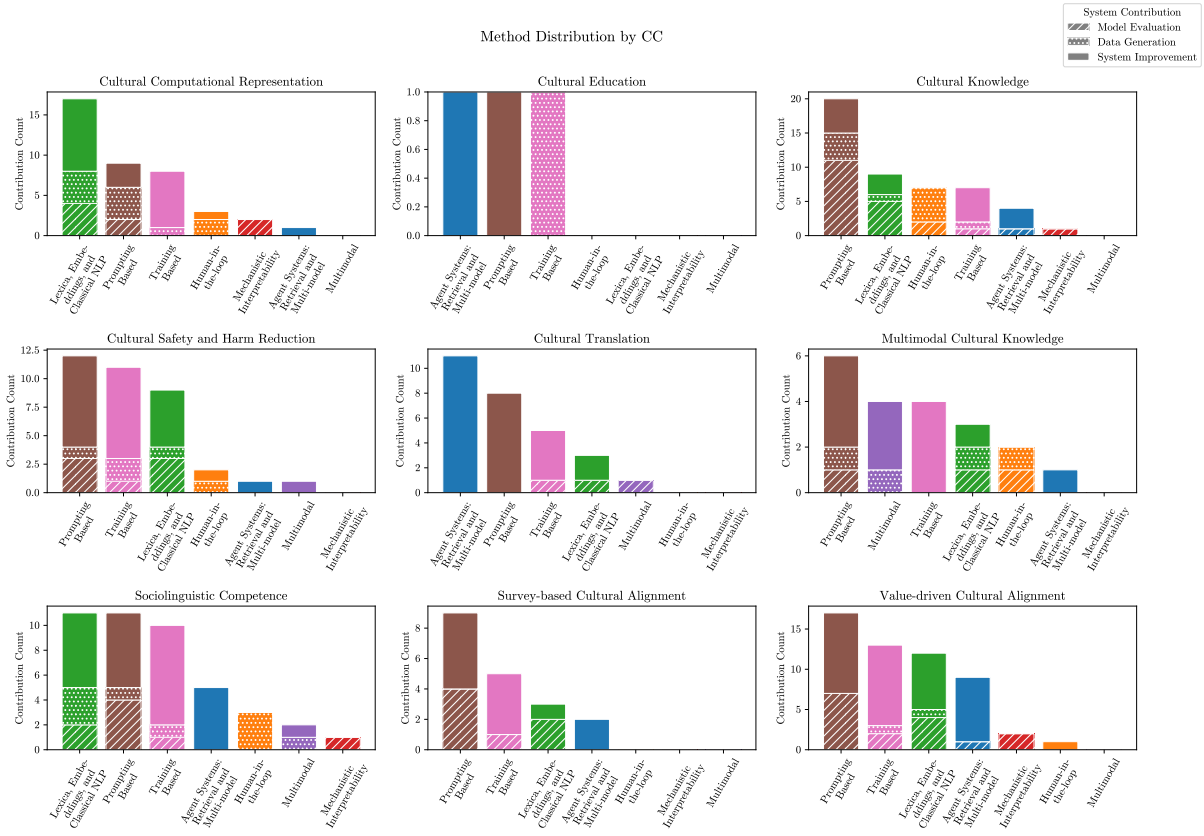


Figure 6: Method distribution across cultural capabilities

graphic Region), and only 5% use some other proxy. As discussed in (Zhou et al., 2025c), these commonly used proxies for culture can be problematic as the groups within them are often composed of multiple cultures. We further analyze the specific cultural groups included in papers that consider at most 10 different groups (235 using language and 50 using geography). Of the papers that use language, there is a very long tail distribution; over half include English and around a third include Chinese, with all but 9 languages—out of 113—being covered less than 20 times (Figure 15). We see a similar trend across 79 total geographic proxies, with around half including US and India, around a third including China, and all but 9 groups covered 5 or fewer times (Figure 16).

F List of Surveyed Papers

For compact presentation, we abbreviate cultural capability categories using acronyms in the following table **??** of surveyed papers. Specifically, CT = Cultural Translation; SCA = Survey-based Cultural Alignment; VCA = Value-driven Cultural Alignment; CK = Cultural Knowledge; MCK = Multimodal Cultural Knowledge; SC = Sociolinguistic Competence; CSHR = Cultural Safety and Harm Reduction; CE = Cultural Education; and CCR = Cultural Computational Representation. These abbreviations are used only for table presentation and correspond directly to the capability categories described in the main text.

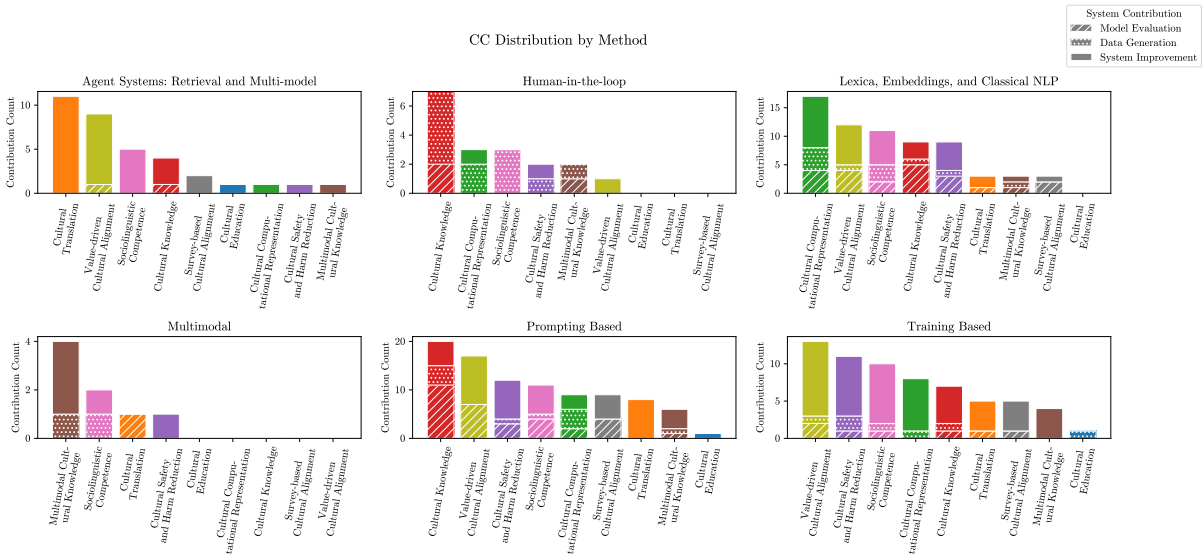


Figure 7: Cultural capability distribution across methods

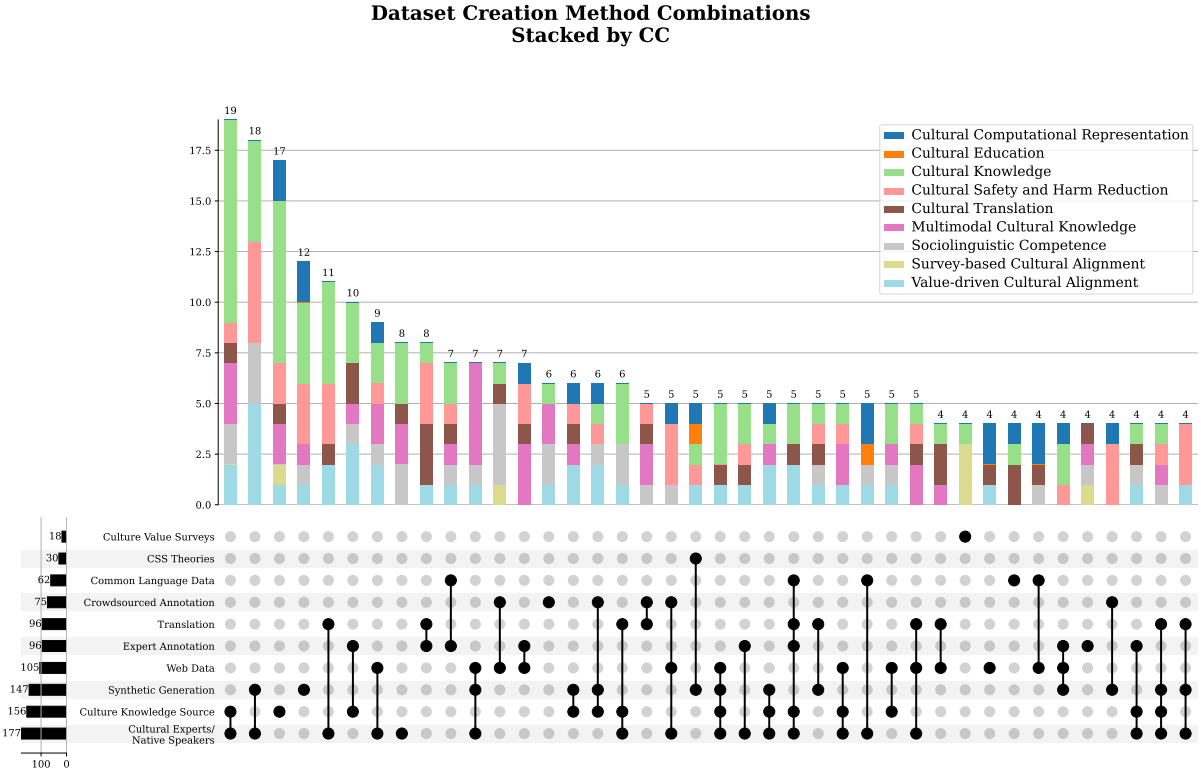


Figure 8: Dataset creation method combinations stacked by cultural capability

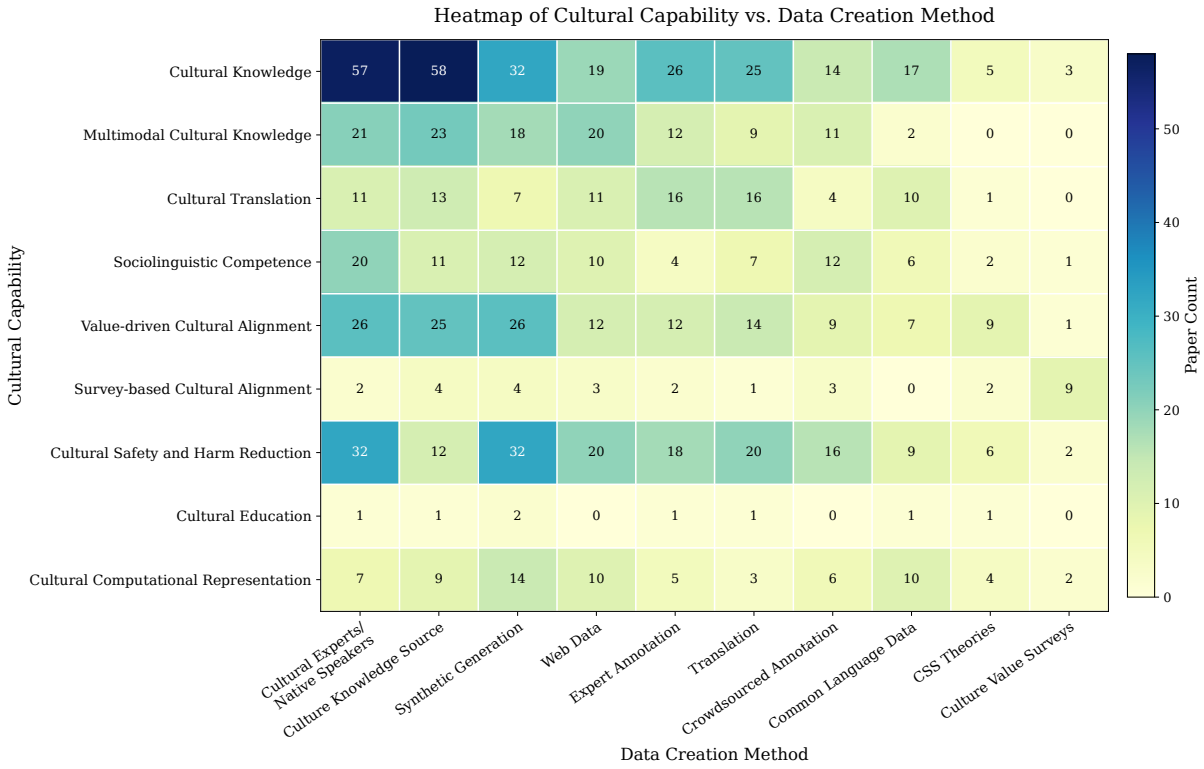


Figure 9: Heatmap of cultural capability by dataset creation method

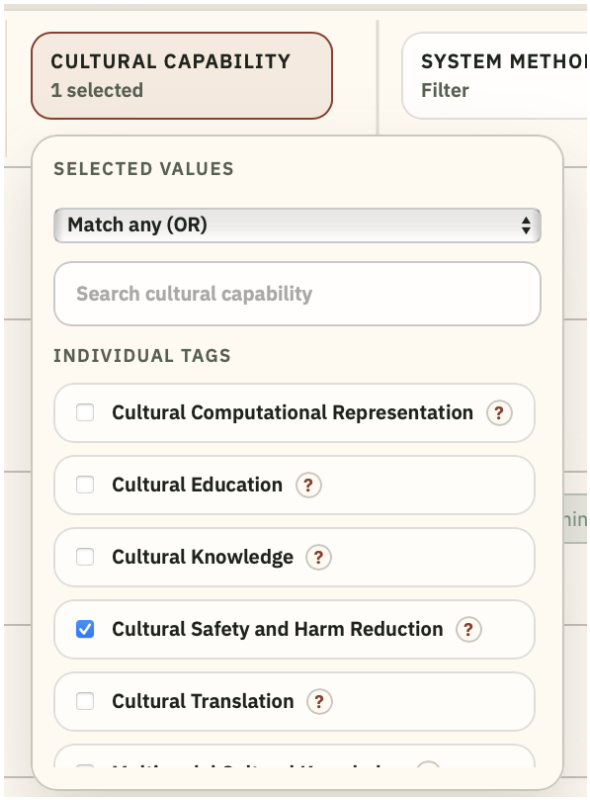


Figure 10: Cultural Capabilities Filter in CULTUREMINE

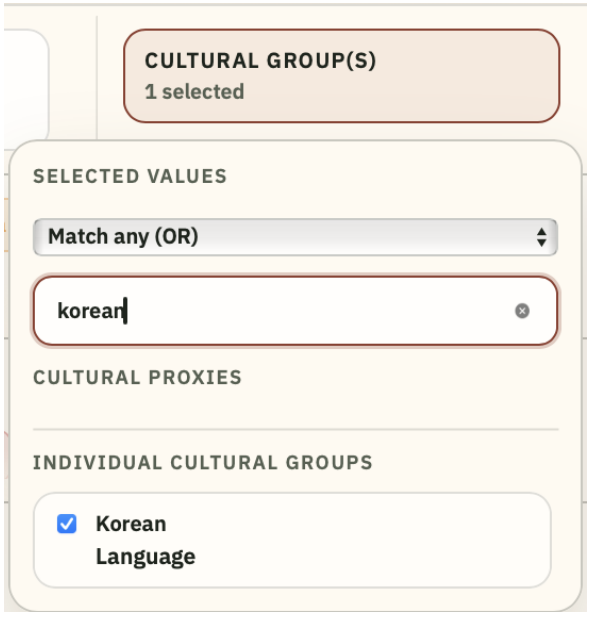


Figure 11: Cultural Groups Filter in CULTUREMINE

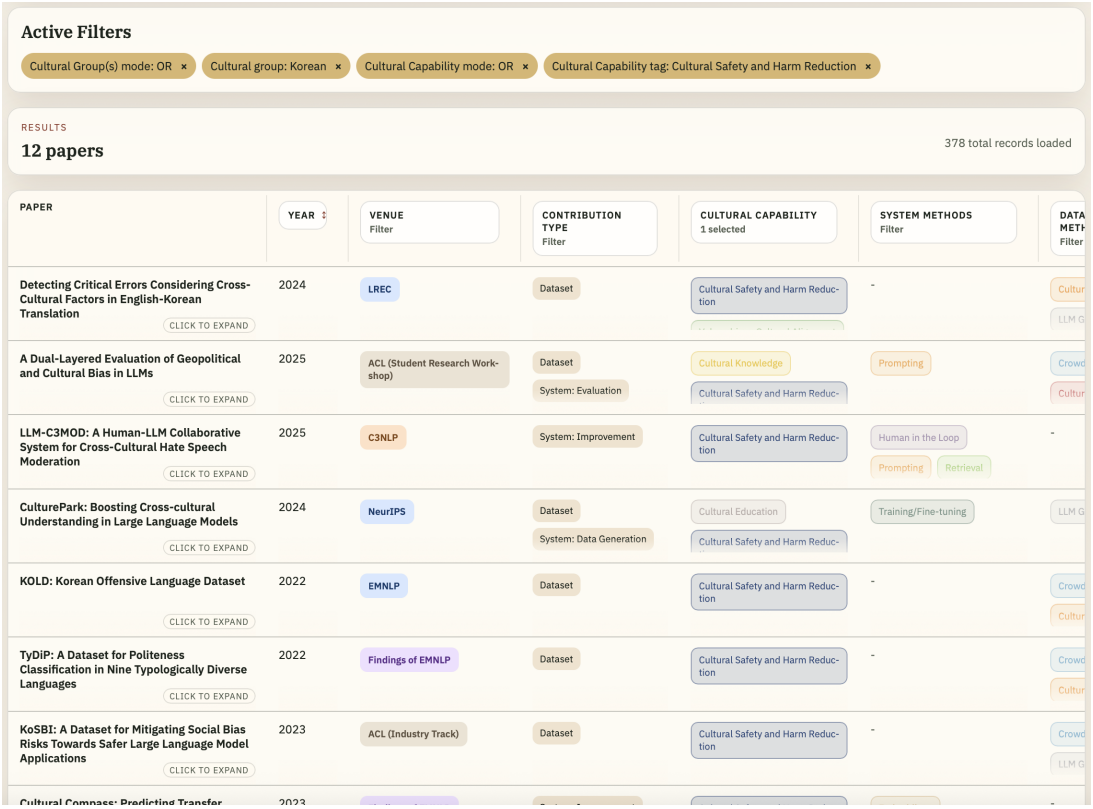


Figure 12: Filtered Paper List in CULTUREMINE

Figure 13: Screenshot of the Plots in CULTUREMINE

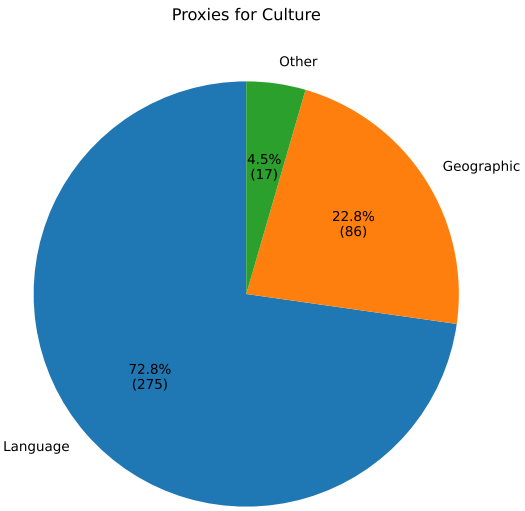


Figure 14: Pie Chart of Cultural Proxies Used

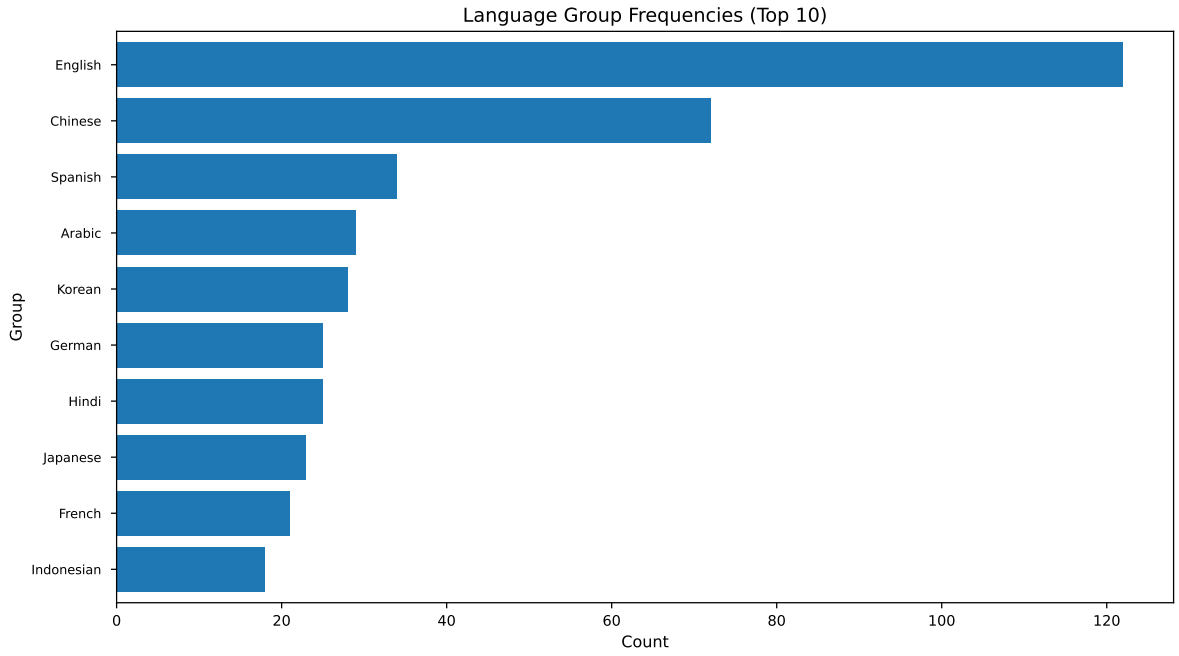


Figure 15: Bar Chart of the Frequencies of the 10 Most Commonly Studied Languages

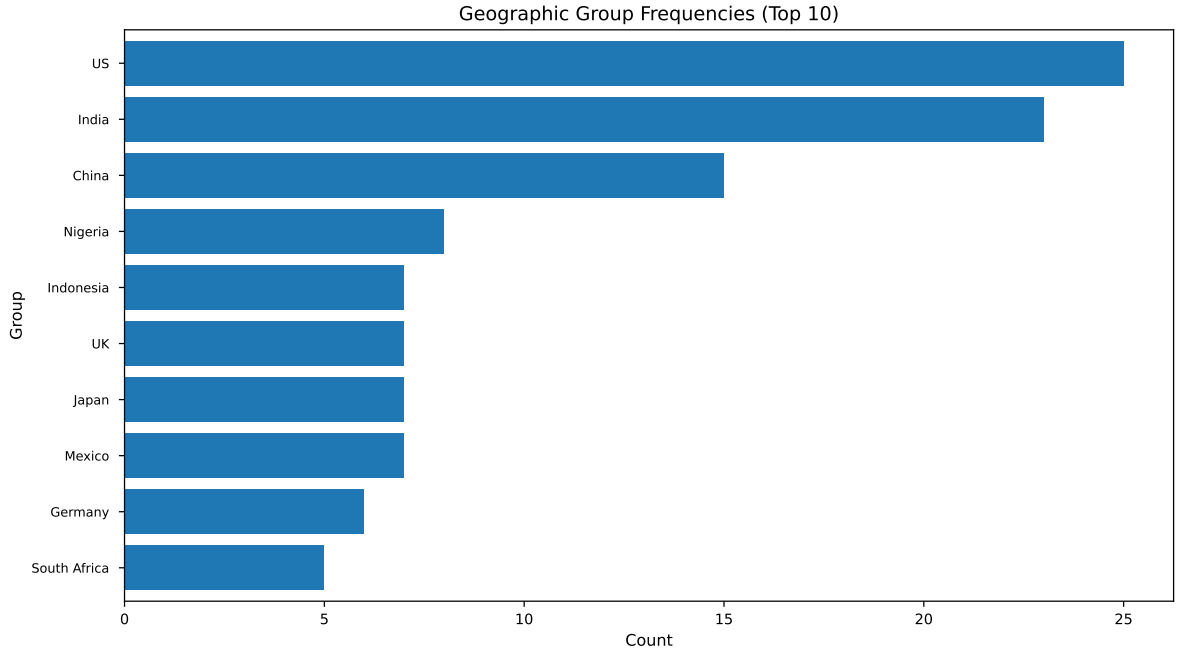


Figure 16: Bar Chart of the Frequencies of the 10 Most Commonly Studied Geographic

Paper	Culture capability
Wang et al. (2024a)	CK, CT, VCA
Singh et al. (2025c)	CK, VCA
Khanuja et al. (2020)	SC
Joshi et al. (2025)	CSHR
Nguyen et al. (2024a)	CCR, SC
Tao et al. (2024)	SCA
Nayak et al. (2024)	MCK
Xu et al. (2024)	VCA, CCR
Arora et al. (2023)	VCA, CCR
Cao et al. (2023)	SCA
Conia et al. (2024)	CT
Sun et al. (2021)	SC
AlKhamissi et al. (2024)	SCA
Masoud et al. (2025)	SCA
Jinnai (2024)	SC
Xu et al. (2025a)	SCA
Ki et al. (2025)	VCA
Sun et al. (2024)	CK
Havaldar et al. (2023b)	VCA
Ahmad et al. (2024)	VCA
Bui et al. (2025b)	CSHR
Bhatia et al. (2024)	MCK
Hale et al. (2025)	VCA, SC
Cecilia Liu et al. (2024)	CK, SC
Schneider and Sitaram (2024)	MCK
Cahyawijaya et al. (2025b)	MCK
Tay et al. (2020)	SCA
Zheng et al. (2022)	VCA
Mohamed et al. (2022)	MCK
Anegundi et al. (2022)	CSHR, CCR
Jha et al. (2023)	CSHR, CCR
Akinade et al. (2023)	CT
Das et al. (2023)	CSHR
Bauer et al. (2023)	VCA
Mukherjee et al. (2023)	CSHR
Shaikh et al. (2023)	SC
Palta and Rudinger (2023)	CK
Hu et al. (2023)	VCA
Choenni et al. (2024)	CCR
Naous et al. (2024)	VCA
Urailertprasert et al. (2024)	MCK
Prieto et al. (2024)	CT
Wang et al. (2024f)	VCA
Zhou et al. (2024)	VCA
Hu et al. (2024)	CT
Khanuja et al. (2024)	MCK
Wuraola et al. (2024)	SC
Li et al. (2024e)	MCK
Putri et al. (2024)	CK
Mohamed et al. (2024)	MCK
Giuliani et al. (2024)	CK
Li et al. (2024d)	VCA
Cao et al. (2024)	VCA
Zhan et al. (2024)	VCA, SC
Hsieh et al. (2024)	CSHR
White et al. (2024)	SC
Yao et al. (2024a)	CT
Bhatt and Diaz (2024)	VCA

Table continues

Paper	Culture capability
Cadotte et al. (2024)	CT
Kim et al. (2024a)	CK
Eo et al. (2024)	CSHR, VCA
Zhao et al. (2024a)	SCA, CK
Fort et al. (2024)	CSHR
Havaldar et al. (2024)	VCA
Lee et al. (2024b)	CSHR
Abdelkadir et al. (2024)	SC
Wang et al. (2024d)	CK
Haberland et al. (2024)	CT
Tonneau et al. (2024)	CSHR
Yakhni and Chehab (2025)	CT
Arora et al. (2025)	CK
Belay et al. (2025)	CK
Park et al. (2025a)	MCK
Chiu et al. (2025)	CK
Yang et al. (2025b)	SC
Havaldar et al. (2025)	CT
Wu et al. (2025)	CK
Zhang et al. (2025a)	CK
Kim and Kim (2025)	CSHR, CK
Ignat et al. (2025)	SC
Park et al. (2025b)	CSHR
Rai et al. (2025a)	SC
Kim et al. (2025b)	CSHR
Kim et al. (2025a)	VCA
Mousi et al. (2025)	CK, SC
Pandey et al. (2025)	CSHR
Cheng et al. (2025)	CK
Umbet et al. (2025)	CK
Qiu et al. (2025)	VCA
Bhatia et al. (2025)	CCR
Yadav et al. (2025)	SCA
Vasilev et al. (2025)	MCK
Li et al. (2025)	VCA, CK
Maji et al. (2025a)	CK
Singh et al. (2025a)	CT
Schneider et al. (2025)	CK, MCK
Hasan et al. (2025)	CK
Kim and Lee (2025)	CK
Zhang et al. (2025f)	MCK
Ghaboura et al. (2025)	MCK
Reshetnikov and Marinescu (2025)	MCK
Anik et al. (2025)	CT
Onohara et al. (2025)	CK, MCK
Seveso et al. (2025)	CK
Bai et al. (2025)	CK
Winata et al. (2025)	MCK
Naous and Xu (2025)	CK
Saha et al. (2025)	SC
Berger and Ponti (2025)	MCK
Jeong et al. (2025)	MCK
Arnardóttir et al. (2025)	CK
Ventura et al. (2025)	CK, MCK
Paniv et al. (2025)	MCK
Ringel et al. (2019)	SC
Lin et al. (2018)	SC
Gutiérrez et al. (2016)	VCA, SC

Table continues

Paper	Culture capability
Sheng et al. (2016)	MCK
Wilson et al. (2016)	SC, VCA
Friedman et al. (2019)	CSHR, CCR
Maji et al. (2025b)	MCK
Cuevas et al. (2025)	CSHR
Sahoo et al. (2025)	CK
Satar et al. (2025)	MCK
Zhang et al. (2025d)	CK, VCA
Limkonchotiwat et al. (2025)	CK
Cao et al. (2025)	SCA
Jiang et al. (2025)	SCA
Ramezani and Xu (2023)	VCA
Cahyawijaya et al. (2025a)	CCR
Yin et al. (2024)	CSHR
Li et al. (2024c)	CE, CSHR, VCA
Ziems et al. (2025)	VCA, CK
Epure et al. (2020)	CCR
Hahm et al. (2020)	SC
Yin et al. (2021)	MCK
Milbauer et al. (2021)	CCR
Liu et al. (2021)	MCK
Ghosh et al. (2021)	CSHR
Kiesel et al. (2022)	SCA, VCA
Kim et al. (2022)	CK
Yin et al. (2022)	CK
Thai et al. (2022)	CT
Jeong et al. (2022)	CSHR
Maronikolakis et al. (2022)	CSHR
Nadejde et al. (2022)	CT, SC
Sharma et al. (2022)	CSHR
Srinivasan and Choi (2022)	CSHR
Herrera et al. (2022)	SC
Chandran Nair et al. (2022)	CCR
Mortensen et al. (2022)	CCR
Deshpande et al. (2022)	CSHR
Ziems et al. (2023)	CCR
Lee et al. (2023)	CSHR
Alshahrani et al. (2023)	CCR
Li and Zhang (2023)	MCK
CH-Wang et al. (2023)	CCR
Havaldar et al. (2023a)	CCR
Lahoti et al. (2023)	CK, VCA
Koto et al. (2023)	CK
Fung et al. (2023)	CCR
Keleg and Magdy (2023)	CCR
Kabra et al. (2023)	CK
Huang and Yang (2023)	CCR
Zhou et al. (2023)	CSHR
Bang et al. (2023)	VCA, CSHR
Wang et al. (2024c)	SCA
Zhang et al. (2024a)	CCR
Alwajih et al. (2024)	MCK
Chen et al. (2024)	CT
Casola et al. (2024)	CSHR, SC
Nguyen et al. (2024b)	CT, CSHR
Falk et al. (2024)	CCR
Zhao et al. (2024b)	CT
Pham et al. (2024)	CCR

Table continues

Paper	Culture capability
Feng et al. (2024a)	CK
Feng et al. (2024b)	SCA, VCA, CCR
Lovenia et al. (2024)	CT, CK, MCK, SC, CSHR, VCA
Watts et al. (2024)	CK
Acquaye et al. (2024)	CK
Aakanksha et al. (2024)	CSHR
Hobson et al. (2024)	VCA
Davani et al. (2024)	CSHR
Dammu et al. (2024)	CSHR, SC
Deas et al. (2024)	CCR
Wu et al. (2024)	CT
Yuan et al. (2024)	CK
Javed et al. (2024)	CCR
Lee et al. (2024a)	CK, SCA
Yu et al. (2024)	CK, VCA
Li et al. (2024f)	CK
Alyafeai et al. (2024)	CK
Kim et al. (2024b)	VCA
Wei et al. (2024)	CK
Plaza-del Arco et al. (2024)	CCR
Liu et al. (2024a)	VCA
Shi et al. (2024)	CK, CCR
Yüksel et al. (2024)	CK
Masala et al. (2024)	CK
Huang and Xiong (2024)	CSHR
Ullah et al. (2024)	CSHR
Seth et al. (2024)	CK
Agarwal et al. (2024)	VCA
Tonja et al. (2024)	CSHR, CT
Yarlott et al. (2024)	CK
Son et al. (2024)	CK
Benkler et al. (2024)	VCA
Grigoreva et al. (2024)	CSHR
Maronikolakis et al. (2024)	CSHR
Lou et al. (2024)	CT
Prabhakaran et al. (2024)	CSHR
España-Bonet and Barrón-Cedeño (2024)	CCR
Shen et al. (2024)	CK
Wang et al. (2024e)	CK
Mukherjee et al. (2024)	CK
Huang et al. (2024)	CK, VCA
Yao et al. (2024b)	VCA
Acharya et al. (2024)	CK
Liu et al. (2024b)	SC, CE
Piccirilli et al. (2024)	SC
Koto et al. (2024)	CK
Romanyshyn et al. (2024)	CK
Kiulian et al. (2024)	CK
España-Bonet et al. (2024)	CCR
Li et al. (2024a)	CCR, SC
Zhang et al. (2024b)	CT
Anastasi et al. (2024)	CSHR
Ng et al. (2024)	CSHR
Liu et al. (2025b)	VCA
Kumar and Jurgens (2025)	VCA
Sadallah et al. (2025)	CK

Table continues

Paper	Culture capability
Fang et al. (2025)	CK
Yu et al. (2025)	SC
Liu et al. (2025e)	MCK, CK
Zhang et al. (2025b)	MCK
Togmanov et al. (2025)	CK
Yari and Koto (2025)	CK
Bui et al. (2025a)	CK
Menis Mastromichalakis et al. (2025)	CSHR
Ying et al. (2025)	CK, SC
Feng et al. (2025)	CK
Isbarov et al. (2025)	VCA, CK
Shetty et al. (2025)	VCA
Zeng et al. (2025)	VCA, CK
Yerukola et al. (2025)	CSHR
Karamolegkou et al. (2025)	MCK
Bhatt et al. (2025)	SC
Farhansyah et al. (2025)	SC, CT
Liang et al. (2025)	CT
Chen et al. (2025d)	MCK
Nawale et al. (2025)	CSHR
Bayramli et al. (2025)	MCK
Montalan et al. (2025)	CK, CSHR
Grandury et al. (2025)	VCA, CK
Alwajih et al. (2025a)	CK, SC
Olaleye et al. (2025)	SC
Ishita and Mamidi (2025)	CT
Alhassoun et al. (2025)	CK, VCA
Bouamor et al. (2025)	SC
Alwajih et al. (2025b)	CK
Krsteski et al. (2025)	VCA
Kim and Johnson (2025)	CSHR
Pokharel and Agrawal (2025)	SC
Altammami (2025)	CT
Yang et al. (2025a)	CT
Shiono et al. (2025)	MCK
Kabir et al. (2025)	SCA
Rooein et al. (2025)	CSHR, CCR
Maji et al. (2025c)	MCK
Xuan et al. (2025)	SC
Sadr et al. (2025)	VCA
Fu et al. (2025)	CK
Shen et al. (2025)	VCA
Wang et al. (2025b)	CK, VCA
Azmi et al. (2025)	CSHR
El Mekki et al. (2025)	CT, VCA, CK
Hu et al. (2025)	CSHR
Ramezani and Xu (2025)	CK
Tanwar et al. (2025)	CK
Singh et al. (2025b)	CK, MCK
Hashmat et al. (2025)	CSHR
Liu et al. (2025d)	SC
Xie et al. (2025)	CK
Yamamoto et al. (2025)	CSHR
Aji and Cohn (2025)	CK, CT
Mia et al. (2025)	MCK, CSHR
Chen et al. (2025c)	CK
Liu et al. (2025c)	SCA
Xu et al. (2025c)	MCK

Table continues

Paper	Culture capability
Mukherjee et al. (2025)	VCA
Zhang et al. (2025c)	MCK
Shafique et al. (2025)	MCK
Al Ghallabi et al. (2025)	CK
Parappan and Henao (2025)	CSHR
Lavrouk et al. (2025)	MCK, CK
Calvo-Bartolomé et al. (2025)	CK
Chen et al. (2025b)	CT
Seweryn et al. (2025)	CK, CSHR, VCA
Ma et al. (2025a)	CSHR
Nyandwi et al. (2025)	MCK
Dai et al. (2025)	SCA
Zhou et al. (2025b)	MCK
Yang et al. (2025c)	CSHR
Mitran et al. (2025)	SCA
Roh et al. (2025)	CK
Zhang et al. (2025e)	CT
Sitaram et al. (2025)	CSHR
Zhong et al. (2025)	VCA
Nimo et al. (2025)	CK, CSHR
Guo et al. (2025)	CK
Vasselli et al. (2025)	SC, CT
Pramodya et al. (2025)	CK
Chen et al. (2025a)	CT
Hong et al. (2025)	SC, VCA
Hsieh et al. (2025)	MCK
R et al. (2025)	SC
Hosseinbeigi et al. (2025a)	CK
Korre et al. (2025)	CSHR
Saffari et al. (2025)	VCA, CSHR
Tapo et al. (2025b)	CE
Wang et al. (2025a)	CT
Lan et al. (2025)	CSHR
Dwivedi et al. (2025)	VCA
Sasu et al. (2025)	SC
Susanto et al. (2025)	CK, VCA, SC
Tsutsumi and Jinnai (2025)	CK
Hosseinbeigi et al. (2025b)	VCA, CK, CT
He et al. (2025)	CK
Das et al. (2025)	MCK, CSHR
Nahin et al. (2025)	CK
Bi et al. (2025)	MCK
Zou et al. (2025)	CT
Wang et al. (2025c)	CCR, CK
Tan et al. (2025)	CT
Zhang et al. (2025h)	SC
Mor-Lan et al. (2025)	CCR
Mubarak et al. (2025)	CSHR
Zhang et al. (2025g)	CK
Chae et al. (2025)	CK, VCA, CSHR
Ma et al. (2025b)	SC
Trager et al. (2025)	CSHR
Gupta et al. (2025)	SC
Dey et al. (2025)	SCA
Nayak et al. (2025)	MCK
Villa-Cueva et al. (2025)	CT
Wibowo et al. (2024)	CK
Alwajih et al. (2025c)	MCK
Zhao et al. (2025)	CK

Table continues

Paper	Culture capability
Malik et al. (2025)	SC, VCA
Caplan et al. (2025)	SC
Mohammadi et al. (2025)	SCA
Bennie et al. (2025)	CSHR
Rachamalla et al. (2025)	CK, CSHR, SC
Liao et al. (2025)	CT
Pujari and Goldwasser (2025)	VCA, CCR
Muhammad et al. (2025)	CSHR
Madhusudan et al. (2025)	CCR
Leteno et al. (2025)	VCA
Son et al. (2025)	CK
Ashraf et al. (2025)	CSHR
Khanuja et al. (2025)	MCK
Banerjee et al. (2025)	CSHR
Miehling et al. (2025)	VCA
Nikandrou et al. (2025)	MCK
Pham et al. (2025)	VCA, CSHR
Zhou et al. (2025a)	CK
Verma et al. (2025)	CK
Rai et al. (2025b)	VCA
Mitchell et al. (2025)	CSHR
Tapo et al. (2025a)	CT
Magdy et al. (2025)	CK
Moosavi Monazzah et al. (2025)	CK
AbuHaija et al. (2025)	CCR
Völker et al. (2025)	CT
Conia et al. (2025)	CT
Park et al. (2025c)	VCA
Srirag et al. (2025)	SC
Sahoo et al. (2024)	CSHR
Özbal et al. (2016)	VCA, SC

Towards More Transparent Online Campaigning: Detecting Political Campaign Content in Election-related Social Media Posts

Abdullah Alabdullah[1], Conor Gaughan[2], Thomas Flavel[3], Shubhanjay Varma[2], Rachel Gibson[2], Marta Cantijoch Cunill[2], Alexandru Cernat[2] and **Riza Batista-Navarro[4]**
[1]School of Informatics, University of Edinburgh, United Kingdom
[2]School of Social Sciences, University of Manchester, United Kingdom
[3]School of Arts, Languages and Cultures, University of Manchester, United Kingdom
[4]Department of Computer Science, University of Manchester, United Kingdom
Correspondence: riza.batista@manchester.ac.uk

Abstract

A large part of political campaigns during elections is now being conducted online, with political actors leveraging their networks on social media platforms. To maintain transparency in political communications, regulations applicable to online campaigning have been put in place in many democracies. While it should be straightforward for voters to determine who produced and funded online advertisements comprising paid political campaigns, it is much more challenging to detect if organic content, i.e., social media posts, pertains to political campaigning, due to possibly subtle yet suggestive language that can be used by certain actors. In this paper, we investigate the feasibility of automatically detecting whether a given tweet posted by a political actor pertains to political campaigning, and if yes, whether it was conveyed in a direct or indirect (subtle) manner. After establishing an annotation scheme for the task of detecting political campaign content in tweets, we fine-tuned three encoder models (BERT, BERTweet and PoliBERTweet) for the same task and evaluated their performance. Our results show that fine-tuning BERTweet leads to the best macro-averaged F1-score (0.776), although all models consistently struggle to detect indirect campaigning.

1 Introduction

Our society is now entering a "fourth era" of political campaigning, defined as a data-driven, digital-first approach using hyper-personalised micro-targeting and networked communication via social media (Magin et al., 2017; Römmele and Gibson, 2020; Strömbäck, 2008). One platform which has been at the forefront of online political campaigning since 2010 is Twitter (now X), enabling political actors to directly communicate with the electorate (Vergeer, 2015). Across numerous national elections, major party candidates have strategically employed Twitter to disseminate political messages and influence public opinion (Jungherr, 2016).

On the one hand, these political messages can be overtly campaigning (e.g. *"Get out and vote Democrat on November 3rd!"* or *"If he's elected, Corbyn will destroy this country"*). On the other hand, many of them can be more subtle, not directly telling the reader to vote one way or the other but indirectly suggesting it via the language that they employ (e.g. *"Climate change is the most pertinent threat to our survival"* or *"We need an economy that works for the many, not just a wealthy few"*).

Automated detection of online political campaigning is crucial not only for academic research, but also for effective regulation. Electoral law in many countries now requires complete transparency around online campaigning by political parties and politicians, including registry requirements and the use of digital imprints. Examples of these include the 2022 UK Elections Act 2022, the 2025 EU Political Advertising Regulation, Section 325 of the Canada Elections Act, and U.S. Federal Election Commission (FEC) disclaimer rules governing paid online political advertisements on social media and digital platforms. These are applied most strictly to paid campaign advertising but can also apply to many other types of digital material including organic content: tweets and other types of social media posts. Identification of text-based online campaigning is not necessarily straightforward, especially when the messaging is implicit rather than explicit.

With the overarching aim of boosting transparency and trust in political messaging on social media, our work seeks to enable the detection of political campaign content in online posts at scale, by developing new natural language processing (NLP) models for automatically classifying text according to whether it pertains to political campaigning or not. Our contributions include:

- A conceptual framework for capturing political campaign content in election-related

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 249–270

tweets, underpinned by an annotation decision tree and guidelines; we report the results of applying this framework on the annotation of tweets by humans and a large language model;

- The development of three baseline transformer-based encoder models fine-tuned for the task of detecting political campaign content; the performance and comparison of these models were systematically evaluated and compared, leading to the selection of the best-performing baseline model which was then applied at scale to US 2020 election tweets.

2 Online Political Campaign Content

Given that this work evaluates models on a US dataset, it would be intuitive to derive our definition of online political campaign content from US online campaigning laws. However, despite similar transparency principles around paid online advertising being conveyed through the Federal Election Commission (FEC) and Federal Election Campaign Act (FECA) disclaimer rules (Fowler et al., 2020), the US framework remains less centralised and less comprehensive than the UK or EU regimes.

Therefore, we construct our framework around statutory guidance laid out by the UK Electoral Commission in accordance with Section 54 of the UK Elections Act 2022 (Electoral Commission, 2026). The benefit of using this framework is that it not only applies to paid digital material, but also to "organic" material that has not been paid for but is political and published by a relevant entity such as a registered party or candidate.

As such, we adopt the following working definition for *online political campaign content* – any digital material posted by political parties or candidates which can be reasonably regarded to influence the public to give support to or withhold support from:

(1) one or more political parties
(2) a candidate or future candidate
(3) an elected office holder
(4) political parties, candidates, future candidates or elected office-holders that are linked by their support for, or opposition to, particular policies, or by holding particular opinions
(5) other categories of candidates, future candidates or elected office-holders that are not based on policies or opinions, e.g., candidates who went to a state school, or Members of Parliament (MPs) who grew up in their constituency.

Here, we distinguish between two forms of campaigning: direct and indirect. Where a party or candidate encourages or discourages support for another political party, candidate or elected office holder (points 1-3 above) by directly mentioning them, we refer to this as **direct campaigning**. Where a party or candidate encourages or discourages support for another political party, candidate or elected office holder through reference to linked policies, opinions or characteristics (points 4-5), we refer to this as **indirect campaigning**. Where content cannot be reasonably regarded to influence the public to support or withhold support from another political party, candidate or elected office holder, irrespective of whether they are directly named or not, we refer to this as **non-campaigning**.

Distinguishing between these three categories is important for flagging digital content posted by political actors that are attempting, either directly or indirectly, to influence the public. Political actors have both a moral and, in many countries, legal obligation to be transparent about campaigning, and this might not always be obvious to voters.

3 Related Work

Computational models for enhancing transparency and accountability in online political communication have been proposed in the past, although majority were concerned with paid advertisements. Sosnovik et al. (2023) collected political advertisements from Facebook during the French election period in 2022 and categorised them according to policy categories using a classification model, while Yoshikawa and Roesner (2025) manually analysed political advertisements shown on news and media websites during the 2024 US elections. To the best of our knowledge, the only work that set out to address the task of automatically detecting political campaign content in organic content (i.e., posts) in social media platforms is that of Achmann-Denkler et al. (2024), which analysed Instagram captions and stories posted during the German elections in 2021, according to whether they contain a "Call to Action" (CTA) – a message that mobilises readers to take specific actions.

4 Task Formulation

In this work, we address the automated detection of political campaign content in tweets authored by political figures during elections. We cast this problem as a text classification task. That is, given a tweet that is known to have come from a political actor, an annotator should label it with one of the three classes in our annotation scheme, namely, `non-campaigning`, `direct_campaigning`, and `indirect_campaigning`, depending on whether it contains campaign content, and if yes, the type of campaigning used.

The next section describes how we constructed a dataset consisting of 27,620 English-language tweets posted by 75 political figures relevant to the 2020 US presidential election. This dataset supports us in our objective to build robust classifiers for campaign content detection, and to systematically assess how domain adaptation, first to Twitter language and then to political and electoral discourse, affects the detection of political campaigning content.

5 Dataset Construction

In this section, we present the steps that we carried out in order to develop a reliable dataset for training and evaluating our text classification models for detecting political campaign content.

5.1 Collection of Social Media Posts

This study uses an in-house dataset of social media posts collected during the 2020 US presidential election campaign period, that was constructed as part of the Digital Campaigning and Electoral Democracy (DiCED) project.[1] It consists of 196,012 tweets from 75 verified political figures, including presidential and vice-presidential candidates and Democratic and Republican senators. For a full list of all political figures involved, see Appendix F.

This dataset was chosen for its focus on official actors, which reduces ambiguity in communicative intent. It thus captures campaign-related discourse more clearly than other existing public datasets (e.g., #Election2020; Chen et al., 2021), which primarily reflect general political discussion rather than campaign communication.

To capture peak campaign activity, we carried out data filtering and included only tweets posted within six weeks before and in the week of the election day (the 3rd of November 2020). After removing predominantly Spanish tweets, our final dataset consists of 27,620 tweets. We consider this dataset to be sufficient in terms of size, given that it is larger than those utilised in prior studies focused on fine-tuning BERT-based encoder models for political tweet classification, such as the work by Grimminger and Klinger, 2021 (3,000 tweets) and Baran et al., 2022 (6,112 tweets). We split the dataset into training (21,984; ~80%), development (2,745; ~10%), and test (2,750; ~10%) sets.

5.2 Annotation Protocol Development

The central challenge of this study is to operationalise the concept of political campaigning that we defined in Section 2 into a multi-level labelling scheme that can be applied consistently by both human annotators and state-of-the-art large language models (LLMs) to detect campaigning intent in tweets. This task is non-trivial, as political tweets are typically short, noisy and often express persuasive intent implicitly rather than through explicit messaging (Vijayaraghavan et al., 2021). To address this, we employ a structured annotation protocol underpinned by a decision tree (DT) and detailed annotation guidelines (see Appendix A).

Our hierarchical decision tree, depicted in Figure 1, decomposes the annotation task into a sequence of binary decisions reflecting increasing levels of interpretive judgment. The DT represents a logical process that a human annotator follows to classify tweets into their respective campaigning classes, rather than a learned DT. Tweets are first evaluated for direct references to political entities (e.g. parties or candidates). If such a reference exists, the annotator determines whether the tweet encourages or discourages support for that entity, resulting in a classification of `direct_campaigning` (`dir_camp`) or `non-campaigning` (`non-camp`). If no political entity is referenced, the tree assesses whether the tweet refers to a related characteristic, such as a policy, ideology, opinion or attribute of a political figure. Tweets that reference such characteristics and express evaluative or persuasive sentiment are classified as `indirect_campaigning` (`ind_camp`). This structure ensures that labels are mutually exclusive and exhaustive while maintaining a clear distinction between direct and indirect campaigning by separating the assessment of entity reference and the discussion of linked characteristics.

[1] https://sites.manchester.ac.uk/diced/

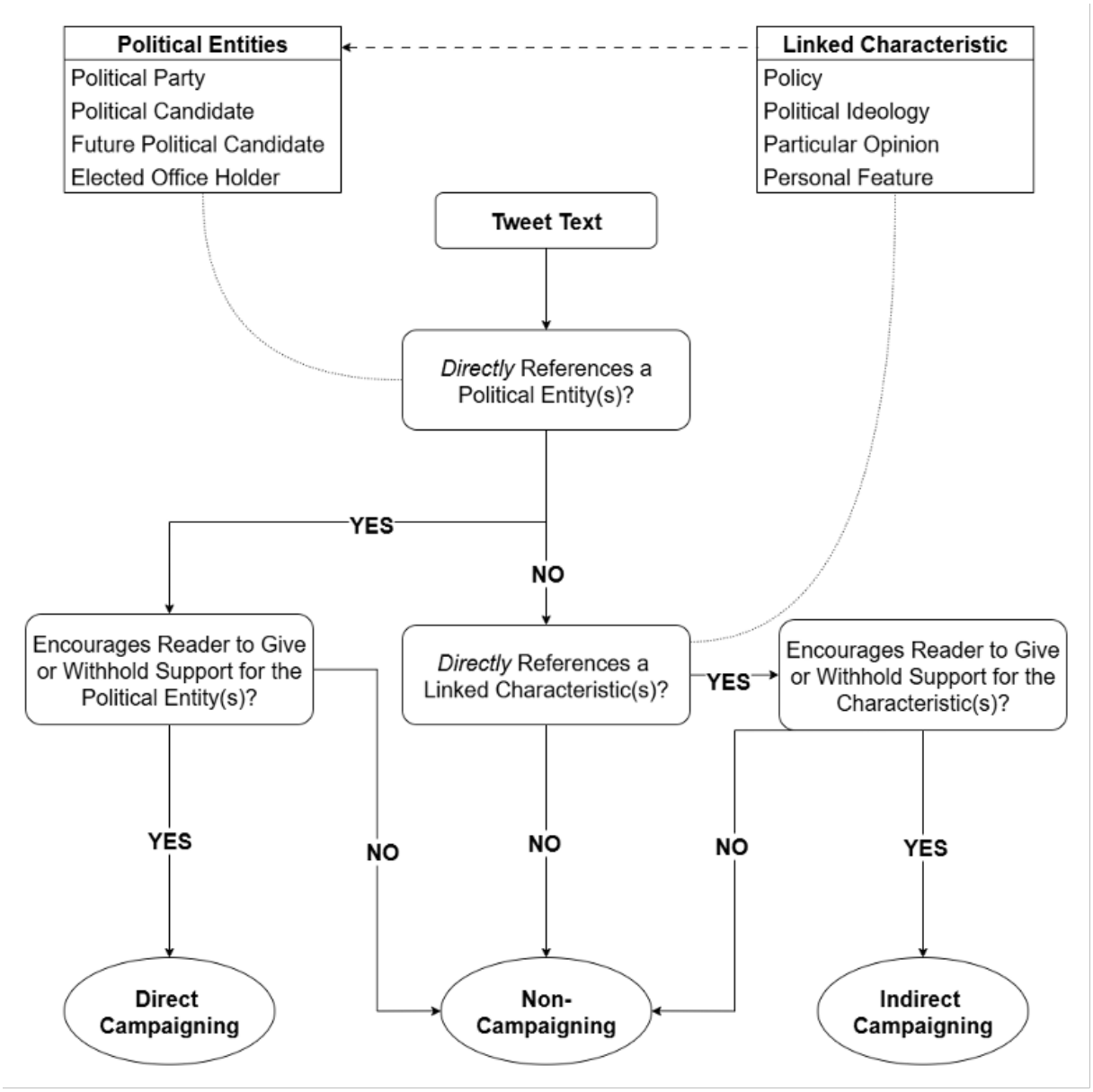


Figure 1: The decision tree guiding the classification of any given tweet according to our annotation scheme.

Our annotation guidelines complement the decision tree by formalising each binary decision with explicit definitions, decision rules and boundary conditions for common sources of ambiguity, such as self-referential language. They specify how political entities and related characteristics are identified within the text of a given tweet and how evaluative language is distinguished from neutral description when assessing persuasive intent.

5.3 Data Annotation

To validate our annotation protocol prior to large-scale labelling using an LLM, we conducted a two-round pilot exercise to assess inter-annotator agreement (IAA) and identify sources of ambiguity. Given the cost and irreversibility of full-scale annotation, this step ensured that the decision tree and guidelines yielded consistent and reliable labels across annotators.

A random sample of 200 tweets was selected for the pilot. In each round, two political science specialists independently annotated 100 tweets following a brief training session with worked examples. Beyond assigning labels, annotators provided structured feedback on the clarity of the DT, instances of hesitation or backtracking, and ambiguities. After each round, the protocol was refined based on class-level agreement patterns and annotator feedback. After the final round, the standard (unweighted) Cohen's κ IAA score between the two human anno-

tators reached 0.57, indicating moderate agreement (Landis and Koch, 1977).

In addition to human annotation, we employed an LLM as a third expert annotator to enable scalable dataset labelling. To ensure consistency with human annotators, the decision tree and annotation guidelines were adapted into a structured prompting scheme, with each decision node represented as a separate prompt. For reproducibility, we provide all prompts in Appendix C.

Among several locally hosted models evaluated, `Llama-3.3-70B-Instruct` achieved the best alignment with human annotations (on the same 100 randomly sampled tweets), attaining an overall accuracy (matching labels) of 0.67 and a Cohen's κ of 0.50.

The agreement between the LLM and one human annotator (κ: 0.50) was comparable to that between the two human annotators (κ: 0.57). Our human–human IAA is lower than that reported by (Grimminger and Klinger, 2021) (human-human Cohen's κ: 0.61–0.88 for stance detection on 2020 US election tweets). It is, however, worth noting that stance detection relies on explicit evaluation towards named candidates, yielding clearer class boundaries, whereas our task in some cases requires inferring persuasive intent without direct political references, making it inherently more subjective. These results indicate that our LLM-generated annotations are sufficiently reliable for large-scale training set labelling.

The final annotation strategy combined human and LLM annotations: The training set was fully annotated by the LLM; the validation set was annotated by one human annotator; and the test set was annotated by two humans with LLM annotations only used for majority vote in cases where the two human annotators disagree.

6 Methodology

6.1 Models and Evaluation Strategy

We fine-tune and evaluate three transformer-based encoders: BERT (Devlin et al., 2019), BERTweet (Nguyen et al., 2020) and PoliBERTweet (Kawintiranon and Singh, 2022). This allowed us to examine how progressive domain adaptation, from general English to Twitter, and then to electoral discourse, affects model performance on the task of campaign content detection.

BERT (110M parameters), which follows the BERT-base architecture and is pre-trained on large-scale English corpora, serves as our general-purpose baseline, though its pre-training data does not reflect the stylistic features of Twitter or political discourse. BERTweet retains the BERT-base architecture but follows the RoBERTa pre-training procedure and is further trained on 850M English tweets, effectively modelling the informal syntax, abbreviations, and platform-specific conventions of Twitter. BERTweet was shown to outperform general-purpose models on tasks such as sentiment analysis of tweets (Nguyen et al., 2020). Meanwhile, PoliBERTweet further pre-trains BERT-base on 83M tweets from the US 2020 presidential election, incorporating election-specific vocabulary and campaigning rhetoric. It has shown strong performance on politics-focused tasks, such as stance detection in relation to presidential candidates (Kawintiranon and Singh, 2022), which, like campaigning detection, requires nuanced interpretation of evaluative and persuasive language.

Our dataset exhibits class imbalance, with the sample of 100 tweets (from the second round of our pilot exercise) showing uneven distribution for `dir_camp` (54%), `ind_camp` (29%) and `non-camp` (17%). As all classes are equally important, we adopt macro-averaged F1-score (macro F1) as the primary metric for model selection and evaluation, ensuring equal weighting across classes. We additionally report per-class metrics for diagnostic analysis and use Cohen's κ to contextualise model performance relative to human–human and human–LLM agreement. This comparison distinguishes genuine model limitations from apparent underperformance due to annotation ambiguity. Given the interpretative nature of the task, model performance can be compared against observed human agreement rather than an assumed noise-free gold standard.

6.2 Training Configurations

All our fine-tuned models were trained under identical configurations. We set `hidden_dropout_prob` = 0.1, `attention_probs_dropout_prob` = 0.1 and `classifier_dropout` = 0.1. Parameter-efficient fine-tuning was applied using LoRA, with rank ($r = 16$), scaling factor ($\alpha = 32$) and dropout of 0.1 applied to the `query`, `key` and `value` projection matrices. LoRA was employed in lieu of full fine-tuning for computational efficiency, as it substantially reduces the number of trainable parameters without significant degradation in downstream task performance (Hu et al., 2021).

Tokenisation was performed using each model's corresponding tokeniser. The maximum sequence length was set to 104 tokens, which was decided by rounding up the 99th percentile of our tokenised sequence lengths to the nearest multiple of 8 for efficient batching. Truncation was enabled and dynamic padding was applied.

To address class imbalance, we applied oversampling using `WeightedRandomSampler`, where class weights were computed as the inverse class frequency, resulting in higher sampling probability for rarer classes. Optimisation was performed using AdamW with a learning rate of 5×10^{-5}, linear scheduling and a warmup ratio of 0.06 followed by linear decay. The batch size was set to 128, with a maximum of 15 training epochs. Early stopping was based on the macro F1 score, with a patience of 4 epochs and a threshold of 1×10^{-4}. Weight decay was set to 0.01, and all experiments were run with a fixed random seed (17) on an NVIDIA A100-SXM4-80GB GPU.

6.3 Hyperparameter Optimisation

While most of our configurations used the recommended values, we experimented with different learning rates, sampling techniques, and cross-entropy loss calculation methods. All hyperparameter optimisation experiments used our human-annotated validation set. We evaluated learning rates in the range of 5×10^{-5} to 3×10^{-4}, with 5×10^{-5} yielding the best validation macro F1 score and was thus selected for all experiments.

Performance on the `ind_camp` class was consistently lower than other classes. We therefore experimented with class-weighted cross-entropy, but this did not improve results. Since loss re-weighting only adjusts gradient magnitudes after batches are formed, batches dominated by the majority class still contained relatively few minority examples per update step. We also explored re-weighting the loss based on measured LLM–human annotation reliability, but this also did not yield improvements, likely because the class performance gap was primarily driven by overlapping decision boundaries rather than uniform label noise that could be mitigated through global weight adjustments. We therefore applied minority class over-sampling during training, where sampling weights were computed as the inverse of class frequency.

7 Results and Discussion

7.1 Comparison of Models

In this section, we compare our different pre-trained transformer models that were fine-tuned to assess the impact of domain-specific pre-training on campaign content detection and to establish a baseline for future work on this task. Table 1 shows the performance of the fine-tuned models on the validation set.

Model	Macro F1
BERT	0.684
BERTweet	**0.703**
PoliBERTweet	0.685

Table 1: Model performance comparison on the validation set.

PoliBERTweet performs almost identically to the base BERT model, while BERTweet outperforms both alternatives. An analysis of per-class performance in Table 2 shows that BERTweet achieves higher scores across all three classes, indicating that its overall ranking is robust and not driven by skewed performance on a single class. Its largest advantage emerges in the `ind_camp` class, suggesting more consistent identification of this more challenging boundary class.

BERTweet outperforms PoliBERTweet despite the latter being pre-trained on a larger set of political tweets. This difference can be explained by variations in pre-training scale, initialisation and training procedure. BERTweet was pre-trained on 850 million English tweets using the RoBERTa pre-training procedure, which includes dynamic masking, removal of next sentence prediction, large batch training and byte-level BPE tokenisation. In contrast, PoliBERTweet was initialised from BERT-base and further pre-trained on approximately 5 million English tweets related to the 2020 US Presidential Election. Although PoliBERTweet benefits from political domain specificity, BERTweet's substantially larger tweet corpus and more robust pre-training strategy likely provide stronger general representations of the language used in Twitter. These advantages appear to outweigh the benefits of political topic specialisation after fine-tuning.

To assess the impact of domain-specific pre-training in a zero-shot setting, we compare the performance of raw BERTweet and PoliBERTweet on the validation set, without fine-tuning (see Table 5 in Appendix D). BERTweet was able to produce

reasonable predictions only for the non-camp class, achieving perfect recall (P = 0.31, R = 1.00), but fails to detect dir_camp and ind_camp, resulting in near-zero precision and recall for those categories. This bias towards non-camp is consistent with the predominantly non-political nature of the general Twitter corpus on which BERTweet was pre-trained, which makes it insensitive to campaign-specific language without fine-tuning.

In contrast, PoliBERTweet (pre-trained on political election tweets) demonstrates substantially stronger zero-shot detection of campaign-related content, particularly for the ind_camp class, where it achieves high recall (R = 0.97). However, it performs poorly on non-camp tweets (R = 0.05), indicating a bias towards predicting political content, likely due to its specialised pre-training on US election data. Where a tweet is not classified as dir_camp, the model defaults to ind_camp. This is likely due to its specialised pre-training on US election data, which likely resulted in limited exposure to non-camp instances, causing the model to default to ind_camp in the absence of clear dir_camp signals.

Overall, these results suggest that political domain pre-training provides a clear advantage in zero-shot campaign detection, although this advantage diminishes once both models are fine-tuned on task-specific data, as fine-tuning exposes both models to the same campaign-related data.

7.2 Analysis of the Best-performing Model

Table 2 presents the detailed performance of the best-performing model. BERTweet achieves strong results on the dir_camp class (F1 = 0.816). However, ind_camp remains the primary bottleneck (F1 = 0.582), with the model substantially underperforming relative to the other classes. Our error analysis indicates that ind_camp is frequently misclassified as either dir_camp or non-camp, with many instances predicted as the former, confirming the blurred boundaries between categories.

This limitation likely stems less from model capacity and more from taxonomy ambiguity and class imbalance, as ind_camp constitutes only 16.14% of the training data. Moderate performance (F1 = 0.710) is demonstrated for the non-camp class, though some confusion with dir_camp remains. Overall, the main challenge lies in the definition and representation of ind_camp rather than in the model architecture.

Class	Precision	Recall	F1
dir_camp	0.778	0.859	0.816
ind_camp	0.559	0.607	0.582
non-camp	0.810	0.632	0.710
Overall Macro F1		0.703	
Overall Accuracy		0.744	

Table 2: Per-class Precision, Recall, and F1 scores obtained by the fine-tuned BERTweet model on the validation set.

7.3 Hierarchical Classification

To better align the model architecture with the human annotation decision tree, we decompose the task into a two-stage (one-vs-rest) setup and fine-tune two instances of BERTweet. Stage A distinguishes between the campaigning and non-campaigning (non-camp) classes, while Stage B takes the tweets predicted as campaign-related and classifies them as either dir_camp or ind_camp.

Performance on the dir_camp class remains the same as the single classifier, while slight improvement can be observed for non-camp (with macro F1 increasing from 0.53 to 0.58). However, performance on ind_camp dropped substantially (from 0.58 to 0.53 macro F1), with many ind_camp instances misclassified as non-camp.

These findings indicate that structurally mirroring the annotation decision tree does not resolve the persistent confusion surrounding the ind_camp class, supporting our hypothesis that improvement requires annotation scheme refinement and enhancing class-specific representation. Consequently, the single multiclass classifier remains preferable, as it provides better overall performance with lower computational cost.

7.4 Performance on the Test Set

Table 3 presents the macro F1 scores of the three models on the held-out test set. BERTweet achieves the highest performance (0.776), outperforming BERT (0.753) and PoliBERTweet (0.762). The ranking observed during validation is maintained on the test set, indicating stable generalisation and confirming BERTweet as the most effective model for our campaign detection task. As previously noted, BERT and PoliBERTweet show comparable performance.

Per-class analysis of BERTweet results (see Table 4) revealed a consistent pattern: dir_camp is the most reliably detected category (F1=0.882), whereas ind_camp remains the most challenging

Model	Macro F1
BERT	0.753
BERTweet	**0.776**
PoliBERTweet	0.762

Table 3: Model performance comparison on the test set.

Class	Precision	Recall	F1
`dir_camp`	0.951	0.823	0.882
`ind_camp`	0.583	0.831	0.685
`non-camp`	0.709	0.819	0.760

Table 4: Per-class Precision, Recall, and F1 scores obtained by the fine-tuned BERTweet model on the test set.

(F1=0.685). BERTweet attains high recall for `ind_camp` (0.831) but lower precision (0.583), indicating that while most indirect instances are captured, false positives persist due to blurred boundaries with neighboring classes. Although the consistency across the validation and test sets demonstrates robust generalisation, the difficulty with `ind_camp` remains, suggesting the models' limitation in understanding this category.

7.5 Political Campaigning in the 2020 US Presidential Election

We applied our best-performing model at scale on the full US 2020 dataset containing 27,620 tweets, covering the six weeks prior to and the week of the Election Day (3 November 2020). This allowed us to examine role-based and temporal patterns in political campaigning behaviour during the final phase of the election.

In Figure 2, one can observe how the distribution of the tweets (in the full dataset) changed over the six weeks in the run-up to the election (in Week 7). Interestingly, political actors seem to have used direct campaigning at a fairly consistent level in the first four weeks of the six-week period; however, this seems to have declined in the two weeks right before the election. Upon producing the breakdown of the weekly distribution according to role groups (see Figures 3 and 4 in Appendix E, we observed that this decline can be attributed to the presidential and vice-presidential candidates seemingly using much less of direct campaigning and more of indirect campaigning two weeks right before the election.

8 Conclusion and Future Work

In this paper, we describe the development of a new annotation scheme for detecting whether any given social media post (i.e., a tweet) can be considered as political campaigning content. We conducted a two-round annotation exercise that confirms that humans are able to apply the scheme consistently on a subset of tweets, and that an LLM can obtain agreement with a human that is close enough to human-human agreement. Afterwards, we fine-tuned three baseline transformer-based encoder models, namely, BERT, BERTweet and PoliBERTweet, for the political campaign content detection task. The fine-tuned BERTweet model obtained the best performance, with a macro-averaged F1-score of 0.776 on the held-out test set.

Indirect campaigning tweets pose a challenge to all baseline models. Future work will explore the use of more advanced models, as well as data augmentation strategies to enhance the representation of this class. Additionally, we will explore detecting a distinct "Call to Action" class, capturing messages mobilising readers to undertake specific actions, which is a more targeted form of support than direct or indirect campaigning.

Limitations

While the two-round pilot test iteratively refined the taxonomy, the moderate IAA reflects an inherent limitation of the classification scheme: the boundary between `ind_camp` and `non-camp` is partly subjective and could not be fully resolved through improved annotation guidelines alone. We consider the current taxonomy to be a sound conceptual framework grounded in electoral law, but further iterations informed by larger-scale annotation studies are needed to sharpen class boundaries. As a consequence, performance differences between models should be interpreted with caution, as some errors may still reflect taxonomy ambiguity rather than model limitations.

The baseline models we used consistently struggle with the indirect campaigning (`ind_camp`) class, likely due to its limited representation in the training data. We did not explore state-of-the-art generative large language models. This is due to our envisioned end-users of automated tools that detect political campaign content: regulatory bodies, NGOs and not-for-profit organisations, who might not necessarily have access to computational or financial resources required by generative LLMs

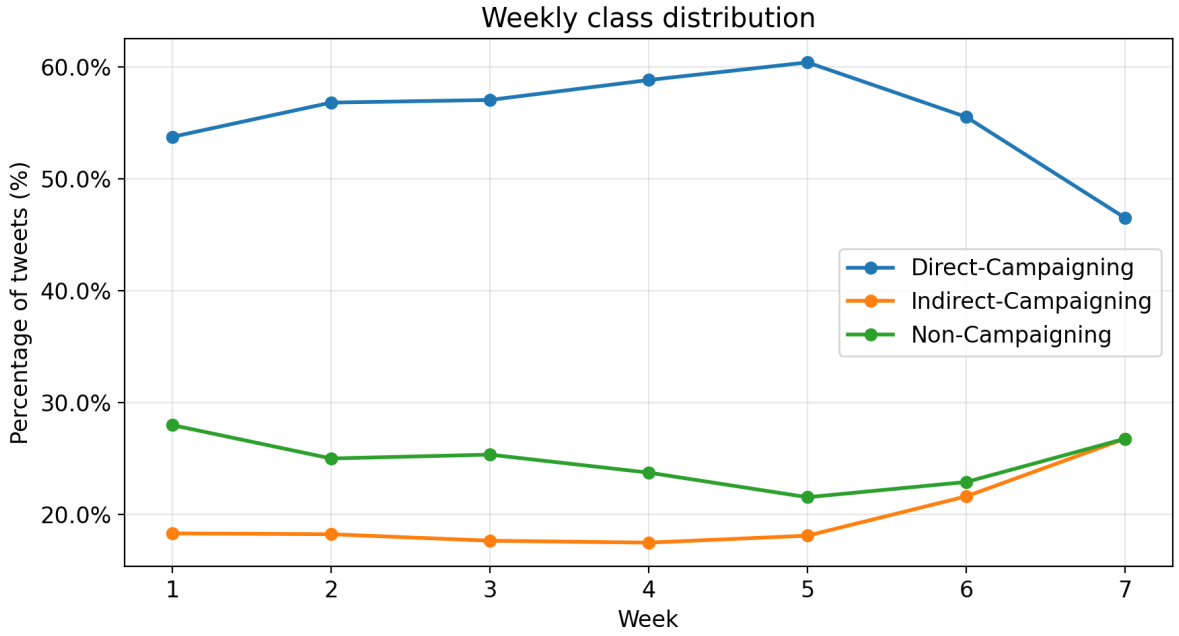


Figure 2: Distribution of the tweets in the weeks right before the US 2020 election, as classified by the best-performing model (fine-tuned BERTweet). Election Day is in Week 7.

in analysing large social media datasets. Lastly, the performance of our best-performing baseline model has not been evaluated across different countries, election cycles or election types beyond the US 2020 presidential elections. Evaluating performance under these settings would help further assess the robustness and generalisability of our proposed approach. We also encourage future work to experiment with using more advanced models like

Ethics statement

The dataset of tweets used in this study was collected as part of a previous project under terms and conditions that prevent us from sharing the data with other researchers. To provide other researchers with insights on how our proposed annotation scheme can be consistently applied, we provided examples of tweets from public, political figures in our annotation guidelines (see Appendix A and B).

References

Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting calls to action in multimodal content: Analysis of the 2021 German federal election campaign on Instagram. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 1–13, Vienna, Austria. Association for Computational Linguistics.

Mateusz Baran, Mateusz Wójcik, Piotr Kolebski, Michał Bernaczyk, Krzysztof Rajda, Lukasz Augustyniak, and Tomasz Kajdanowicz. 2022. Electoral agitation dataset: The use case of the Polish election. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 32–36, Marseille, France. European Language Resources Association.

Emily Chen, Ashok Deb, and Emilio Ferrara. 2021. election2020: the first public twitter dataset on the 2020 us presidential election. *Journal of Computational Social Science*, 5(1):1–18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Electoral Commission. 2026. Statutory guidance on digital imprints.

Erika Franklin Fowler, Michael M Franz, and Travis N Ridout. 2020. Online political advertising in the united states. *Social media and democracy: The state of the field, prospects for reform*, pages 111–138.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Andreas Jungherr. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1):72–91.

Kornraphop Kawintiranon and Lisa Singh. 2022. PoliBERTweet: A pre-trained language model for analyzing political content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367, Marseille, France. European Language Resources Association.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Michael Magin, Nicole Podschuweit, Jan Haßler, and Uwe Russmann. 2017. Campaigning in the fourth age of political communication: A multi-method study on the use of facebook by german and austrian parties in the 2013 national election campaigns. *Information, Communication & Society*, 20(11):1698–1719.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *Preprint*, arXiv:2005.10200.

Andrea Römmele and Rachel K. Gibson. 2020. Scientific and subversive: The two faces of the fourth era of political campaigning. *New Media & Society*, 22(4):595–610.

Vera Sosnovik, Romaissa Kessi, Maximin Coavoux, and Oana Goga. 2023. On detecting policy-related political ads: An exploratory analysis of meta ads in 2022 french election. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 4104–4114, New York, NY, USA. Association for Computing Machinery.

Jesper Strömbäck. 2008. Four phases of mediatization: An analysis of the mediatization of politics. *The International Journal of Press/Politics*, 13(3):228–246.

Mark Vergeer. 2015. Twitter and political campaigning. *Sociology Compass*, 9(9):745–760.

Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2021. Automatic detection and categorization of election-related tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):703–706.

Emi Yoshikawa and Franziska Roesner. 2025. Exploring political ads on news and media websites during the 2024 u.s. elections. *Preprint*, arXiv:2503.02886.

Appendix

A Annotation Guidelines

STEP 1

BINARY QUESTION: Does the tweet text directly reference a political entity?

• This strictly includes political parties, political candidates, future/prospective political candidates, and elected office holders (e.g. MPs, Mayors, Councillors, Governors, Senators, Prime Ministers, Presidents, etc.)
• The entity(s) must be directly named in the text, and this can also include @mentioned user accounts and relevant hashtags. For example, #trump2020 or #DemsLose would classify as direct references for Donald Trump and the Democratic Party respectively.
• Where specific political entities are not refenced directly by name but by their specific role/title (e.g. "Prime Minister", "the President", "POTUS", "Governor for X"), this classifies as a direct reference.
• It does not matter how many entities are referenced as long as there is at least one.
• Make sure to double-check where possible if the referenced entity meets one of the four entity categories. Former political candidates and elected office holders (at the time of the post) do not account as a direct reference.
• There are certain cases where the use of self-referential pronouns may also count as a direct reference, but this is contextual based on the language of the text. Given that every account in the tweet data was a US senator, where they refer to themselves as "I" "Me", "My" this counts as a direct reference to a political entity. E.g.
"I am the best person for the job. Lend me your vote and I promise it will not go to waste!"
"I have fought against the growing tide of fascism for over decade, and I will not give in today. No matter how hard my opponents try to slander to me."
"My job is to make sure that the voices of ordinary people are heard when it comes to concerns over immigration."
• Where they use plural self-referential pronouns like "we" or "us" or "our", whether this classifies as a direct reference is contingent on the context they are used. If it appears they are using the term in reference to their political party, this will count as a direct reference. E.g.
"We are the party on the side of the people; they are the party of the rich and corrupt."
"If you give us your vote this November, we promise not to let you down!"
"We are working tirelessly to improve transport up and down the country."
• However, if it used in reference to society or to make an appeal to the general population more broadly, this would not count as a direct reference. E.g.
"We deserve better as a nation. This is not what America stands for." "We should not accept this new policy proposal from our Government; we need to do more to help protect the poorest communities in our society."
"It's up to us to save the planet, together, united."

If answer to STEP 1 is YES: proceed to STEP 2.A
If answer to STEP 1 is NO: proceed to STEP 2.B

STEP 2.A

BINARY QUESTION: Does the tweet text encourage the reader to give to or withhold support from the directly reference entity(s)?
• Encouraging/discouraging support for a political entity can be explicit: e.g.

"The only way to secure a fairer future is with the Green Party. Join us this election—vote Green, vote for change!"
"Our community needs strong leadership. That's why I'm voting for Sarah Thompson on May 5th. She's the only candidate who will fight for working families. ThompsonForMayor"
"President Alvarez has delivered on jobs, healthcare, and education. Let's keep building on that progress—re-elect Alvarez this November!"
• Or encouraging/discouraging support for a political entity can be more suggestive based on sentiment: e.g.
"Hard not to notice how much better things have gotten since Mayor Patel took office. Feels like someone's finally listening to us."
"Sure, let's give the Democrats that caused the housing crisis another term. What could possibly go wrong?"
"Another broken promise from the Conservatives today... starting to lose count. Guess some things never change."
• It can also include instances where the author is attempting to elicit support via non-political or non-partisan dimensions like appeals to family, religious, national or cultural values:
"There is nothing more important to me in this world than my faith. I am proud to represent all devoted Hindus up and down the country."
"As a true American patriot, Senator Green is committed to upholding the values embedded in our constitution."
"The Republicans claim to be the party of family, but really they're the party of greed and self-interest."

If answer to STEP 2.A is YES: This is DIRECT CAMPAIGNING
If answer to STEP 2.A is NO: This is NON-CAMPAIGNING

STEP 2.B

BINARY QUESTION: Does the tweet text directly reference a linked characteristic of a political entity(s)?
• This includes policies, political ideologies, or political opinions, as well as other categories not based on policies or opinions such as sociodemographics or personal qualities.
• The referenced characteristic should be reasonably attributable to one of the four political entities:
• References to policies can include:
o Explicitly named policies like "Online Safety Act 2023" or the "One Big Beautiful Bill Act 2025".
o Policy areas like "healthcare", "crime", "foreign aid", "welfare", "education", "gun ownership", "human rights", "taxation" and so on.
• References to political ideologies can include:
o Explicitly named ideologies like "socialism", "conservatism", "capitalism", "liberalism", "populism", "fascism" and their variations.
o Or general political leanings like "left-wing", "right-wing", "centrists" "far-right", "far-left" and their variations.
o It can also include broader references to the facets of an ideology such as "free markets", "small government", "personal liberties" and so on.
• References to political opinions can include:
o Subjective takes, judgments, reactions, values, or feelings without direct reference to a policy or political entity. I.e.
"It seems that politics over the last few years has become more about bickering than dealing with pressing issues facing this country."

"The country is more divided now than ever before"
"The media seems more interested in pushing an agenda than actually covering the issues that people care about"
"Well, that debate was a complete joke. No wonder people are losing faith in the system!"
• References to personal features or characteristics can include:
o Explicit socio-demographic or personal features like gender, ethnicity, age, educational background, geographic location or occupation.
o Or more vague references to positive and negative traits about certain groups. E.g.
"The country is being ran into the ground by a small group of corrupt individuals who think they are above the law"
"99% of the population are hardworking, honest, and decent individuals. It's the 1% that are the problem"
"There are some people in politics who take their jobs seriously, and others who see it as a chance to progress their own careers."
• It does not matter how many characteristics are referenced as long as there is at least one.
• Characteristics can either be directly referenced within the tweet text or via relevant hashtags. For example, BuildTheWall or PlayIn4Climate can be considered as support for Trump's policy to build a wall along the US-Mexico border and support for reduction in air pollution respectively. Forthe99% or AgainstCapitalists would be considered support references to personal characteristics and an ideology.

If answer to STEP 2.B is NO: This is NON-CAMPAIGNING
If answer to STEP 2.B is Yes: proceed to STEP 3

STEP 3

BINARY QUESTION: Does the tweet text encourage the reader to give to or withhold support from the refenced characteristic(s)?
• Encouraging/discouraging support for a linked characteristic can be explicit: e.g.
"Support the Clean Air for Schools Act — every child deserves classrooms free from toxic pollution."
"Do not stand with those who dismiss the struggles of others — empathy is essential in public life."
"Stand behind the Affordable Homes Guarantee Bill to ensure safe housing is within everyone's reach."
"Reforms to vital medical provisions for the elderly is cruel and callous."
• Or encouraging/discouraging support for a characteristic can be more suggestive based on sentiment: e.g.
"Free markets shouldn't be allowed to decide everything — community values must matter too."
"Our right to the freedom of expression is crucial to everything this country stands for. It needs to be protected at all costs."
"Politics is dominated by private school kids who have no idea about the real world. We need more people who know about the real world."
"Veterans fought for this country and now they have been thrown to wayside. Don't let their sacrifices be in vain."

If answer to STEP 3 is YES: This is INDIRECT CAMPAIGNING
If answer to STEP 3 is NO: This is NON-CAMPAIGNING

B Example Tweets for All Decision Tree Outcomes

Pathway 1: Direct Campaigning

Directly References an Entity(s): YES
Tweet Encourages to Give/Withold Support: YES

"Senator Davidson is a fantastic leader and an even better person. The people of Arizona are lucky to have him!"
"I have always stood for working people and I always will"

Pathway 2: Direct Reference, Not Campaigning (Non-Campaigning)

Directly References an Entity(s): YES
Tweet Encourages to Give/Withold Support: NO

"I have scheduled a live Q&A session on Twitch tonight at 7pm"
"Senator Grey has passed away this evening, aged 78."

Pathway 3: Indirect Campaigning

Directly References an Entity(s): NO
Directly References Linked Characteristic: YES
Tweet Encourages to Give/Withold Support: YES

"Rolling back protections in the Digital Privacy Reform Act puts people at risk. That change should not go unchallenged. @PrivacyWatch #DataRights"
"Communities thrive when solidarity is put ahead of unchecked markets. Let's not lose sight of those values. #FairEconomy"

Pathway 4: Indirect Reference, Not Campaigning (Non-Campaigning)

Directly References an Entity(s): NO
Directly References Linked Characteristic: YES
Tweet Encourages to Give/Withold Support: NO

"The Green Transport Act was passed in 2021."
"The National Housing Strategy includes five chapters."

Pathway 5: No Reference (Non-Campaigning)

Directly References an Entity(s): NO
Directly References Linked Characteristic: NO

"Merry Christmas to all those who celebrate!"
"Anyone know why the Twitter server was down this afternoon?"

C LLM Annotation Prompts

system prompt STEP 1

You are an impartial and precise text classification model trained to detect whether a tweet (that is written by a political figure) directly references a political entity. You must read the tweet carefully, apply the rules below, and respond only with "YES" or "NO", no extra text.

Rules:
A political entity includes:
- Political parties
- Political candidates
- Future or prospective political candidates
- Elected office holders (e.g. MPs, Mayors, Councillors, Governors, Senators, Prime Ministers, Presidents, etc.)

Answer YES if any of the following conditions are met:

1) The tweet explicitly names a political entity. Example: "Democrats," "Donald Trump," "President Biden."
2) The tweet includes an @mention or hashtag referring to a political entity. Example: @JoeBiden, VoteLabour, trump2020, DemsLose
3) The tweet refers to a political entity by role or title. Example: "The Prime Minister", "the President", "Governor of Texas", "POTUS", "Governor for X".
4) The tweet author (who is always a political entity) refers to themselves using singular pronouns ("I", "me", "my"). Example: "I am honored to serve the people of Ohio", "I am the best person for the job. Lend me your vote and I promise it will not go to waste!".
5) The author uses plural pronouns ("we," "our," "us") in reference to their political party or campaign. Example: "We are the party of progress," "If you give us your vote this November, we won't let you down."
Answer NO in all other cases, including but not limited to the following cases:
- The tweet refers to "we," "our," or "us" in a societal or national sense. Example: "We deserve better as a nation", "It's up to us to save the planet, together, united.".
- The tweet mentions "government," "Congress," or similar institutions without specifying a political entity or title. Example: "The government should do more to protect the environment."

Respond strictly with "YES" or "NO"

system prompt STEP 2.A

You are an impartial and precise text classification model trained to detect whether a tweet (that directly references a political entity) encourages or discourages support for that entity. You must read the tweet carefully, apply the rules below, and respond only with "YES" or "NO", no extra text.

Rules:
A tweet encourages or discourages support for a political entity when it expresses, implies, or suggests that readers should:
- Give support to the referenced political entity (e.g. vote for, endorse, trust, or praise them), OR
- Withhold support from the referenced political entity (e.g. criticize, oppose, or reject them).

Answer YES if any of the following conditions are met:
1) The tweet explicitly urges the reader to support or reject the political entity.
Example: “Vote Green this election!”, “Re-elect Alvarez this November!”, “Our community needs strong leadership. That’s why I’m voting for Sarah Thompson.”
2) The tweet praises or criticizes a political entity in a way that clearly implies endorsement or opposition.
Example: “President Alvarez has delivered on jobs, healthcare, and education. Let’s keep building on that progress—re-elect Alvarez this November!”, “Another broken promise from the Conservatives today... starting to lose count.”
3) The tweet expresses sentiment or opinion that can reasonably be interpreted as encouraging or discouraging support for the entity.
Example: “Hard not to notice how much better things have gotten since Mayor Patel took office.”, “Sure, let’s give the Democrats that caused the housing crisis another term. What could possibly go wrong?”
4) The tweet appeals to non-political or cultural values (e.g. family, religion, patriotism) to elicit support or opposition for the political entity.
Example: “There is nothing more important to me in this world than my faith. I am proud to represent all devoted Hindus up and down the country.”, “As a true American patriot, Senator Green is committed to upholding the values embedded in our constitution.”, “The Republicans claim to be the party of family, but really they’re the party of greed and self-interest”

Answer NO in all other cases, including but not limited to the following:
- The tweet only provides factual or neutral information about the political entity without encouraging or discouraging support.
Example: “The President has confirmed that the annual conference will be held in NewYork this September.”, “Senator Grey has passed away this evening.”
- The tweet refers to a political entity in a ceremonial or administrative context without evaluative language.
Example: “The President met with politications in California to discuss trade policy.”

Respond strictly with “YES” or “NO”

system prompt STEP 2.B

You are an impartial and precise text classification model trained to detect whether a tweet (that does not directly reference a political entity) directly references at leaast one linked characteristic of a political entity. You must read the tweet carefully, apply the rules below, and respond only with “YES” or “NO”, no extra text.

Rules:
A tweet directly references a linked characteristic of a political entity if it mentions or implies a policy, political ideology, political opinion, or personal/sociodemographic characteristic that can be reasonably attributed to one or more political entities.

Answer YES if any of the following conditions are met:
• The tweet includes a reference to policies, political ideologies, or political opinions, as well as other categories not based on policies or opinions such as sociodemographics or personal qualities.
• The referenced characteristic should be reasonably attributable to one of the four political entities.
• Characteristics can either be directly referenced within the tweet text or via relevant hashtags. For example, BuildTheWall or PlayIn4Climate can be considered as support for Trump’s policy to

build a wall along the US-Mexico border and support for reduction in air pollution respectively. Forthe99% or AgainstCapitalists would be considered support references to personal characteristics and an ideology.

Explanation:
• References to policies can include:
o Explicitly named policies like “Online Safety Act 2023” or the “One Big Beautiful Bill Act 2025”.
o Policy areas like “healthcare”, “crime”, “foreign aid”, “welfare”, “education”, “gun ownership”, “human rights”, “taxation” and so on.
• References to political ideologies can include:
o Explicitly named ideologies like “socialism”, “conservatism”, “capitalism”, “liberalism”, and their variations.
o Or general political leanings like “left-wing”, “right-wing”, “centrists” “far-right”, and their variations.
o It can also include broader references to the facets of an ideology such as “free markets”, “small government”, “personal liberties” and so on.
• References to political opinions can include:
o Subjective takes, judgments, reactions, values, or feelings without direct reference to a policy or political entity. Examples:
“It seems that politics over the last few years has become more about bickering than dealing with pressing issues facing this country.”
“The media seems more interested in pushing an agenda than actually covering the issues that people care about”
“Well, that debate was a complete joke. No wonder people are losing faith in the system!”
• References to personal features or characteristics can include:
o Explicit socio-demographic or personal features like gender, ethnicity, age, educational background, geographic location or occupation.
o Or more vague references to positive and negative traits about certain groups. Examples:
“The country is being ran into the ground by a small group of corrupt individuals who think they are above the law”
“99“There are some people in politics who take their jobs seriously, and others who see it as a chance to progress their own careers.”

Answer NO in all other cases, including but not limited to the following:
- The tweet does not contain any reference to a policy, ideology, opinion, or personal/sociodemographic characteristic.
- The tweet is purely personal, social, or unrelated to politics or governance.

Respond strictly with “YES” or “NO”

system prompt STEP 3

You are an impartial and precise text classification model trained to detect whether a tweet (that directly references a linked characteristic of a political entity) encourages or discourages support for that characteristic. You must read the tweet carefully, apply the rules below, and respond only with “YES” or “NO”, no extra text.

Rules:
A tweet encourages or discourages support for a linked characteristic when it expresses, implies, or

suggests that readers should:
- Give support to the referenced characteristic (e.g. agree with, promote, or defend it), OR
- Withhold support from the referenced characteristic (e.g. criticize, oppose, or reject it).

Answer YES if any of the following conditions are met:

• Encouraging/discouraging support for a linked characteristic can be explicit: e.g.
"Support the Clean Air for Schools Act — every child deserves classrooms free from toxic pollution."
"Do not stand with those who dismiss the struggles of others — empathy is essential in public life."
"Stand behind the Affordable Homes Guarantee Bill to ensure safe housing is within everyone's reach."
"Reforms to vital medical provisions for the elderly is cruel and callous."
• Or encouraging/discouraging support for a characteristic can be more suggestive based on sentiment: e.g.
"Free markets shouldn't be allowed to decide everything — community values must matter too."
"Our right to the freedom of expression is crucial to everything this country stands for. It needs to be protected at all costs."
"Politics is dominated by private school kids who have no idea about the real world. We need more people who know about the real world."
"Veterans fought for this country and now they have been thrown to wayside. Don't let their sacrifices be in vain."

Answer NO in all other cases, including but not limited to the following:
- The tweet merely states or describes a characteristic without encouraging or discouraging support.
- The tweet provides factual or neutral information about a policy, ideology, or social characteristic without evaluative or persuasive language.

Respond strictly with "YES" or "NO"

D Raw Model Performance

Model	Class	Precision	Recall	F1
BERTweet	dir_camp	0.000	0.000	0.000
	ind_camp	0.000	0.000	0.000
	non-camp	0.312	1.000	0.476
PoliBERTweet	dir_camp	0.125	0.001	0.003
	ind_camp	0.181	0.967	0.305
	non-camp	0.315	0.055	0.093

Table 5: Per-class Precision, Recall, and F1 scores obtained on the validation set by the raw BERTweet and PoliBERTweet models, i.e., without any fine-tuning.

E Visualisation of Results from the Application of the Best-performing Model at Scale

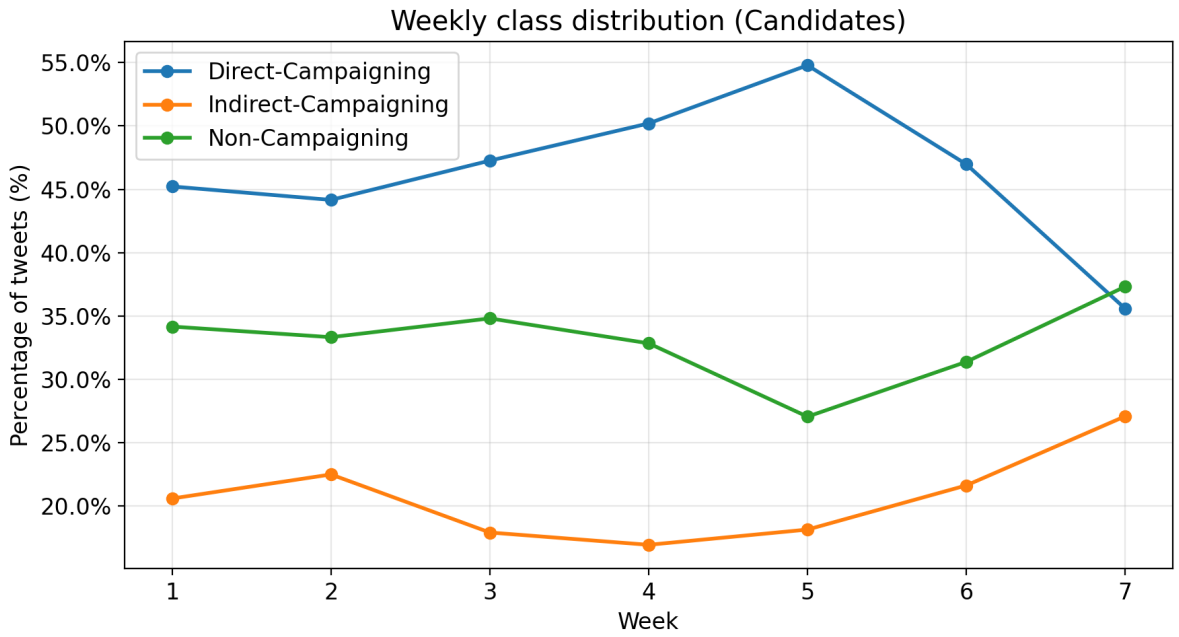


Figure 3: Distribution of the tweets from presidential and vice-presidential candidates in the weeks right before the US 2020 election, as classified by the best-performing model (fine-tuned BERTweet). Election Day is in Week 7.

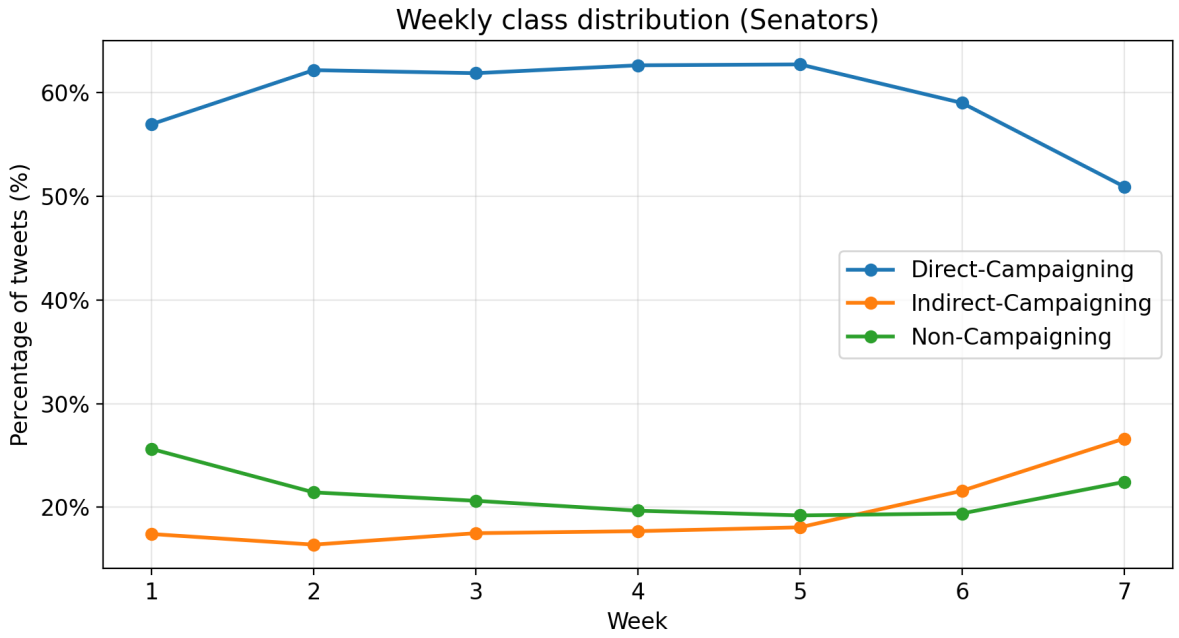


Figure 4: Distribution of the tweets from Democratic and Republican senators in the weeks right before the US 2020 election, as classified by the best-performing model (fine-tuned BERTweet). Election Day is in Week 7.

F Political Figures in the Dataset

Account Holder	Twitter Account	Position
Joe Biden	@JoeBiden	Presidential Candidate
Donald Trump	@realDonaldTrump	Presidential Candidate
Howie Hawkins	@HowieHawkins	Presidential Candidate
Jo Jorgensen	@Jorgensen4POTUS	Presidential Candidate
Kamala Harris	@KamalaHarris	Vice-Presidential Candidate
Mike Pence	@Mike_Pence	Vice-Presidential Candidate
Angela N. Walker	@AngelaNWalker	Vice-Presidential Candidate
Spike Cohen	@RealSpikeCohen	Vice-Presidential Candidate

Table 6: Twitter accounts of presidential and vice-presidential candidates included in the dataset.

Account Holder	Twitter Account
Abby Broyles	@abbybroyles
Dan Ahlers	@ahlers_dan
Amy McGrath	@AmyMcGrathKY
Barbara Bollier	@BarbaraBollier
Marquita Bradshaw	@Bradshaw2020
Cal Cunningham	@CalforNC
Mark Kelly	@CaptMarkKelly
Chris Coons	@ChrisCoons
CJ Ellison	@CJSenate2020
Cory Booker	@CoryBooker
Al Gross	@DrAlGrossAK
Paulette Jordan	@electpaulette
Theresa Greenfield	@GreenfieldIowa
Jaime Harrison	@harrisonjaime
John Hickenlooper	@Hickenlooper
Mark Warner	@MarkWarner
Monica Ben-David	@MBenDavid2020
Mike Espy	@MikeEspyMS
MJ Hegar	@mjhegar
Jon Ossoff	@ossoff
Paula Jean Swearengin	@paulajean2020
Ben Ray Luján	@repbenraylujan
Raphael Warnock	@ReverendWarnock
Ricky Torres	@RickyForSenate
Sara Gideon	@SaraGideon
Dick Durbin	@SenatorDurbin
Jeanne Shaheen	@SenatorShaheen
Doug Jones	@SenDougJones
Gary Peters	@SenGaryPeters
Jack Reed	@SenJackReed
Jeff Merkley	@SenJeffMerkley
Ed Markey	@SenMarkey
Tina Smith	@SenTinaSmith
Steve Bullock	@stevebullockmt

Table 7: Twitter accounts of Democratic senators included in the dataset.

Account Holder	Twitter Account
Bill Hagerty	@BillHagertyTN
Bryant “Corky” Messner	@CorkyForSenate
Cynthia Lummis	@CynthiaMLummis
Steve Daines	@DainesforMT
Daniel Gade	@gadeforvirginia
Jim Inhofe	@JimInhofe
John James	@JohnJamesMI
Kevin O’Connor	@KOCforSenate
Lauren Witzke	@LaurenWitzkeDE
Jason Lewis	@LewisForMN
Lindsey Graham	@LindseyGrahamSC
Mark Ronchetti	@MarkRonchettiNM
Mitch McConnell	@McConnellPress
Shane Perkins	@PerkinsForUSSen
Doug Collins	@RepDougCollins
Rik Mehta	@RikMehta_NJ
Susan Collins	@SenatorCollins
Kelly Loeffler	@SenatorLoeffler
Jim Risch	@SenatorRisch
Mike Rounds	@SenatorRounds
Shelley Moore Capito	@SenCapito
Cory Gardner	@SenCoryGardner
Dan Sullivan	@SenDanSullivan
David Perdue	@sendavidperdue
Cindy Hyde-Smith	@SenHydeSmith
Joni Ernst	@SenJoniErnst
Martha McSally	@SenMcSallyAZ
Ben Sasse	@SenSasse
Thom Tillis	@SenThomTillis
Tom Cotton	@SenTomCotton
John Cornyn	@TeamCornyn
Tommy Tuberville	@TTuberville
Joan Waters	@watersforsenate

Table 8: Twitter accounts of Republican senators included in the dataset.

Mapping the Landscape of Unregulated eXplicit Contents on Reddit

MSVPJ Sathvik[1,*], **Manan Roy Choudhury**[2,*], **Rishita Agarwal**[3], **Sathwik Narkedimilli**[4], **Thao Ha**[2], **Liesel Sharabi**[2], and **Vivek Gupta**[2,†]
[1]University of Birmingham, [2]Arizona State University
[3]Indian Institute of Technology Guwahati [4]National University of Singapore
[*]Equal contribution (co-first authors) [2,†]Corresponding Author
msvpjsathvik@gmail.com, mroycho1@asu.edu, rishita@iitg.ac.in,
sathwik.narkedimilli@ieee.org, thaoha@asu.edu, liesel.sharabi@asu.edu
[2,†]**Correspondence:** vgupt140@asu.edu | **Project Website:** Link

Abstract

The rise of online platforms has facilitated the dissemination of explicit content, posing significant challenges for detection and regulation. Often using coded language to bypass moderation, this content erodes user trust and may be associated with scam-related risks, posing direct financial and personal risks. In this study, we map the landscape of online explicit content posts, focusing on their categorization, linguistic strategies, and temporal and behavioral patterns as they appear within the mainstream platform Reddit. We investigated five distinct content categories including Virtual Services (VS), Physical Services (PS), Exhibitionism (Ex), Couples and Group Interactions (CGI), and Content Creation and Sales (CCS) and performed large-scale experimentation using state-of-the-art large language models (LLMs) such as GPT-4, LLaMA 3.3-70B-Instruct, Gemini 1.5 Flash, Mistral 8×7B, Qwen 2.5 Turbo, and Claude 3.5 Haiku. Our work demonstrates that a nuanced classification of these services requires moving beyond simple keywords, and we establish that expressive signals, such as sentiment, emotion, and tone, are critical for accurate detection. Our analysis reveals distinct behavioral and psychosocial expression patterns for each service category, providing a robust framework for future moderation.

1 Introduction

Modern social media platforms are vast, multipurpose ecosystems that enable unprecedented self-expression and community formation. However, this openness also creates vulnerabilities that facilitate the proliferation of unregulated explicit-service solicitations, posing systemic challenges to online safety (Marche et al., 2023; Raponi et al., 2022). While such solicitations may appear across diverse communities, identifying and studying them requires reliable ground-truth examples. To establish a foundational benchmark, we analyze explicit-service posts within dedicated Reddit communities where solicitation behaviors are observable and consistently labeled (Baumgartner et al., 2020; Gothard, 2021). These posts frequently employ coded language, euphemisms, and suggestive signals designed to bypass moderation, and are sometimes associated with scams and fraudulent schemes that pose financial and personal risks. This controlled setting enables systematic analysis of linguistic, behavioral, and engagement patterns that can inform future detection efforts in broader, heterogeneous social media environments (Trager et al., 2022; Teitelbaum, 2020).

The scope of these solicitations spans multiple modalities, including Virtual Services (VS), such as live video calls; Physical Services (PS), involving in-person transactions; Exhibitionism (Ex); Couples and Group Interactions (CGI); and Content Creation and Sales (CCS). These categories reflect the adaptability of actors promoting illicit services and expose the limitations of moderation strategies that rely solely on surface-level keyword filtering rather than contextual and behavioral indicators.

The risks associated with such content are multilayered. Minors face heightened exposure and vulnerability to grooming and exploitation, while adult users may encounter scams, financial loss, and erosion of trust in digital spaces(Amirkhani et al., 2026; Sanchez and Genelza, 2025; Gupta, 2025). At the governance level, these solicitations strain moderation infrastructure, often necessitating hybrid human–AI approaches to manage scale and contextual ambiguity (Peter and Valkenburg, 2016; Ferguson and Hartley, 2022; Mitchell et al., 2003). Furthermore, the issue intersects with complex legal and regulatory frameworks (Government of NCT of Delhi, 2023). Many jurisdictions criminalize aspects of digital solicitation, while others regulate certain forms of sex work under strict licensing regimes (Government of India, 2000; Ministry of Electronics and Information Technology, Gov-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 271–292

ernment of India, 2021). This global patchwork underscores the need for scalable, context-aware moderation mechanisms that can adapt across legal, cultural, and linguistic contexts.

In response and also seeing recent efforts to stress-test LLMs through domain-specific benchmarks (Choudhury et al., 2025, 2026), this paper introduces REDDIX-NET, a structured benchmark dataset designed to map and analyze explicit-service solicitations on a mainstream platform. We do not frame this work as a direct detection task in mixed-content environments; rather, we provide a systematic behavioral and linguistic characterization that serves as a precursor for developing detection systems in broader settings. Beyond categorization, we demonstrate that expressive signals, i.e., sentiment, emotion, and tone, play a meaningful role in classification and in understanding engagement dynamics across service types, thereby establishing a more context-aware moderation benchmark.

In this paper, we present the analysis, experimentation, and dataset construction underlying REDDIX-NET. By leveraging multilingual data and evolving online trends, the dataset supports AI-driven categorization of distinct solicitation modalities that frequently operate outside regulated channels and platform policies.

The contributions of this work are as follows:

- We provide the first large-scale analysis of solicited explicit content within service-oriented communities on a mainstream social media platform.
- We introduce REDDIX-NET, a curated benchmark dataset that fills a critical resource gap for moderation and safety research.
- We analyze sentiment and user expression patterns to derive psychosocial engagement indicators that extend beyond surface-level content filtering.

2 Related Work

Sex Trafficking and Escort Advertisement Detection. Prior work on online sexual services primarily targets sex trafficking networks and illicit escort advertisements on dedicated platforms. (Ibanez and Suthers, 2016b,a) used network and content analysis to identify trafficking indicators from structured metadata such as phone numbers and social links. Keskin et al. (2021) analyzed large-scale escort ads to model mobility and service circuits, while Giommoni and Ikwu (2021) extracted trafficking indicators from UK-based advertisements. These approaches focus on overt ads and structured signals rather than conversational content.

Supervised Classification and Ordinal Risk Modeling. Machine learning methods have also been applied to classify illicit businesses and trafficking-related advertisements. Diaz and Panangadan (2020) trained supervised models on review-site data to distinguish legitimate from illicit services, and Wang et al. (2020a,b) proposed Ordinal Regression Neural Networks for predicting trafficking likelihood scores from ad text. These formulations emphasize binary or ordinal risk detection over explicit advertisements, not nuanced solicitation behavior in social media contexts.

NSFW Filtering and Content Moderation. Traditional NSFW detection relies on keyword filtering and image-based recognition (Davidson et al., 2017; Founta et al., 2018; Vidgen et al., 2021; Li et al., 2026; Bao et al., 2025; Jangra et al., 2025; Yang et al., 2025). While effective for overt content, such systems struggle with coded language, euphemisms, and contextual ambiguity, and typically treat detection as surface-level classification rather than modeling expressive linguistic and interaction dynamics.

User Behavior and Engagement in Sexual Content Communities. Studies on online sexual content communities indicate that engagement patterns differ from general social media, particularly in anonymous environments like Reddit. Users often engage in sensitive self-disclosure and receive varied responses such as support and validation (Andalibi et al., 2018), while also seeking sexual advice through intent-driven interactions (PettyJohn et al., 2025). Such disclosures tend to attract higher engagement and repeated participation (Haq et al., 2025). Anonymity facilitates open expression (Shelton, 2015), and community norms shape interaction behavior (Brown, 2018). Overall, interactions combine social, informational, and transactional elements, where comments may include requests, verification, or negotiation, offering insights into user intent beyond the original post.

Our Contribution. In contrast, we examine explicit-service solicitations on a mainstream platform (Reddit), specifically within dedicated service-oriented communities that provide high-

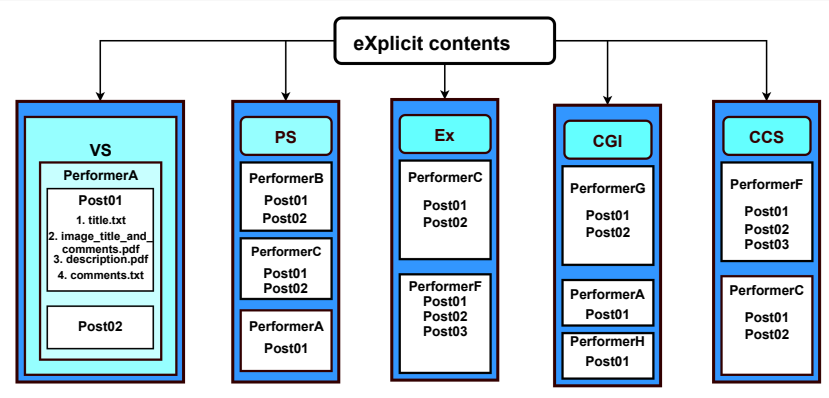


Figure 1: Structure of the proposed dataset, categorizing online sexual services across five distinct categories and their further subdivisions.

confidence ground-truth examples of solicitation behavior. This controlled setting enables systematic analysis of linguistic, behavioral, and engagement patterns that can inform future detection systems in broader environments. Methodologically, we move beyond surface lexical cues by leveraging expressive signals, i.e., sentiment, emotion, and tone, benchmarking state-of-the-art LLMs for context-aware classification. We further analyze comment-level interactions and temporal dynamics, offering a computational social science perspective on solicitation behavior within explicit-service communities on a mainstream platform.

3 REDDIX-NET CONSTRUCTION

This section outlines the methodology and sources used to build REDDIX-NET, providing a comprehensive, data-driven foundation. It highlights the strategic approach taken to construct a robust and insightful dataset.

3.1 Data Collection

The dataset was collected from three large subreddit channels (each with over 50K members) focused on online sexual services with users from all over the world. Although these communities explicitly focus on adult services, they exist within a mainstream social media platform and retain conversational, user-generated characteristics distinct from those of structured advertising marketplaces. This allows us to analyze solicitation behavior in a realistic social interaction context while maintaining high labeling confidence. Our dataset includes posts from 2014 to 2023. Due to privacy policies, their names are withheld. Using the Reddit API with PRAW API, we gathered posts offering services like paid meetups, nude video calls, and couple swaps. More details on data labeling, collection, cleaning, and pre-processing are provided in Appendix 4.

3.2 Data Annotation

For the experiments, three annotators, including the authors, manually reviewed and categorized the posts. The team was selected to mitigate gender and regional bias, comprising two male and one female annotator from three different states in India, bringing diverse linguistic and socio-cultural perspectives across English and several Indian languages. All annotators possessed domain familiarity and contextual sensitivity to handle euphemistic or implicit references to adult services. The annotation process was conducted carefully to ensure reliable and balanced classification.

The posts were divided into the following five categories:

1. **Content Creation and Sales (CCS)**: This category includes services related to the production and sale of adult content, such as videos, photos, or written material, often tailored to specific user requests.

2. **Couples and Group Interactions (CGI)**: Services in this class involve collaborative engagements, typically between two or more individuals, either for personal interaction.

3. **Exhibitionism (Ex)**: This category refers to services where individuals perform live or recorded acts for an audience, often emphasizing the act of showcasing themselves in a sexual or provocative context.

4. **Physical Services (PS)**: Services involving physical interaction, such as in-person meetings, escort services, or any physical contact-based activities, are classified here.

5. **Virtual Services (VS)**: This category includes services provided online, such as video chats, private messages, or virtual performances, which do not involve any physical meetings but are sexual or adult-oriented in nature.

The five service categories were derived through iterative qualitative coding grounded in prior online solicitation and digital sexual commerce studies. Two annotators independently reviewed sampled Reddit posts, consolidated recurring behavioral codes, and refined five solicitation modalities. Pilot annotation validated separability, minimized overlap, and ensured contextual consistency. Figure 1 presents the five-class folder structure reviewed.

3.3 Data annotation guidelines

Posts were annotated into five behavioral categories:

Virtual Services (VS), Physical Services (PS), Exhibitionism (Ex), Couples and Group Interactions (CGI), Content Creation and Sales (CCS).

Annotators reviewed the post title, body, any associated (anonymized) media, and user comments. 1) Guidelines included decision heuristics and examples to disambiguate cases (e.g., distinguishing VS from CCS based on service type and platform mention). 2) Posts were labeled via majority vote from three annotators. Uncategorizable posts were removed. 3) Inter-annotator agreement was measured using Krippendorff's alpha, Cohen's kappa, and Fleiss' kappa. 4) Ethical training and precautions were implemented for the annotator's well-being.

3.4 Statistical Analysis

Table. 1 provides an overview of posts. The dataset comprises 8,146 posts across five categories, with Ex containing the most posts (2,302) and CGI the fewest (1,103), indicating a moderate imbalance in class representation. The overall word count across the dataset is 557,764, with Ex again contributing the most words (164,131) and CGI the fewest (73,751). Despite these variations in volume, the average word length per post remains relatively consistent, ranging between 65 and 75 words across categories. Engagement levels, measured through comment activity, follow a similar pattern: the dataset contains 60,240 comments in total, with Ex receiving the largest share (17,923), although the average number of comments per post remains stable across all categories (7.3–8.0), suggesting uniform interaction behavior. Lexical density, reflected in the average tokens per post, ranges from 63 in CCS to 75 in Ex, indicating only minor variation in text granularity.

3.5 Inter-annotator Agreement Score

To assess annotation consistency, we conducted an Inter-Annotator Agreement (IAA) analysis using Krippendorff's Alpha (K), Cohen's Kappa (C) for pairwise comparisons, and Fleiss' Kappa (F) for group agreement across the full dataset. Pairwise Krippendorff scores were $K(1,2) = 0.6633$, $K(1,3) = 0.7470$, and $K(2,3) = 0.6783$, while the overall agreement among all three annotators was $K(1,2,3) = 0.6963$ (Table. 2). These results indicate substantial inter-annotator reliability and consistent labeling across annotators.

4 Analysis of REDDIX-NET

This section outlines key experiments on REDDIX-NET that use LLMs and PLMs for user classification, sentiment analysis, comment classification, and metadata-temporal analysis. Details follow in subsequent subsections.

4.1 REDDIX-NET User Classification

This experiment aims to identify users offering specific services based on their posts using LLMs. Users often employ coded language and suggestive presentation to evade moderation. While our analysis focuses primarily on textual signals (titles, descriptions, and comments), the dataset retains contextual references to associated media, enabling future multimodal extensions. Users' services are treated as ground truth, and LLMs are prompted to classify posts into predefined service categories.

Posts that remained uncategorized were embedded using Sentence-BERT and clustered via K-means (k = 5, aligned with the five service categories). Cluster centroids were manually reviewed and mapped to the closest category based on dominant semantic patterns. We evaluated multiple state-of-the-art LLMs for this task, including GPT-4 (Wiggers, 2022), LLaMA 3.3-70B-Instruct (Touvron et al., 2023), Gemini 1.5 Flash (Google, 2023), Mistral 8×7B (Jiang et al., 2023), and Claude Haiku.

Metric	VS	PS	Ex	CGI	CCS	Overall
No. of Posts per Category	1588	1501	2302	1103	1547	8146
No. of Words per Category	103844	102917	164131	73751	102490	557764
No. of Comments per Category	11373	10984	17923	7776	11176	60240
Avg No. of Comments per Post per Category	7.5	7.6	8	7.3	7.5	-
Avg No. of Tokens per Post per Category	72	74	75	69	63	-
Avg No. of Posts per Performer per Category	150	162	120	97	175	-
No. of Performers per Category	11	10	19	12	9	-

Table 1: Statistical summary table of the REDDIX-NET dataset, detailing post counts, word counts, comment volumes, and average engagement metrics across the five defined service categories (VS, PS, Ex, CGI, CCS)

Annotator	Krippendorff(K)	Cohen(C)	Fleiss(F)
(1,2)	0.6633	0.554	—
(1,3)	0.7470	0.681	—
(2,3)	0.6783	0.740	—
(1,2,3)	0.6963	—	0.608

Table 2: Inter-Annotator Agreement Scores

4.2 REDDIX-NET Expression Analysis

This analysis examines user responses to posts to understand how engagement patterns reflect emotional and psychosocial reactions rather than direct mental health outcomes.

We performed sentiment analysis on REDDIX-NET using both pre-trained and GoEmotions-fine-tuned (Demszky et al., 2020) BERT-based models (PLMs) and LLMs (Qwen 2.5 Turbo, GPT-4o). The models predict fine-grained emotion labels, which are aggregated into higher-level psychosocial categories (e.g., sadness/anxiety → mental health concern; joy/trust → positive experience; anger/disgust → exploitation indicators). LLMs extract sentiment polarity (positive, neutral, negative, mixed), emotional spectrum, tonal variation (casual, formal, playful, aggressive), confidence scores, and key phrases. A stratified sample of LLM-generated sentiment outputs was manually reviewed to verify consistency with the intended polarity and emotion labels.

GoEmotions fine-tuned BERT models produced probability distributions across emotion categories, later aggregated into broader psychosocial indicators using predefined affective computing mapping rules. Emotional spectrum, confidence scores, keyphrases, and tonal variation were derived through weighted emotion analysis, normalized classifier certainty, transformer-based attention ranking, and stratified manual validation, with all features stored as aggregated statistics from direct Reddit post responses. The fine-tuned BERT model further evaluates emotional dependency, varied emotional states, exploitation signals, user experience, and mental health-related expressions. Using Hugging Face transformers, BERT maps star ratings (e.g., "5 stars" → satisfaction; "1 star" → aggression) to discrete emotions via a custom dual-mapping framework.

4.3 REDDIX-NET Comments Classification

While sentiment analysis captures emotional tone, it fails to represent interaction intent. Therefore, comments were categorized into 19 thematic buckets (Table 6) derived through an iterative qualitative analysis of representative samples. The taxonomy captures recurring intents, including purchase intent, negotiation, verification, harassment, emotional dependency, and authenticity concerns, while remaining moderation-oriented for analysis of explicit service interactions.

Classification Methodology. Comments were classified using GPT-4 with a few-shot prompting setup (Appendix). Each comment was processed individually, and a multi-label scheme was adopted to account for overlapping intents. Comments not confidently assigned were embedded and clustered, then re-evaluated using the same few-shot prompt to map clusters to predefined buckets.

Handling Ambiguity and Validation. Ambiguous comments received all semantically relevant labels. A stratified random sample validated labeling consistency and taxonomy reliability. Nineteen interaction buckets were derived through grounded, qualitative coding by three annotators, who iteratively merged overlapping themes. Validation included manual review and inter-annotator agreement analysis. The final dataset features store normalized proportions of aggregated comment-label distributions.

4.4 Time-based Analysis on REDDIX-NET

We analyzed metadata to examine temporal engagement trends, tracking fluctuations in posts and comments to identify peak and low activity periods. Hour-of-day analysis revealed engagement cycles

in sexual service-related discussions, offering insights into user behavior, participation patterns, and content visibility dynamics.

5 Results and Analysis

In this section, we will discuss all the results that we have obtained from the experiments that we mentioned in the previous section.

5.1 REDDIX-NET User Classification

Table. 3 summarizes the performance of multiple large language models for classifying posts into predefined service categories using precision, F1-score, accuracy, and error-based metrics including MSE, MAE, and JSD. Performance varies considerably across categories, with Virtual Services (VS) and Physical Services (PS) generally achieving higher F1-scores than Exhibitionism (Ex) and Couples and Group Interactions (CGI), indicating stronger textual separability. GPT-4 demonstrates relatively stable performance across categories, whereas Claude 3.5-Haiku attains the highest F1-score in VS, and LLaMA-3.3-70B-Instruct performs competitively in CCS classifications overall.

The accuracy values appear comparatively high because classification was performed on balanced confidence-solicitation samples from dedicated communities rather than in realistic mixed-content environments. Therefore, F1-scores provide a more reliable estimate of category separability under contextual ambiguity. Preliminary ablation trends, influenced by lexical leakage from repeated promotional phrases, were excluded from the final evaluation. For methodological clarity, exploratory clustering analyses should be separated from supervised classification results. Furthermore, reporting macro-averaged precision, recall, and confusion matrices would improve interpretability and better characterize misclassification behavior.

Cat.	Pre (%)	F1 (%)	MSE	MAE	JSD	Acc (%)
GPT-4						
VS	**45.1**	**62.0**	**0.05**	**0.14**	0.44	85.7
PS	39.3	56.1	0.07	0.18	0.56	81.8
Ex	25.0	39.6	0.06	**0.14**	0.49	**85.8**
CGI	28.6	43.2	0.13	0.20	**0.70**	79.5
CCS	33.3	49.6	**0.05**	0.15	0.48	84.8
Llama-3.3-70B-Instruct						
VS	49.0	64.0	0.05	**0.14**	0.51	84.7
PS	**50.1**	62.2	0.05	0.15	0.51	**85.2**
Ex	43.0	58.0	0.11	0.23	**0.71**	77.4
CGI	42.0	54.0	0.11	0.20	0.67	79.9
CCS	49.0	**67.0**	**0.04**	0.15	0.46	85.1
Mistral 8×7B						
VS	42.0	56.0	**0.03**	**0.12**	0.42	**87.9**
PS	**47.0**	**61.0**	0.06	0.18	0.64	82.2
Ex	40.0	55.0	0.10	0.22	**0.70**	78.4
CGI	41.0	56.0	0.08	0.19	0.63	81.1
CCS	44.0	58.0	0.04	0.13	0.43	86.9
Gemini 1.5 Flash						
VS	**48.0**	63.2	0.05	**0.14**	0.47	85.9
PS	47.5	**64.3**	**0.04**	0.13	0.48	**87.2**
Ex	32.5	48.2	0.06	0.17	0.61	82.7
CGI	31.0	46.2	0.11	0.20	**0.71**	80.1
CCS	39.5	56.1	0.05	0.16	0.48	84.3
Claude 3.5-Haiku						
VS	**52.0**	**66.0**	0.06	0.16	0.54	84.1
PS	46.3	56.9	0.05	**0.13**	0.46	**86.6**
Ex	35.8	51.8	0.08	0.19	0.63	81.0
CGI	36.7	52.4	0.13	0.21	**0.68**	79.3
CCS	37.8	52.4	**0.04**	0.15	0.46	85.5

Table 3: Comparative evaluation of different large language models (LLMs) across various service categories (VS, PS, Ex, CGI, CCS). The models are assessed based on multiple performance metrics, including Precision (Pre), F1-score, Mean Squared Error (MSE), Mean Absolute Error (MAE), Jensen-Shannon Divergence (JSD), and Accuracy. The category is abbreviated as Cat.

5.2 REDDIX-NET Expression Analysis

We analyze user expressions using sentiment, emotion, and tone, employing both LLM- and PLM-based approaches.

Sentiment Analysis: As shown in Figure 2, positive sentiment constitutes the largest proportion across all categories (approximately 40–50%), followed by neutral sentiment (around 30%). Negative and mixed sentiments appear less frequently. This distribution reflects the predominance of positive or neutral expressions in the dataset's user comments. Variations across categories are evident; for example, VS and PS exhibit slightly higher proportions of mixed or negative sentiment than CCS. These observations describe sentiment patterns within REDDIX-NET and do not necessarily generalize beyond this context.

Emotion Type Analysis: Table 4 presents the distribution of emotion categories derived from user comments. Emotions such as desire and joy appear frequently across multiple categories, particularly in VS and Ex. PS shows comparatively higher proportions of emotions such as lust and disgust, while CGI and CCS exhibit more balanced distributions, including neutral expressions. These results reflect model-inferred emotional signals and provide an approximate characterization of user responses within the dataset.

Tonality Analysis: Table 5 illustrates the distribution of tonal expressions. The casual tone is most prevalent across all categories, especially in VS and Ex. PS shows relatively higher propor-

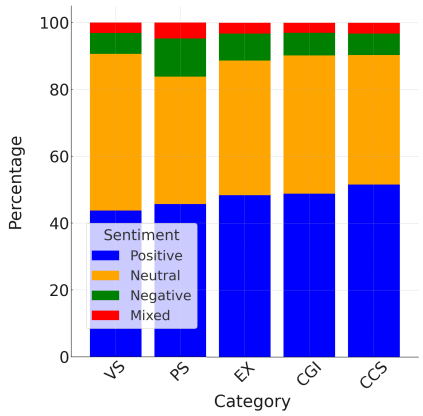


Figure 2: Distribution of sentiment classifications across five different service categories (VS, PS, Ex, CGI, CCS).

Emo(%)	VS	PS	EX	CGI	CCS
Des	31.8	25.3	32.6	27.6	31.3
Joy	31.8	11.4	27.2	22.7	23.7
None	7.8	0.3	11.0	15.1	10.7
Int	9.2	1.5	9.8	10.1	9.2
Sur	5.6	0.1	6.0	6.5	6.9
Ant	4.4	10.1	4.3	5.4	4.0
Lus	2.5	16.4	3.3	2.1	4.6
Ang	1.8	6.3	1.4	3.2	2.9
Exc	1.4	9.5	2.2	2.5	2.1
Ind	1.1	0.6	0.8	1.7	0.9
Pla	0.8	1.4	0.1	1.0	0.8
Dis	0.7	12.6	0.5	0.8	1.4
Inf	0.1	0.1	0.3	1.1	1.0
Frus	1.0	4.4	0.5	0.2	0.5

Table 4: Detailed breakdown percentages of top 14 emotions (Emo) types across five service categories (VS, PS, EX, CGI, CCS). The emotions include Desire (Des), Joy, Interest (Int), Surprise (Sur), Anticipation (Ant), Lust (Lus), Anger (Ang), Excitement (Exc), Indifference (Ind), Playfulness (Pla), Disgust (Dis), Informational (Inf), and Frustration (Frus), along with instances labeled as None.

tions of erotic and playful tones compared to other categories. Formal and neutral tones are consistently present across categories, while aggressive and explicitly sexual tones occur less frequently. These patterns highlight differences in linguistic style across service types.

Cross-Correlation Analysis: Correlation analysis (Figure 6 in Appendix) indicates that relationships between sentiment and tone are generally weak or inconsistent. In contrast, certain emotions align more closely with specific tones (e.g., anger with an aggressive tone, joy with a playful tone). Category-level variations are also observed, with PS showing weaker correlations than other categories. These findings describe associations within the dataset rather than causal relationships.

Tone(%)	VS	PS	EX	CGI	CCS
Cas	62.3	28.6	61.8	61.0	62.8
For	13.8	8.9	14.9	16.3	13.3
Inf	6.9	8.9	6.4	6.3	5.5
Neu	6.2	10.3	7.7	8.1	8.1
Play	5.5	13.6	4.3	3.3	5.2
Agg	1.7	4.2	1.7	2.4	2.0
Ero	1.4	14.5	1.3	1.2	0.9
Flir	1.2	9.1	1.1	0.8	1.5
Se	1.0	1.9	0.9	0.7	0.7

Table 5: Illustration of the distribution of top nine tonal expressions across five service categories. The tones include Casual (Cas), Formal (For), Informal (Inf), Neutral (Neu), Playful (Play), Aggressive (Agg), Erotic (Ero), Flirtatious (Flir), and Sexual (Se).

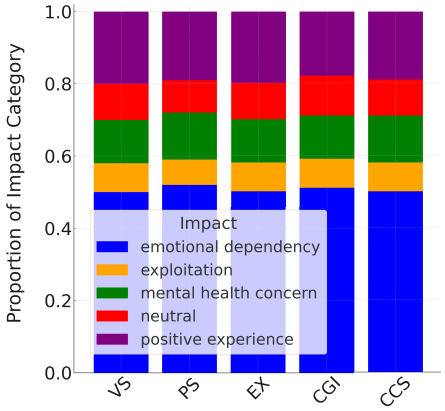


Figure 3: Category-wise impact proportion highlighting the proportion of emotional dependency, exploitation, mental health concerns, neutral perceptions, and positive experiences across different online prostitution categories. This provides insights into the psychosocial expression patterns associated with various engagement types.

5.3 REDDIX-NET Comments Classification

The baskets/categories are defined as: b_{t_1}: Payment or delivery complaints. b_{t_2}: Fantasy and violent demands. b_{t_3}: Legal and ethical concerns. b_{t_4}: Competition or self-promotion. b_{t_5}: Emotional support requests. b_{t_6}: Unclassified comments. b_{t_7}: Price or service negotiations. b_{t_8}: Verification and identity inquiries. b_{t_9}: Specific content requests. $b_{t_{10}}$: External link sharing. $b_{t_{11}}$: Unsolicited requests or harassment. $b_{t_{12}}$: Reviews and recommendations. $b_{t_{13}}$: Service demands (intent to purchase/engage). $b_{t_{14}}$: Skepticism or authenticity questions. $b_{t_{15}}$: Multilingual comments. $b_{t_{16}}$: Positive engagement

Bucket(%)	VS	PS	Ex	CGI	CCS
b_{t_1}	0.4	0.5	0.3	0.7	0.7
b_{t_2}	0.4	0.5	0.9	1.6	1.1
b_{t_3}	0.5	0.9	1.7	1.7	1.8
b_{t_4}	0.7	1.1	0.4	1.9	1.7
b_{t_5}	1.3	1.1	4.5	3.8	3.6
b_{t_6}	1.3	1.8	1.7	0.9	0.7
b_{t_7}	0.5	0.8	0.7	1.9	0.9
b_{t_8}	1.7	2.7	2.8	4.5	6.7
b_{t_9}	12.3	11.9	6.1	3.9	4.5
$b_{t_{10}}$	0.7	3.2	5.0	4.5	4.5
$b_{t_{11}}$	2.7	5.5	4.5	3.2	3.1
$b_{t_{12}}$	5.3	4.6	3.3	6.4	3.6
$b_{t_{13}}$	7.9	3.6	21.2	16.7	20.6
$b_{t_{14}}$	6.2	6.4	7.8	6.4	5.4
$b_{t_{15}}$	1.3	2.0	2.2	9.7	5.4
$b_{t_{16}}$	24.6	27.3	13.9	12.8	14.3
$b_{t_{17}}$	2.6	2.3	3.4	5.2	4.5
$b_{t_{18}}$	19.3	19.3	10.6	7.1	8.1
$b_{t_{19}}$	10.1	4.6	8.9	7.1	8.9

Table 6: This table presents the comments classification of the posts of the different users. Each of the 19 bucket types $b_{t_i} \in \{1, 2, \ldots, 19\}$ captures a distinct user comment or behavior.

(enjoying the post). $b_{t_{17}}$: Self-assertive or confident expressions. $b_{t_{18}}$: Sexual propositions or explicit requests. $b_{t_{19}}$: Ambiguous or multi-response comments.

Table 6 summarizes comment category distributions across service types using a multi-label classification scheme, where percentages are reported independently for each bucket. Positive engagement ($b_{t_{16}}$) is more common in VS and PS, while service demand ($b_{t_{13}}$) is more prominent in Ex, CGI, and CCS. Content requests (b_{t_9}) and explicit propositions ($b_{t_{18}}$) occur more frequently in VS and PS. Lower-frequency categories (e.g., b_{t_1}–b_{t_4}) remain rare across all service types. Overall, the results provide a descriptive overview of interaction patterns and labeled comment behaviors rather than mutually exclusive user intents.

5.4 REDDIX-NET Temporal Analysis

Temporal activity patterns are illustrated in Figures 4 and 5, which show post and comment distributions across hourly intervals. Engagement levels increase from early hours, peak during mid-day to evening periods (approximately 12–19 hours), and decline thereafter. A relatively high comment-to-post ratio is observed during peak periods, indicating increased interaction levels. Since timestamps are aggregated without user location normalization, these patterns represent global activity trends within the dataset rather than region-specific behaviors. Overall, the temporal analysis highlights recurring activity cycles in REDDIX-NET and provides insights into when higher interaction volumes occur within the observed data.

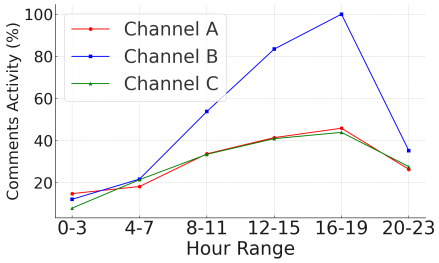


Figure 4: Illustration of comment activity trends over different hourly ranges in a day, highlighting peak engagement times for Channels A, B, and C. The x-axis categorizes days into six time periods, while the y-axis measures the proportion of total posts in each window.

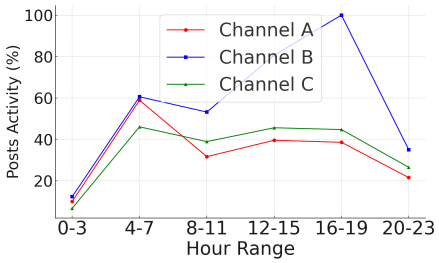


Figure 5: Visualization of temporal posts activity patterns across distinct hourly intervals, emphasizing peak engagement periods for Channels A, B, and C.

6 Conclusion

This study introduces REDDIX-NET, a benchmark dataset for mapping and moderating solicited explicit content on platforms such as Reddit. Using state-of-the-art LLMs, we classify interactions into five behavioral classes through scalable AI-driven moderation. Findings highlight challenges, including evasion tactics and contextual complexity, demonstrating that accurate classification depends on both post content and behavioral patterns derived from sentiment and comment analysis.

Our framework does not make legal judgments, as service legality is jurisdiction-dependent. Instead, it identifies and categorizes unregulated solicitation on mainstream platforms that often violates platform policies and bypasses regulatory oversight. The resulting behavioral classifications, including PS, VS, and CCS, are intended to support moderation and policy enforcement rather than provide definitive legal assessment.

Limitations

This study, while offering a novel dataset and benchmark for online sexual service moderation, is subject to certain limitations. Contextual ambiguity in online discussions makes classification difficult, even for advanced LLMs. While we employed diverse annotation strategies, human bias may still affect labeling. The dataset is Reddit-specific, limiting generalizability to other platforms. Also, the dataset primarily contains explicit-service posts, so model performance on this benchmark should not be interpreted as indicative of real-world moderation accuracy in mixed-content environments where explicit posts represent a small minority. Additionally, evolving evasion tactics pose ongoing challenges for AI moderation, requiring frequent updates.

Future research should focus on several key directions. These include enhancing sentiment analysis with more deterministic methods, integrating AI systems with human-in-the-loop approaches to improve contextual understanding, and further investigating how emotionally charged engagement patterns may relate to psychosocial indicators in online discourse. Our study also offers several important real-life insights, which are discussed in detail in Appendix-6. Crucially, all these analyses done in this paper are confined to data sourced from specific Reddit channels and do not purport to replicate real-world conditions. The findings reflect only the content of this dataset, and interpretations are subject to its inherent limitations and may vary across regional and individual perspectives. Another future work will be to extend this benchmark by incorporating mixed-content datasets with negative examples to evaluate detection performance in real-world moderation scenarios.

Ethics Statement

The ethical dimensions of research concerning online sexual services are multifaceted and deeply sensitive, demanding rigorous safeguards and deliberate ethical oversight. Our study seeks to enhance online safety and inform content moderation strategies while remaining acutely aware of the potential for misuse, exploitation, or harm. To uphold the highest ethical standards, we established a comprehensive framework that emphasizes participant privacy, robust data security, and principled use throughout the research lifecycle.

To promote transparency and responsible engagement, we will release a detailed datasheet alongside the REDDIX-NET dataset. This datasheet outlines the dataset's structure, data collection pipeline, annotation methodology, and usage limitations. We explicitly state that REDDIX-NET is intended strictly for benchmarking and not for model training or any application that could facilitate harm or exploitation.

We prioritized annotator safety and well-being by collaborating with a technical, community-driven organization to recruit and manage our annotation team (One of the authors is a part of this organization). This organization (due to anonymity considerations, we are unable to disclose the organization's name. However, upon request, all relevant details will be shared with authorized personnel) independently oversaw ethical review procedures and secured formal IRB approval for the study. One female annotator was intentionally included in the process to ensure sensitivity toward gender dynamics and to mitigate implicit biases in the annotation of content related to online sexual services. All annotators were clearly informed of the emotionally sensitive nature of the data and were provided with mental health resources and protocols for emotional self-care. Annotators were encouraged to take breaks and access professional support as needed throughout their work.

We implemented a layered, automated anonymization protocol to ensure complete de-identification of users and channels. All usernames, profile links, subreddit identifiers, timestamps, and geolocation metadata were stripped or replaced with generic placeholders such as `[USER]`, `[PROFILE]`, or `[SUBREDDIT]`. Personally identifiable information (PII), including phone numbers, email addresses, and physical locations, was redacted or tokenized as `[INFO REDACTED]`. Images were processed through automated pipelines to blur or crop identifiable regions, removing any visual cues to re-identification. A secondary manual verification step will precede any public release to ensure comprehensive compliance with anonymization requirements.

The dataset was sourced from Reddit, a publicly accessible platform, and all data collection adhered to Reddit's terms of service. We emphasize that despite these precautions, both Reddit and the LLMs used may carry inherent biases. Future users are encouraged to critically evaluate and mitigate such biases in their downstream applications.

REDDIX-NET will be distributed under

a restrictive Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) license, requiring users to accept an ethical usage agreement. Access will be granted only after a valid research use case is submitted and approved via a dedicated form. We release only a curated subset of the dataset, not the full corpus, to minimize risks. Additionally, all dataset access will be logged, ensuring accountability and traceability in case of any ethical breaches or re-identification attempts.

To further reinforce ethical use, we outline a responsible usage checklist:

- **Permitted Uses:** Academic, non-commercial research with recognized IRB or ethical committee approval; development of harm-reduction strategies; bias detection and fairness audits; and responsible studies involving sensitive content.

- **Prohibited Uses:** Attempts to re-identify users or profiles; any activity contributing to sexual exploitation or violating relevant legal frameworks such as the Immoral Traffic (Prevention) Act (1956); or commercial use without explicit written approval.

A clarification is warranted regarding the annotation team: the annotation effort was directly coordinated by one of the authors, a member of the IRB granting organization's research team. All annotators were also members of this organization, and the work was conducted collaboratively with the research team. Although no financial compensation was provided, the annotators volunteered because they were driven by their strong belief in and commitment to the project's objective. The annotators were fully informed of the study's goals and aligned with its ethical standards as outlined in the approved IRB protocol.

By adhering to these principles, we aim to enable responsible research that promotes societal benefit, respects individual dignity, and avoids infringement upon consensual adult expression. This project exemplifies our commitment to ethical innovation and accountable AI deployment in sensitive domains. We acknowledge the use of AI assistants to draft portions of the paper and support related tasks, such as editing and formatting, with all content reviewed and finalized by the authors.

References

Sima Amirkhani, Mahla Fatemeh Alizadeh, Dave Randall, Gunnar Stevens, and Douglas Zytko. 2026. My parents expectations were overwhelming: Online dating romance scams targeting minors in iran through exploitation of parental pressure. *arXiv preprint arXiv:2601.16321*.

Nazanin Andalibi and 1 others. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on reddit. In *Proceedings of the ACM on Human-Computer Interaction*.

Han Bao, Qinying Wang, Zhi Chen, Qingming Li, Xuhong Zhang, Changjiang Li, Zonghui Wang, Shouling Ji, and Wenzhi Chen. 2025. Vmoda: An effective framework for adaptive nsfw image moderation. *arXiv preprint arXiv:2505.23386*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

D.K. Brown. 2018. Reddit's veil of anonymity: Predictors of engagement. *Social Media + Society*.

Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand, and Vivek Gupta. 2026. Better call CLAUSE: A discrepancy benchmark for auditing LLMs legal reasoning capabilities. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5776–5818, Rabat, Morocco. Association for Computational Linguistics.

Manan Roy Choudhury, Anirudh Iyengar Kaniyar Narayana Iyengar, Shikhhar Siingh, Sugeeth Puranam, and Vivek Gupta. 2025. TABARD: A novel benchmark for tabular anomaly analysis, reasoning and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21783–21817, Suzhou, China. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Preprint*, arXiv:2005.00547.

Maria Diaz and Anand Panangadan. 2020. Natural language-based integration of online review datasets for identification of sex trafficking businesses. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 259–264. IEEE.

Christopher J Ferguson and Richard D Hartley. 2022. Pornography and sexual aggression: Can meta-analysis find a link? *Trauma, Violence, & Abuse*, 23(1):278–287.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Luca Giommoni and Ruth Ikwu. 2021. Identifying human trafficking indicators in the uk online sex market. *Trends in Organized Crime*, pages 1–24.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://arxiv.org/abs/2312.11805.

Kelly Caroline Gothard. 2021. *The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community*. The University of Vermont and State Agricultural College.

Government of India. 2000. The information technology act, 2000.

Government of NCT of Delhi. 2023. Implementation of immoral traffic (prevention) act.

Vivek Kumar Gupta. 2025. The digital childhood dilemma: Reconciling children's rights, online safety and legal safeguards against exploitation in an era of cyber vulnerabilities. *Journal of Teachers and Teacher Education*, 2(1):01–11.

Ehsan Ul Haq and 1 others. 2025. Exploring self-disclosure norms and engagement dynamics on reddit. *arXiv preprint arXiv:2502.10701*.

Michelle Ibanez and Dan Suthers. 2016a. Detecting covert sex trafficking networks in virtual markets. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 876–879.

Michelle Ibanez and Daniel D Suthers. 2016b. Detecting covert sex trafficking networks in virtual markets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 876–879. IEEE.

Shalini Jangra, Zaid Almahmoud, Suparna De, Gareth Tyson, Ehsan Ul Haq, and Nishanth Sastry. 2025. Understanding the complexities of responsibly sharing nsfw content online. *arXiv preprint arXiv:2511.15726*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Burcu B Keskin, Gregory J Bott, and Nickolas K Freeman. 2021. Cracking sex trafficking: Data analysis, pattern recognition, and path prediction. *Production and Operations Management*, 30(4):1110–1135.

Xian Li, Yuanning Han, Di Liu, Pengcheng An, and Shuo Niu. 2026. When generative ai is intimate, sexy, and violent: Examining not-safe-for-work (nsfw) chatbots on flowgpt. *arXiv preprint arXiv:2601.14324*.

Claudio Marche, Ilaria Cabiddu, Christian Giovanni Castangia, Luigi Serreli, and Michele Nitti. 2023. Implementation of a multi-approach fake news detector and of a trust management model for news sources. *IEEE Transactions on Services Computing*.

Ministry of Electronics and Information Technology, Government of India. 2021. Information technology (intermediary guidelines and digital media ethics code) rules, 2021.

Kimberly J Mitchell, David Finkelhor, and Janis Wolak. 2003. The exposure of youth to unwanted sexual material on the internet: A national survey of risk, impact, and prevention. *Youth & Society*, 34(3):330–358.

Jochen Peter and Patti M Valkenburg. 2016. Adolescents and pornography: A review of 20 years of research. *The Journal of Sex Research*, 53(4-5):509–531.

Mary E. PettyJohn and 1 others. 2025. Teens seeking information and advice about sexual behaviors on reddit. *Journal of Adolescent Health*.

Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, 16(3):1–34.

Skye Roisen P Sanchez and Genesis G Genelza. 2025. A systematic literature review on sexual exploitation and abuse of children: Prevalence, risk factors, and societal responses. *Universe International Journal of Interdisciplinary Research*, 5(10):207–219.

M. Shelton. 2015. A qualitative study of disclosure within and beyond reddit.

Ayliana Teitelbaum. 2020. *The Ethics of Reddit and an Artificial Moral Compass*. Ph.D. thesis, New York, NY. Stern College for Women. Yeshiva University.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 21 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, and 1 others. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Longshaokan Wang, Eric Laber, Yeng Saanchi, and Sherrie Caltagirone. 2020a. Sex trafficking detection with ordinal regression neural networks. *Preprint*, arXiv:1908.05434.

Longshaokan Wang, Eric Laber, Yeng Saanchi, and Sherrie Caltagirone. 2020b. Sex trafficking detection with ordinal regression neural networks. *arXiv preprint arXiv:1908.05434*.

Kyle Wiggers. 2022. Gpt-2 output detector demo. https://openai-openai-detector.hf.space/. [Online; accessed 2023-01-02].

Fan Yang, Yihao Huang, Jiayi Zhu, Ling Shi, Geguang Pu, Jin Song Dong, and Kailong Wang. 2025. Seeing it before it happens: In-generation nsfw detection for diffusion-based text-to-image models. *arXiv preprint arXiv:2508.03006*.

Appendix-1: Prompts

Classification/Evaluation Prompt

Classification Prompt

Here is the content of a post. It has the following attributes:
1. Title: "title"
2. Image Description: "image_description"
3. Comments: "comments"

Your task is to classify this post into one of the following five categories of services related to online prostitution: 1. Physical Services: Posts offering in-person sexual services or physical prostitution. 2. Virtual Services: Posts offering virtual interactions such as video calls, virtual sex, or promoting platforms like OnlyFans. 3. Exhibitionism: Posts showcasing exhibitionistic behavior, such as public displays or other forms of showcasing oneself. 4. Content Creation and Sales: Posts promoting or selling photos, videos, or other content without direct interaction. 5. Couples and Group Interactions: Posts seeking interactions involving couples, threesomes, or group scenarios. 6. Miscellaneous Fun/Exploration: Posts describing non-specific fun, exploration, or interactions that do not fall into the other categories.
Carefully analyze the title, image description, and comments. Then determine which of the five categories the post best fits. Respond with only the category name. If the information is insufficient, respond with 'Uncategorizable'. Return only the category name or "Uncategorizable" based on your analysis, in the response.

Hyperparameters

Key hyperparameters we used in our experiments include the temperature, which controls the randomness of predictions and is typically set between 0.2 and 0.5 to ensure a more deterministic setting, and the max token length, chosen based on the average post length. Additionally, the models are fine-tuned with a learning rate range of 1e-5 to 1e-3 and a batch size of 16-64, with the number of training epochs determined by the loss function's convergence. During inference, models may use nucleus sampling (top-p) with a probability threshold of 0.9. The evaluation metrics for this task include Distribution Accuracy, Accuracy, F1 Score, Precision, Mean Absolute Error (MAE), and Shannon Entropy of the distribution.

Prompt 1 instructs an AI to classify a post into one of five distinct categories related to online prostitution services based on the post's title, image description, and comments. It delineates clear definitions for each category, ranging from physical and virtual services to exhibitionism, content creation, couples and group interactions, and miscellaneous fun/exploration, ensuring that the classification process is both structured and comprehensive. The prompting strategy emphasizes a careful, context-based analysis of the provided attributes and requires the AI to return only the appropriate category name or "Uncategorizable" if the information is insufficient, thereby promoting precise, deterministic decision-making in the classification process.

Sentiment Analysis Prompt

Sentiment Analysis

You are an expert AI performing sentiment analysis.
Analyze the following text and provide the following insights: 1. Sentiment: Positive, Neutral, or Negative, with a confidence score (0-1). 2. Emotion Classification: Identify the dominant emotion (e.g., joy, anger, sadness, surprise, etc.). 3. Keywords: Extract the main keywords or phrases relevant to the context. 4. Tone: Determine the tone (e.g., formal, casual, playful, persuasive, etc.).

Output Format:

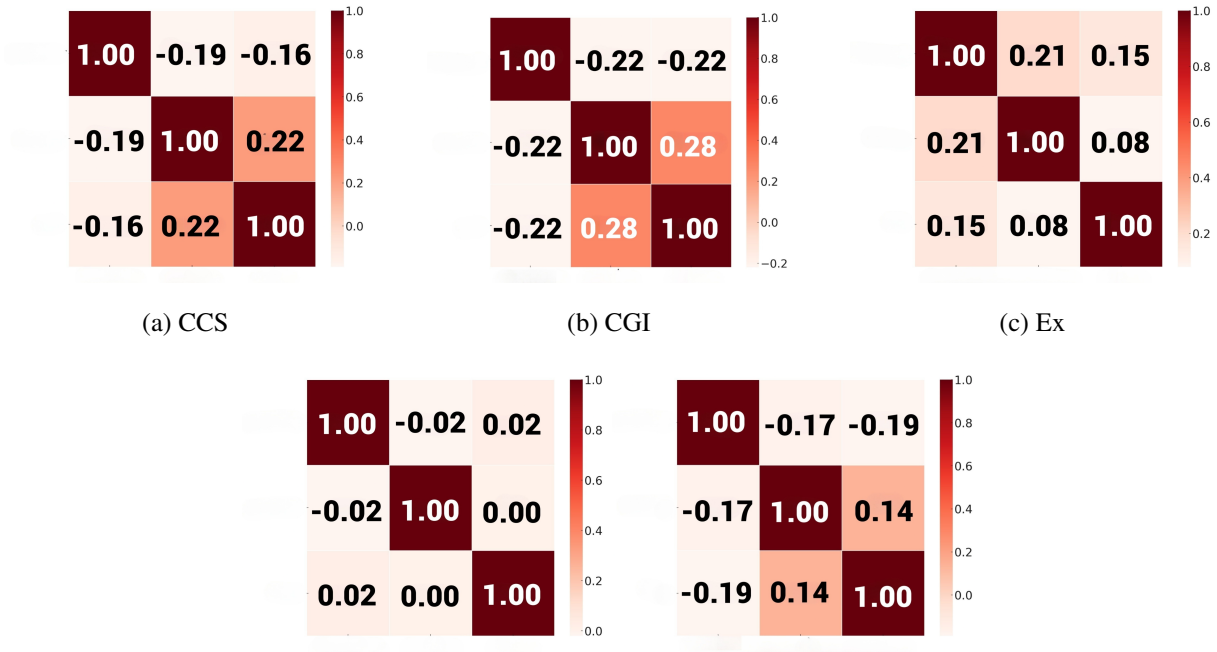
- Sentiment: [label], Confidence: [score]
- Emotion: [emotion]
- Keywords: [keywords]
- Tone: [tone]

The provided sentiment analysis prompt instructs an expert AI to perform a comprehensive evaluation of a given text by extracting multidimensional insights. Specifically, it requires the AI to determine the overall sentiment - positive, neutral, or negative while providing a confidence score, classify the dominant emotion (such as joy,

anger, or sadness), extract key phrases or keywords pertinent to the context, and assess the tone (e.g., formal, casual, or playful) of the text. The prompt also specifies a structured output format, ensuring results are returned consistently and in a standardized manner. This prompting strategy is designed to facilitate a detailed, context-aware analysis that leverages both qualitative and quantitative dimensions, thereby enhancing the interpretability and reliability of the sentiment analysis process.

Comments Classification Prompt

```
You are an expert content analyst. For each
comment provided, classify it into exactly one of
the following 18 classes:

1. Users Who Are Enjoying the Post and Its
Contents (Engagement & Positive Sentiment)
   - "Wow! You look absolutely stunning!"
   - "Absolutely mesmerizing, I'm hooked!"
   - "Stunning visuals, keep up the great work!"

2. Users Who Are Demanding Such Services (Intent
to Purchase/Engage)
   - "How much do you charge for this service?"
   - "Where can I reach you for more details?"
   - "Are you available for a private session?"

3. Users Who Are Requesting Specific Content
(Content Demand Trends)
   - "Can you do a video in a red dress?"
   - "Id love to see more dance moves from you!"
   - "Could you post more outdoor shoots?"

4. Users Who Are Skeptical or Questioning
Authenticity (Trust & Credibility Issues)
   - "Is this actually you or just edited?"
   - "Has anyone actually met her? Looks fake."
   - "Seems too polished—are these authentic?"

5. Users Who Are Providing Reviews &
Recommendations (Word-of-Mouth & Service
Feedback)
   - "Shes super professional and amazing to work
   with!"
   - "Had a great time, shes very professional!"
   - "Overpriced, not worth it."

6. Users Who Are Discussing Legality & Ethics
   - "Isn't this kind of thing banned here?"
   - "Should this even be allowed on this
   platform?"
   - "Im concerned about the legality of this
   content."

7. Users Who Are Competing or Self-Promoting
Services
   - "Check out my profile if you like this!"
   - "I offer exclusive content at a discount!"
   - "I can do this for half the price. DM me!"

8. Users Who Are Negotiating Prices or Services
   - "Can you do this for \$40 instead?"
   - "Is there any discount if I book multiple
   sessions?"
   - "How about a special rate for returning
   customers?"

9. Users Who Are Complaining About Payment or
Delivery Issues
   - "I paid but never got my order!"
   - "She stopped replying after I sent the
   payment!"
   - "This is a scam, don't fall for it!"

10. Users Who Are Making Unsolicited Requests or
Harassment
    - "Send me something for free first!"
    - "I'll find you if you don't reply!"
    - "Do this for me, or else!"

11. Comments that are Multi-Lingual
    - "Teri nazaron mein vo jadu hai."
    - "Include comments from languages which are
    not English."
    - "acha hai"

12. Comments that are Fantasy and Violent Demands
    - "Show me an online act that blends erotic
    fantasy with a violent edge."
    - "I demand you to enact a dark fantasy scene
    with intense aggression."

13. Comments that are Emotional Support
    - "Your posts always brighten my day!"
    - "I appreciate your openness; it helps me
    feel less alone."

14. Comments that are Verification and Identity
Inquiries
    - "Is this really you or just an
    impersonator?"
    - "Can you prove that this is your real
    account?"

15. Comments that are External Link Sharing
    - "Check this link out for exclusive content:
    [external link]"
    - "Visit my page for more: [link]"

16. Comments that are Illicit Propositions /
Explicit Requests
    - "Can you send me private pics? Ill pay
    extra."
    - "Do you do custom videos with nudity?"

17. Comments that are Self-Assertive/Confidence
Expressions
    - "Im the best at what I do, no one compares!"
    - "I always get what I want, and this is no
    different."

18. Comments that are Ambiguous or Multi-Response
Comments
    - "I'm not sure what to think about this..."
    - "Interesting... I wonder whats really going
    on."

For each comment provided, classify it into
exactly one of the above categories and return
the output as a JSON object with each original
comment as a key and its classification as the
value.
```



(a) CCS (b) CGI (c) Ex

(d) PS (e) VS

Figure 6: Correlation heatmaps depicting the interplay between sentiment, emotion, and tone across multiple dataset categories. Darker shades denote stronger correlations, while lighter hues indicate weaker associations.

Appendix-2: Sentimental Analysis

Sentimental Analysis using LLMs

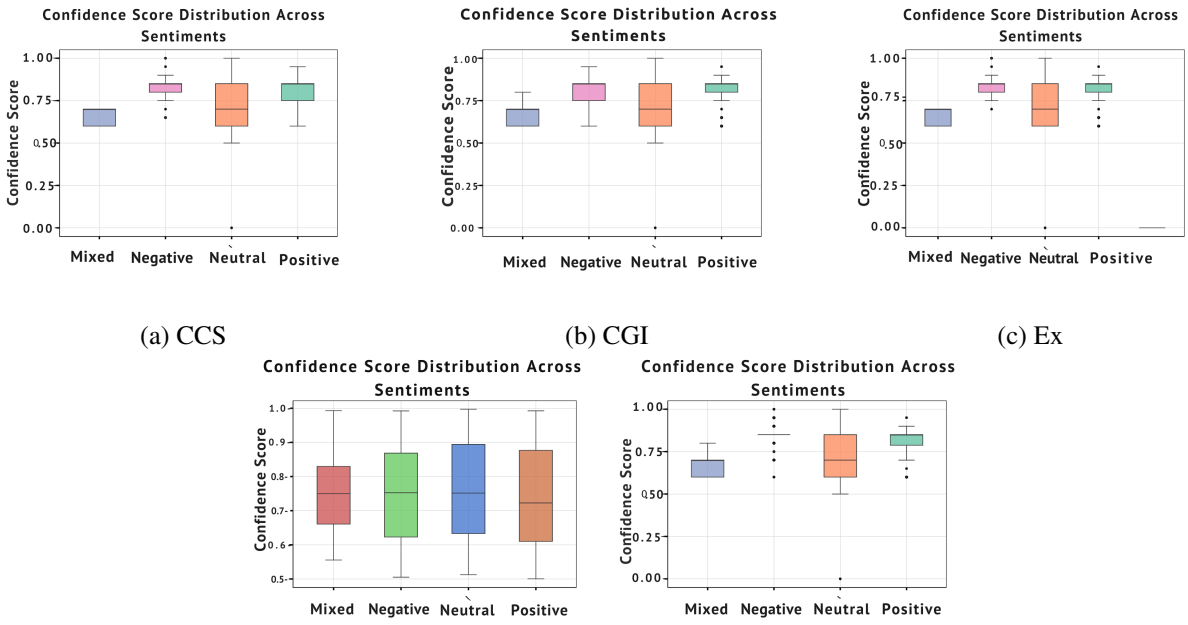
The above prompt positions the AI as an expert content analyst, using a persona-based approach to accurately classify user comments into 18 predefined categories. This structured prompt incorporates few-shot examples for each category, ranging from positive engagement and service inquiries to external link sharing and ambiguous expressions, thereby guiding the AI with concrete instances of desired outputs. By instructing the AI to evaluate each comment and return a JSON object mapping each comment to its classification, the strategy leverages contextual cues and demonstration-based learning to ensure consistent, precise categorization.

Distribution of Tone Types Across Sentiment Types: Based on the two nested pie charts (Figure 7) showing sentiment analysis alongside emotions and tone distribution, here's a comprehensive analysis: The sentiment distribution in both charts shows a predominantly positive and neutral outlook, with 48% positive sentiment the largest segment, followed by 42.9% neutral and 7.21% negative. This indicates that the overall communication style in couples and group interactions tends to maintain a constructive, balanced emotional atmosphere, with very few instances of negative exchanges.

Looking at the emotional aspects in the first chart, Desire (29.4%) and Joy (23.9%) emerge as the dominant emotions, collectively accounting for over half of the emotional expressions. This is followed by a notable segment of "None" (14.8%) and "Interest" (9.03%), suggesting that while interactions are generally emotionally engaged, there are also periods of neutral or emotionally reserved communication. The presence of other emotions, such as Anticipation, Surprise, and Excitement, in smaller proportions indicates a rich diversity of emotional expression, though negative emotions like Anger remain minimal (3.04%).

The tone analysis in the second chart provides interesting insights into the communication style: a Casual tone strongly dominates at 65.7%, followed by a Formal tone at 17.2%. This suggests that most interactions maintain a relaxed, comfortable atmosphere while still preserving some level of formality when needed. The presence of Neutral (7.89%), Informal (5.26%), and Playful (2.28%) tones, with minimal Aggressive tone (1.67%), indicates that the communication environment is generally conducive to open and comfortable interaction while maintaining appropriate boundaries and respect.

Confidence Score Distribution by Sentiment: From Figure. 7, the four sentiment categories (Pos-



(a) CCS (b) CGI (c) Ex

(d) PS (e) VS

Figure 7: The Confidence Score Distribution by Sentiment: Box Plot Comparison of Positive, Neutral, Mixed, and Negative Categories

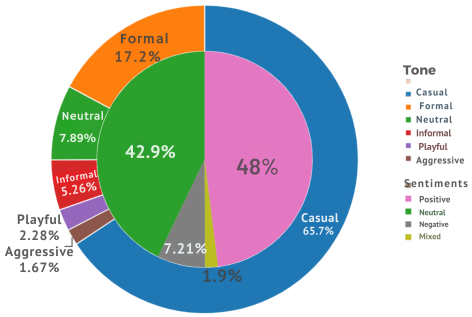
itive, Neutral, Negative, Mixed). Overall, each box plot reveals moderate to high median confidence values, suggesting that the underlying model generally assigns sentiment labels with a notable degree of certainty. However, the presence of outliers and varying interquartile ranges across subplots indicates that classification confidence can fluctuate depending on the specific context or linguistic cues in the data.

A closer inspection of the individual subplots highlights subtle differences in how sentiments are classified. In contrast, other classes (e.g., (a) CCS and (c) Ex) exhibit a broader spread for Neutral or Mixed sentiments, suggesting that the model occasionally encounters more ambiguity when distinguishing between emotionally neutral content and text that blends multiple affective tones. Negative sentiment typically shows a slightly wider distribution, suggesting potential variability in the strength with which negative cues are detected.

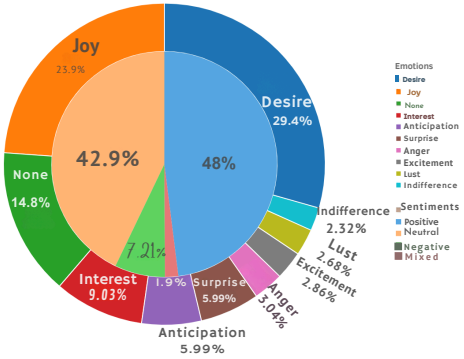
Collectively, these findings underscore a robust, yet context-sensitive classification process. While Positive sentiment often emerges with higher, more consistent confidence scores, Neutral, Mixed, and Negative categories reveal more diverse confidence intervals, reflecting the nuanced nature of human language and emotional expression. The recurring outliers across subplots further emphasize that certain instances may challenge the model's ability to assign a definitive sentiment category. Overall, the distribution of confidence scores across these five classes illustrates a generally reliable classification framework, albeit one that must navigate the inherent complexities of the sentiment-laden text.

Sentimental Analysis using PLM (BERT)

In reviewing Figure 9 describing the sentiment proportions across five categories, a clear trend emerges. Positive sentiments such as "appreciation" and "satisfaction" account for a substantial share across most categories, indicating that user feedback skews favorably. "Neutral" sentiment also appears consistently, though at varying levels, suggesting a notable fraction of content that neither leans strongly positive nor negative. In contrast, "aggression" and "frustration" are relatively lower, suggesting that overtly negative expressions are less common in the overall dataset.



(a) Sentiment and Tones Distribution across sentiment types for different classes in the dataset for CGI Category.



(b) Sentiment and Emotions Distribution across sentiment types for different classes in the dataset for CGI Category.

Figure 8: Distribution of tone and emotions across sentiment types for different classes in the dataset.

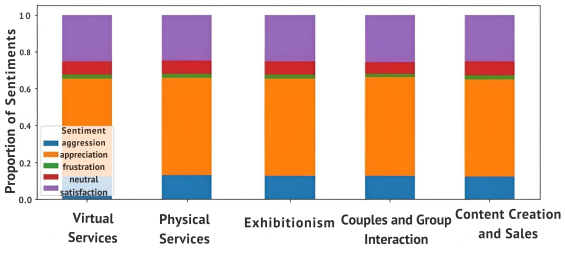


Figure 9: Sentiment Analysis using BERT Model

A closer look reveals subtle differences in sentiment composition among the categories. For example, Physical Services, Content Creations, and Sales exhibit a higher prevalence of "appreciation," indicating more frequent expressions of gratitude or praise. These observations are derived from a BERT-based model that leverages contextual embeddings to classify text with a high degree of nuance. Consequently, the analysis highlights both the generally positive nature of user communications and the importance of contextual factors in shaping sentiment.

Appendix-3: REDDIX-NET Metadata Analysis

Figure 10 provides a clear visual summary of how performers are distributed and overlap across multiple categories, including Miscellaneous Fun, Content Creation and Sales, Physical Services, Virtual Services, Couples and Group Interaction, and Exhibitionism. Each row corresponds to a category, and the black dots indicate shared performers among these categories. The bar chart at the top shows the number of performers participating in each specific combination of categories. Notably, "Exhibitionism" and "Couples and Group Interaction" are the most prevalent, as evidenced by the tallest bars, suggesting their high popularity or frequent reporting. Overall, the figure underscores that while many performers concentrate on a single category, a noteworthy subset engages in multiple overlapping areas, highlighting the importance of cross-category involvement. Looking at the intersection sizes in the top bar chart, it is evident that most performers participate in only one or two categories. However, a distinct group of five performers is active across a broader range of categories, indicating significant overlap in their services and interactions. Additionally, other smaller clusters—such as a group of three performers—reveal that although single-category involvement is common, a considerable minority diversifies their participation.

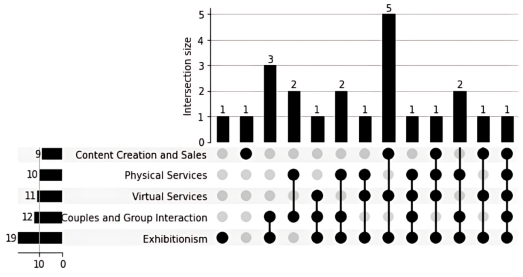


Figure 10: Total No. of performers per category and their overlap

Appendix-4: Labelling Process

The dataset was annotated through a structured labeling process executed by a team of three trained annotators (two male, one female). The

composition of the annotation team was deliberately designed to include a female perspective, a methodological choice intended to mitigate potential gender-based observational biases. This ensures a more balanced and representative interpretation of the illicit services offered across different genders within the corpus.

The annotation workflow proceeded as follows: Each annotator was assigned a subset of user profiles and their associated activities within the designated subreddits. The fundamental unit of analysis was the user's explicit service offering, distilled from their posts, comments, and profile descriptions. Annotators were tasked with performing a qualitative analysis of this content to identify the nature of the advertised service and classify it according to a predefined annotation schema.

This schema was developed iteratively based on a preliminary content analysis of the data. It consists of five mutually exclusive categories that encapsulate the primary types of illicit services observed. The categories are formally defined as:

- **Content Creation and Sales (CCS):** This category encompasses the production and sale of digital media. It includes, but is not limited to, the sale of pre-made or custom photosets, video clips, and subscriptions to platforms like OnlyFans or Fansly, where explicit content is monetized. The key differentiator is the transactional nature of acquiring a static or pre-recorded media product.

- **Couples and Group Interactions (CGI):** This label applies to services that explicitly involve more than one participant, such as live performances by couples, or services marketed towards groups. This category is distinct from individual services and highlights collaborative or multi-person offerings.

- **Exhibition (Ex):** This category is defined by performative acts where the service is a live or public display for an audience. This primarily includes real-time webcam services (camming) or other forms of live, voyeuristic exhibition where the interaction is centered on the act of being watched.

- **Physical Services (PS):** This label is assigned to any service that requires direct, in-person physical contact or meetups. Annotators identified these services through explicit keywords related to location, availability for "in-call" or "out-call" appointments, and other language indicating a non-virtual transaction.

- **Virtual Services (VS):** This category covers interactive, one-on-one services conducted remotely. Examples include synchronous activities such as sexting, live video calls, and "Girlfriend/Boyfriend Experience" (GFE/BFE) simulations that occur in real time but without physical presence. This is distinct from CCS as the service is an interactive experience rather than a media product.

Upon completion of the independent annotation phase, the labels were compiled for a quantitative reliability assessment. To validate the consistency and reproducibility of our schema and the annotators' judgments, we calculated Inter-Annotator Agreement (IAA) scores. This analysis was conducted for all possible annotator pairs (1,2), (1,3), and (2,3), as well as for the complete triad of annotators (1,2,3) to provide a comprehensive measure of concordance across the dataset.

Procedure Involving Data Collection and Construction

1. **Data Collection** The data for this study was collected from three specific subreddits identified as primary hubs for discussions related to illicit services. Data extraction was performed using the Reddit API, facilitated by the PRAW (Python Reddit API Wrapper) library, which enabled the retrieval of both posts and comments from these subreddits.

2. **Data Cleaning** The initial dataset underwent cleaning to remove irrelevant or extraneous content. Posts and comments deemed non-substantive, such as greetings (e.g., "Hi," "Hello!"), were removed to ensure the dataset focused solely on meaningful exchanges related to the research topic.

3. **Data Preprocessing** To protect the anonymity of individuals involved, several preprocessing steps were implemented. Posts containing visible faces were excluded from the dataset, as most posts naturally blurred such identifying features. Additionally, all usernames and Reddit IDs were stripped from the data, retaining only the content of the posts and comments for analysis.

Appendix-5: Ablation studies on feature importance

The evaluation is summarized in the table. 7 provides insights into the relative contribution of different feature sets to classification performance.

Exclusion of Emotion Features: The model maintains near-perfect performance (approximately 0.99 across Accuracy, Precision, Recall, and F1 Score) even when emotion-related features are removed. This indicates that emotion features are complementary rather than strictly required for classification accuracy, suggesting that other expressive signals already capture sufficient discriminative information for the core prediction task.

Use of Individual Feature Sets in Isolation: When the model is trained using only sentiment features, only emotion features, or only tone features, performance drops substantially, with metric values ranging between approximately 0.07 and 0.14. This demonstrates that while each feature set contains a useful signal, none is sufficient on its own to support robust classification. Effective performance, therefore, depends on combining multiple expressive cues.

Exclusion of Comment Features: The removal of comment-related features results in extremely poor performance (approximately 0.07 across metrics), indicating a near-total failure of classification. This sharp decline underscores the central role of comment-derived contextual information in distinguishing service categories, highlighting that engagement context is foundational to the model's predictive capability.

Relative Contribution of Expressive Features: The ablation results indicate that expressive features contribute unequally to model performance. Removing sentiment, tone, or metadata features leads to substantial degradation in classification accuracy, whereas removing emotion features alone does not significantly affect performance in this experimental setup (Accuracy 0.99, F1 0.99). This suggests that sentiment polarity and tonal cues serve as the primary discriminative signals for category prediction, while emotion features play a complementary role, supporting psychosocial interpretation rather than being strictly necessary for classification accuracy.

Feature Set	Accuracy	Precision	Recall	F1 Score
No Sentiment Features	0.03	0.04	0.03	0.03
No Emotion Features	0.99	0.98	0.99	0.99
No Tone Features	0.02	0.03	0.02	0.01
No Comment Features	0.07	0.06	0.09	0.06
No Metadata Features	0.03	0.03	0.02	0.03
Sentiment Features Only	0.14	0.12	0.13	0.11
Emotion Features Only	0.07	0.08	0.07	0.07
Tone Features Only	0.09	0.09	0.08	0.09

Table 7: Table representing ablation study on feature importance.

Overall, these findings indicate that REDDIXNET benefits from a balanced combination of contextual comment features and LLM-derived expressive signals, particularly sentiment and tone. Together, these feature groups provide the strongest predictive contribution, while emotion features enhance interpretability and support downstream psychosocial analysis.

Appendix-6: Psychological and Social Implications Study

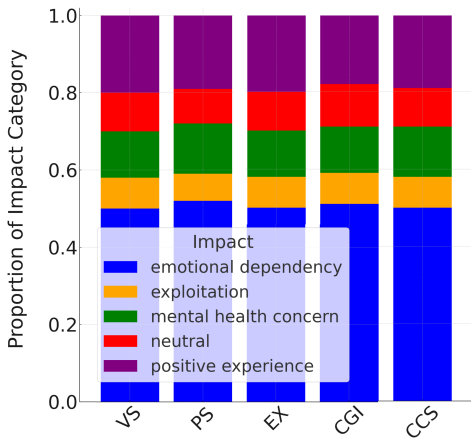


Figure 11: Category-wise impact proportion highlighting the proportion of emotional dependency, exploitation, mental health concerns, neutral perceptions, and positive experiences across different online prostitution categories. This provides insights into the psychosocial expression patterns associated with various engagement types.

For interpretability, emotion predictions were grouped into five impact categories: emotional dependency, exploitation indicators, mental health concerns, neutral perception, and positive experience, which form the basis of Figure 11. The aggregation into higher-level psychosocial impact categories was performed using a few-shot LLM-based classification setup, where emotion predictions and contextual comment text were provided to the LLM to assign one or more impact categories. A strati-

Model	VS				PS				Ex				CGI				CCS			
	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN
LLaMA	0.19	0.01	0.60	0.20	0.19	0.01	0.61	0.19	0.19	0.01	0.55	0.25	0.18	0.02	0.55	0.3	0.2	0.0	0.6	0.2
GPT-4	0.20	0.00	0.56	0.24	0.20	0.00	0.49	0.31	0.20	0.00	0.21	0.59	0.19	0.01	0.31	0.5	0.2	0.0	0.4	0.4
Gemini	0.19	0.01	0.59	0.21	0.20	0.00	0.58	0.22	0.20	0.00	0.39	0.41	0.20	0.00	0.36	0.4	0.2	0.0	0.5	0.3
Claude	0.19	0.01	0.62	0.18	0.18	0.02	0.59	0.21	0.20	0.00	0.45	0.35	0.20	0.00	0.46	0.3	0.2	0.0	0.5	0.3
Mistral	0.19	0.01	0.54	0.26	0.19	0.01	0.59	0.21	0.19	0.01	0.51	0.29	0.19	0.01	0.52	0.3	0.2	0.0	0.6	0.2

Table 8: True positive, False positive, True negative, and False negative for each model across categories.

fied subset of the automatically assigned labels was manually verified by annotators to ensure consistency and reliability. Figure 11 presents the distribution of impact categories across service types (VS, PS, EX, CGI, CCS). The stacked bar chart shows that emotional dependency (blue) makes up the largest proportion across all categories, consistently occupying nearly half of each bar. Exploitation (yellow) and mental health concern (orange) appear in smaller proportions, with exploitation being slightly higher in EX and PS compared to other categories. Neutral perception (red) is relatively minor but present across all categories at comparable levels. Positive experience (purple) consistently accounts for the second-largest segment, reaching close to one-third in some categories, such as CCS and CGI. Overall, the chart highlights that emotional dependency and positive experiences dominate the distribution, while exploitation, mental health concerns, and neutral perceptions remain comparatively lower.

Note: We emphasize that these categories represent inferred psychosocial indicators derived from linguistic expression and should not be interpreted as clinical assessments of users' mental health.

Appendix-7: Additional Observational Insights

This appendix presents additional observational insights derived from expressive and engagement patterns within REDDIX-NET. These observations are based on aggregated linguistic signals and user interaction behavior and should be interpreted as indicators of expressed psychosocial reactions rather than direct measures of psychological or mental health outcomes.

1. Engagement Intensity Across Service Categories. Different service categories exhibit distinct engagement dynamics. Categories such as Exhibitionism (Ex) and Virtual Services (VS) show higher proportions of emotionally expressive and positively engaged comments, suggesting stronger interaction intensity. These patterns reflect differences in how users linguistically respond to different solicitation types.

2. Expressive Signals and Interaction Context. Sentiment, tone, and emotion distributions reveal variation in how users frame their interactions. For example, certain categories demonstrate elevated levels of desire, joy, or transactional tone, while others show more neutral or skeptical expressions. These expressive signals function as indicators of conversational context rather than evidence of underlying psychological states.

3. Psychosocial Indicator Patterns. The aggregated impact categories (emotional dependency, exploitation indicators, mental health-related expressions, neutral perception, and positive experience) reflect patterns in how users articulate reactions within comments. These categories capture observable linguistic markers associated with engagement behavior and do not imply clinical diagnosis or causal psychological effects.

4. Platform-Level Behavioral Rhythms. Temporal engagement patterns highlight cyclical activity in comment and post interactions. These patterns reflect platform-level usage rhythms and collective participation dynamics rather than region-specific or individual behavioral conclusions.

5. Moderation-Oriented Implications. The expressive and engagement patterns identified in this study may assist platform moderators and policy designers in understanding how solicitation posts generate interaction signals. Such insights can inform the development of context-aware moderation tools without implying direct psychological impact assessment.

Overall, these findings provide descriptive and computational insights into how users linguistically engage with explicit-service content within dedicated communities on a mainstream platform. They should be interpreted as platform-level behavioral observations rather than claims about individual-level psychological consequences.

Metric	G-L	G-C	G-M	L-C	L-M	C-M	G-L-C	G-L-M	G-C-M	L-C-M
Pre (%) VS	70.59	70.59	70.59	61.33	61.33	65.66	67.89	72.12	74.23	63.40
F1 (%) VS	15.89	15.89	15.89	22.44	22.44	26.08	19.00	19.38	18.77	23.57
Pre (%) PS	53.11	53.11	53.11	52.59	52.59	52.02	52.88	54.51	54.01	52.13
F1 (%) PS	47.42	47.42	47.42	51.71	51.71	51.87	51.40	47.15	47.51	50.52
Pre (%) Ex	71.10	80.45	63.50	75.88	69.83	67.03	74.63	70.87	83.98	90.69
F1 (%) Ex	40.10	40.10	40.10	17.11	17.11	20.52	28.44	33.59	36.61	17.76
Pre (%) CGI	85.00	85.00	85.00	86.18	86.18	86.86	85.61	86.54	86.01	87.95
F1 (%) CGI	17.19	17.19	17.19	25.17	25.17	23.20	22.14	25.84	23.84	27.68
Pre (%) CCS	58.47	58.47	58.47	60.00	60.00	59.26	61.96	55.64	56.78	59.78
F1 (%) CCS	19.03	19.03	19.03	14.74	14.74	5.05	16.31	20.00	18.48	15.74
MSE	0.464	0.464	0.464	0.486	0.486	0.475	0.471	0.465	0.461	0.482
MAE	0.464	0.464	0.464	0.486	0.486	0.475	0.471	0.465	0.461	0.482
JSD	0.128	0.128	0.128	0.142	0.142	0.144	0.134	0.126	0.125	0.139
Acc (%)	62.43	62.43	62.43	55.88	55.88	56.10	60.19	62.28	63.24	57.07

Table 9: Evaluation Results of aggregation of LLMs (ensemble methods). Here, G→ Gemini 1.5 Flash, L→ LlaMA 3.3-70B-Instruct, M→ Mistral 8×7B, Q→ Qwen 2.5 Turbo, C→ Claude 3.5 Haiku.

Appendix-8: Ensemble Ablation Analysis

An evaluation of dyadic (two-model) and triadic (three-model) ensembles was conducted to measure the impact of model aggregation. The results, shown in Table 9, highlight several key findings regarding the performance of these combined models.

The analysis reveals that triadic combinations consistently outperform dyadic pairs, particularly in overall accuracy and F1 scores. The G-C-M (Gemini-Claude-Mistral) ensemble is the top performer, achieving the highest accuracy (63.24%) and the lowest Jensen-Shannon Divergence (JSD) of 0.125, which indicates the best alignment with the ground truth distribution. This configuration also registered the lowest overall error (MSE/MAE of 0.461). A notable trade-off between precision and recall was observed for the 'Exhibition' (Ex) category, where some ensembles maintained perfect precision but at the cost of a significantly lower F1 score.

In summary, aggregating three diverse models, especially the G-C-M combination, yields more robust and accurate predictions than simpler two-model ensembles. Additionally, we have conducted an ablation study on feature importance, focusing on sentiment, emotion, and tone, which yielded valuable insights detailed in Appendix 5.

Note: It is important to note that the above observations are based on a limited and domain-specific dataset; therefore, the interpretations should be viewed as preliminary hypotheses rather than definitive conclusions.

Appendix-9: Error Analysis

We performed a comprehensive error analysis, summarized in the table. 8 (for our multi-label classification task across six distinct categories, assessing performance based on per-category True Positives, False Positives, False Negatives, and True Negatives. A predominant challenge across all models was the high rate of FNs, largely due to the failure to interpret subtle, coded, and slang-based language. For instance, Virtual Services and Exhibitionism have missed classifications when posts lacked explicit keywords, relying instead on implicit phrases like online sessions" or suggestive imagery. False Positives typically arose from contextual misinterpretations, such as flagging benign terms like meet and greet" as Physical Services or generic pronouns like "we" as Couples and Group Interactions. Significant confusion was also observed due to the inherent ambiguity of the Miscellaneous Fun category and the substantial thematic overlap between Virtual services and Content Creation and Services, making it hard for models to distinguish between live interactions and content sales.

Model-specific behaviors revealed distinct trade-offs between precision and recall. GPT-4 and Llama adopted a more conservative, high-precision approach, minimizing FPs but resulting in high FNs by overlooking nuanced cues, particularly in the Ex, MF, and CGI categories. Conversely, Mistral demonstrated stronger recall for VS and CCS by recognizing industry-specific terms (e.g., OnlyFans"), but this came at the cost of more FPs in the PS category. Gemini proved adept at identifying subtle Ex cues but tended to over-classify CGI and

CCS content. Claude provided a more balanced performance but struggled to resolve ambiguity between borderline PS and MF classifications. In summary, all models were consistently challenged by contextual ambiguity in vague terms like arrangements," the nuances of slang, and pervasive overlaps between service categories.

From Adoption to Adaptation: Tracing the Diffusion of New Emojis on Twitter

Yuhang Zhou
University of Maryland
College Park, USA
tonyzhou@umd.edu

Xuan Lu
University of Arizona
Tucson, USA
luxuan@arizona.edu

Wei Ai
University of Maryland
College Park, USA
aiwei@umd.edu

Abstract

The frequent introduction of new emojis in each Unicode release creates a dynamic shift in social media content, providing a unique opportunity to explore the evolution of digital language. Analyzing a large dataset of sampled English tweets, we examine how newly released emojis gain popularity and evolve in meaning. We find that community size of early adopters and emoji semantics are positively correlated with their popularity. Certain emojis experienced notable shifts in the meanings and sentiment associations during the diffusion process. Additionally, we propose a novel framework utilizing language models to extract words and pre-existing emojis with semantically similar contexts, which enhances interpretation of new emojis. The framework demonstrates its effectiveness in improving downstream text classification performance by substituting unknown new emojis with familiar ones. This study offers a new perspective in understanding how new language units are adopted, adapted, and integrated into the fabric of online communication.

1 Introduction

The language landscape of the Web era is ever-evolving, characterized by the emergence and evolution of new language units. Individuals have creatively crafted out-of-vocabulary language units, such as Internet memes and viral hashtags, to encapsulate and convey complex ideas, sentiments, and cultural phenomena, fostering shared lexicons that resonate across digital communities. Understanding the adoption and adaptation of these language units is crucial for gaining insights into the dynamic nature of online communication, the information diffusion process in social networks, and the underlying social trends and movements. However, analyzing the dynamics of these emerging language units in online communication presents unique challenges. These units lack universal conventions and standards and their characteristics may vary during diffusion, making it complex to track their initial appearances, early adoption, and frequency of use.

As a recent addition to this landscape, emojis offer a distinctive opportunity to explore the diffusion and evolution of new language units. Emojis are visual symbols that are embedded into text. These non-verbal symbols go beyond a single word or phrase, encapsulating rich semantics spanning a wide spectrum of emotions, actions, objects, and concepts. Originating as emoticons in the early Internet culture, emojis have evolved into a standardized and universally recognized visual language. Unlike other language units like hashtags or internet memes, emojis undergo a standardized process prior to their inclusion in the language. They are proposed to the Unicode consortium, formally defined by the Unicode standard, uniquely coded as Unicode strings, such as U+1F603 for emoji 😃, and then rendered by various platforms.

Since the Unicode started to adopt emojis in 2010, we have witnessed emojis' remarkable rise on the Web, with their adoption consistently increasing across multiple platforms (Rong et al., 2022; Lu et al., 2018; Kejriwal et al., 2021; Halverson et al., 2023). New emojis continue to be introduced in response to user requests. From 2018 to 2022, five new versions of emojis (Unicode 11.0 to 15.0) have been released, some of which, like the pleading face emoji (🥺) and the partying face emoji (🥳), have gained widespread adoption among Twitter users.

This standardized approach ensures that emojis maintain a stable form throughout their journey within social networks, enabling precise tracking of emoji adoption and diffusion. These attributes – precise definition, standardized implementation, stable form, and accurate release information – underscore the unique suitability of emojis for study-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 293–306

ing the evolution of language in social networks.

We take the initiative to study the diffusion of emerging language units on social media through the unique perspective of Unicode-versioned emojis. We investigate the diffusion of newly created emojis introduced in Unicode versions 11.0 to 13.0, across the Twitter platform, particularly the usage frequency and semantic shift as the new emojis cascade through social media.

Because of the rich semantics, the Unicode definition (such as "pleading face") is far from enough to interpret the meaning of an emoji. To this end, we propose an interpretation framework that leverage language models (LMs) to identify words and existing emojis that share similar semantic contexts with the new emojis. Finally, we evaluate the practical implications of our framework by substituting new emojis with semantically similar ones in sentiment classification and irony detection tasks, which demonstrates the effectiveness of our approach in helping NLP models interpret emerging language units for downstream tasks. We summarize our major contributions as follows:

- We explore the pattern of emoji diffusion initial adoption to widespread usage, with a focus on frequency and sentiment aspects.
- We introduce an interpretation framework to interpret the semantics of new emojis by exploring the words or old emojis with similar semantics.
- To validate the effectiveness of our interpretation framework, we replace emojis in texts with surrogates and improve the model performance in the sentiment classification and irony detection task.

2 Related Work

Our work builds upon three research areas: emoji understanding and applications, innovation diffusion on social media, and temporal effects in machine learning models.

Emoji Understanding and Applications. The functions and applications of emojis are widely studied, including their role in conveying sentiment (Ai et al., 2017), irony (Hu et al., 2017), topics (Lu et al., 2016), and identity (Ge, 2019). Emojis have been incorporated into various downstream NLP tasks, notably sentiment analysis (Chen et al., 2019; Felbo et al., 2017; Eisner et al., 2016) and irony detection (Hayati et al., 2019). While the emoji development process (Lu et al., 2018) and requests (Feng et al., 2020) have been examined, the diffusion dynamics and evolving interpretations of *recently introduced* emojis remain largely unexplored. Furthermore, existing methods often fail to generalize effectively to these new, unseen emojis.

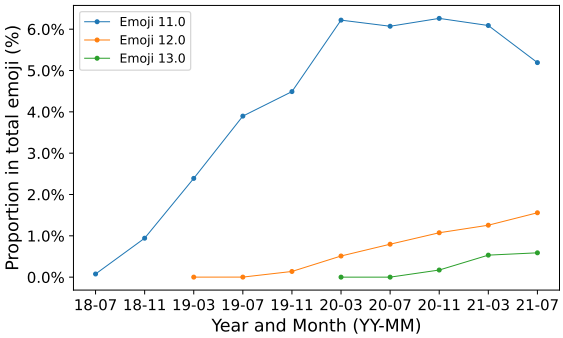


Figure 1: Frequency trends per emoji version over three years post-release. Y-axis: proportion of total emoji usage from each version (shown every four months).

Innovation Diffusion on Social Media. Research has explored the diffusion patterns of various online content like hashtags, memes, and news (Ma et al., 2014; Johann and Bülow, 2019; Kümpel et al., 2015; Zhou et al., 2021), including studies on general language innovation (Grieve et al., 2018; Kershaw, 2018). We argue that the standardized, versioned nature of Unicode emojis presents a unique, controlled setting to study language unit adoption. Therefore, this work specifically investigates the diffusion of new emojis through the lens of innovation diffusion theory (Rogers et al., 2014; Ma et al., 2014), examining factors influencing their uptake and popularity.

Temporal Effects on Models. Language evolution on social media causes a temporal shift, degrading the performance of models trained on older data when applied to newer text (Röttger and Pierrehumbert, 2021; Huang and Paul, 2018; Agarwal and Nenkova, 2022). Newly introduced emojis, appearing as out-of-vocabulary tokens to models trained prior to their release, exacerbate this issue. While methods like continual training can adapt models (Röttger and Pierrehumbert, 2021; Ke and Liu, 2022; Su et al., 2022), they incur significant computational costs. We propose a novel emoji substitution approach as a computationally efficient alternative to help models interpret new

emojis without requiring retraining.

3 Emoji Diffusion: Frequency and Semantics

We explore the diffusion of new emojis (Unicode 11.0-13.0, released 2018-2020) on Twitter from May 2018 to May 2022, focusing on usage frequency and semantic evolution, particularly through the lens of sentiment context. Our dataset comprises English tweets collected via the Twitter API[1].

3.1 Emoji Usage Frequency Dynamics

Newly released emojis generally exhibit increasing usage over the subsequent two years (Figure 1), though adoption rates vary dramatically both between versions and within versions. The popularity often follows a power-law distribution, similar to established emojis (Lu et al., 2016), with top emojis being orders of magnitude more frequent than others (Figure 2). Adoption speeds also vary; for example, 🥺 initially lagged behind others in Emoji 11.0 but later surged in popularity, suggesting diffusion across different community boundaries over time. We also observe that older emoji versions (e.g., Emoji 11.0) can eventually plateau or decrease in usage, potentially as users adopt newer alternatives.

While overall usage grows, short-term frequency can be influenced by external events. Fine-grained weekly analysis revealed temporary spikes in usage coinciding with events like New Year's Day and Valentine's Day, confirmed by examining associated keywords during those periods (details in Appendix A).

3.2 Semantic Evolution during Diffusion

Given increasing adoption, we investigate if emoji meanings adapt during diffusion. As a proxy for semantic context, we analyze the sentiment of tweets containing new emojis using Vader (Hutto and Gilbert, 2014), averaging scores over two-month periods.

For most top emojis studied (from Emoji 11.0), the average sentiment score remained relatively constant, suggesting stable interpretations (Figure 3). However, a notable exception is 🥺 (pleading face), whose average sentiment score progressively increased (became more positive) during the first year post-release.

[1] https://developer.twitter.com/en/docs/twitter-api

This shift suggests an evolution in usage. Initially, 🥺 appeared frequently in negative contexts, but over time, its usage in positive contexts grew significantly (see Appendix Figure 6 for score distributions). Analysis of highly associated words (via PMI) supports this: early associated sentimental words included negative terms like '*sad*' and '*hate*', while a year later, positive terms like '*prettiest*' and '*cutest*' became prominent (Table 1). This demonstrates that the perceived meaning and application context of an emoji can evolve as it diffuses through the user population.

Time Period	Top 10 PMI Sentimental Words	Score Avg.
2018-10	cry sad please miss hate sorry idk wish bad stop	-0.198
2019-10	sobbing protect cry prettiest pls precious hug cutest sad heart	0.221

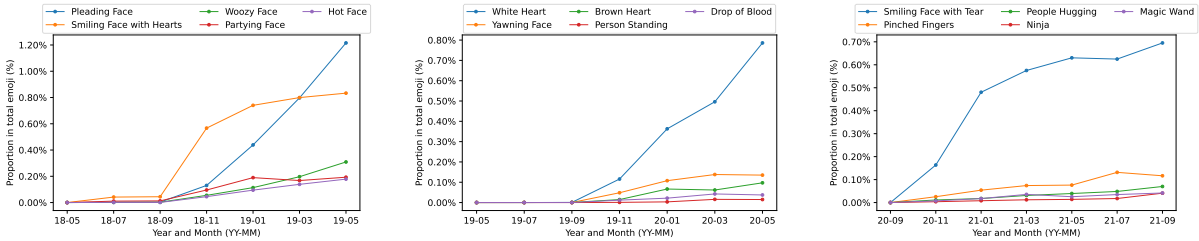
Table 1: Top 10 associated words with positive or negative sentiments of emoji 🥺 in October 2018 and October 2019. Red and blue highlight positive and negative words, respectively. The darker the background color, the more sentimental the words.

4 Influencing Factors of Emoji Diffusion

The observed popularity discrepancies (Section 3) raise the question of what influences the diffusion process. Viewing emojis as digital innovations, we apply Rogers' diffusion of innovation theory (Rogers et al., 2014), examining factors related to the diffusion network (early adopters) and the innovation itself (emoji semantics). We defined early and late stages as two months and fourteen months post-release, respectively.

Diffusion theory suggests larger early adopter communities accelerate spread (Ma et al., 2014; Steffes and Burgee, 2009). We proxied community size by the popularity of hashtags co-occurring with new emojis early on (Yang et al., 2012; Zhou and Ai, 2022) (details in Tables 5 and 6 in Appendix B.1). We found positive Spearman correlations between early hashtag counts and late emoji counts for the top 10 emojis across versions (e.g., $\rho = 0.580, p = 0.08$ for Emoji 11.0; $\rho = 0.530, p = 0.11$ for 12.0), indicating a tendency for larger initial communities to predict greater eventual popularity.

The innovation's characteristics also matter (Rogers et al., 2014). We proxied semantic



(a) Emoji 11.0 (b) Emoji 12.0 (c) Emoji 13.0

Figure 2: Frequency trends of the top 5 popular emojis in each emoji version over the two years following their first appearance. We show the frequency of emojis every two months, and the y-axis represents the proportion of each emoji in the total of emojis in that month.

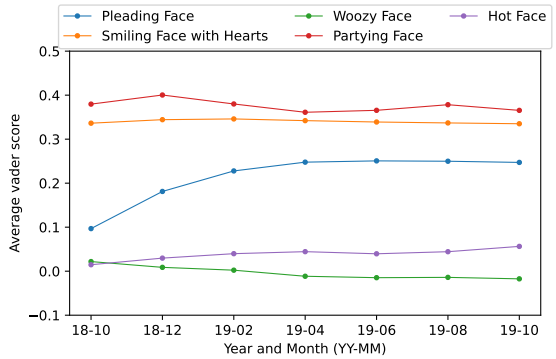


Figure 3: Average Vader scores of the tweets containing the top 5 popular emojis in Emoji 11.0 from October 2018 to October 2019.

popularity using the early-stage frequency of semantically similar words generated by GPT-4 (OpenAI, 2023) (details in Tables 7 and 8 in Appendix B.2). We found significant positive correlation for Emoji 12.0 (Spearman $\rho = 0.780, p < 0.01$; Pearson $r = 0.812$) and positive tendency for 11.0, suggesting emojis representing frequently discussed concepts tend to become more popular. The weaker correlation for 13.0 might reflect limitations in keyword proxies for certain emoji semantics. Overall, both early community size and semantic relevance show associations with emoji diffusion success, aligning with innovation diffusion principles.

5 Interpret Emojis with Language Models

The analysis so far shows that the semantic of emojis not only affect their diffusion process, but also evolves during the diffusion process, both of which highlights the importance of an effective framework to interpret new emojis dynamically.

Although ChatGPT is capable of showing us similar words to emojis in Emoji 11.0 to 13.0, it relies on the stereotypical associations to infer emojis' semantics and lacks the understanding based on the application scenario, which cannot capture the semantic evolution during diffusion (Zhou et al., 2024). Secondly, the pre-training data of ChatGPT may not cover the recently-released emojis or emojis not included in the Unicode organization. For example, GPT4 generates hallucinations when prompting it to explain the semantics of 🫠 (broken) and 🤪 (onlooker), specifically used on the WeChat app (Zhou et al., 2024).

To address these challenges, we utilize the corpus containing new emojis and open-source language models (LMs) to investigate the application scenario and semantic meaning of emojis. Previous researchers used LMs in interpretation work (Lin et al., 2023; Zhou et al., 2023; Romanou et al., 2023) and relied on the attention mechanism to explore the inner association of the fine-tuning data (Wang et al., 2021). In this section, we use the attention score method and the cross-dataset inference method to explore words and old emojis with semantics similar to new emojis.

5.1 Interpretation with High-attention Words

Attention scores are a well-studied interpretability method to identify important tokens for LMs to make the decision (Clark et al., 2019; Wang et al., 2021). To understand what words are specifically associated with newly created emojis, we design an emoji prediction task and extract high-attention words to interpret the emoji's meaning.

5.1.1 Attention Calculation

Formally, we first construct an emoji classification dataset with the input space $x \in \mathcal{X}$ and the pre-defined output space $y \in \mathcal{Y} = \{0, 1, \cdots, n\}$, where x is a tweet excluding the emoji and y represents the emoji label. To better distinguish the word association between new and old emojis, the

tweet with label 0 means that the tweet contains the old emoji, and with label $1 \leq k \leq n$ means that the tweet contains the new emoji k. Denote f as the fine-tuned model in the emoji classification dataset (specifically the Roberta model for our experiments) (Liu et al., 2019). For each input tweet $x_i \in \mathcal{X}$ with tokens $\{t_i^1, t_i^2, \cdots, t_i^m\}$, where m is the token number, we adopt f in the input i and obtain the attention scores $\{a_i^1, a_i^2, \cdots, a_i^m\}$ and the emoji prediction $f(i)$. Since for Roberta model, the embeddings of the [CLS] token in the last layer are used to make the prediction, we compute the scores a_i^m as the average attention scores of the m^{th} token to the [CLS] token across different heads. For the overall attention scores $\overline{a_{kt}}$ of the token t for the emoji k, we extract the sentences with prediction $f(x_i) = k$ and average the attention scores of the token t in these extracted sentences, which can be formulated as

$$\overline{a_{kt}} = \frac{\sum_{x_i \in \mathcal{X}} \mathbb{1}(f(x_i) = k) \cdot \sum_{j=1}^{m} (a_i^j \cdot \mathbb{1}(t_i^j = t))}{\sum_{x_i \in \mathcal{X}} \mathbb{1}(f(x_i) = k) \cdot \sum_{j=1}^{m} \mathbb{1}(t_i^j = t)}$$

where $\mathbb{1}$ is the indicator function. With the overall attention scores, for each emoji k, we can extract the keywords with the highest attention scores $\overline{a_{kt}}$ to understand the semantics of the new emojis.

5.1.2 Experiment Setup and Results

To verify the effectiveness of the attention method, we first extracted tweets containing emojis from April 2022 to May 2022 using the Twitter API. We randomly selected six emojis from Emoji 13.0, each appearing in more than 1,000 tweets: 🥲 (smiling face with tears), 🥷 (ninja), 🪄 (magic wand), 🫰 (pinched fingers), 🪙 (coin), and 🫂 (people hugging). A balanced dataset was constructed with a total of 50,000 tweets, divided into 7 labels (label 0 for old emojis and labels 1 to 6 for the new emojis), ensuring each emoji had a sufficient number of tweets for accurate fine-tuning.

We fine-tuned the RoBERTa model pre-trained on tweets (Barbieri et al., 2020) (`twitter-roberta-base`[2]) on the emoji prediction dataset with an 8:1:1 train-validation-test split, achieving a training accuracy of 78.29% and a test accuracy of 66.98%. The top 10 words with the highest attention scores are presented in the second column of Table 2, with words recognized by the authors as having similar semantics highlighted in bold.

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base

Emoji	Top 10 Attention Score Words	Top 3 Inference Score Old Emojis
🥲	same, its, so, me, **no**, **why**, this, **please**, **oh**, i	😔 (16.7), 🥺 (8.5), 😩 (6.1)
🥷	days, account, **ninja**, both, **assassins**, who, **samurai**, coin, **gaming**, website	✨ (13.3), 👀 (13.2), 😊 (10.2)
🪄	**magic**, **wish**, **magician**, **wizard**, follow, hours, tweet, special, **light**, recent	✨ (22.6), 💥 (10.1), 🔥 (9.9)
🫰	this, puff, the, that, **kiss**, another, you, **art**, **perfect**, **please**	😍 (10.2), 🔥 (7.8), 💯 (6.1)
🪙	start, **billion**, **coins**, **token**, **coin**, **money**, proof, gob, hours, follow	💥 (25.1), 🎉 (16.4), 🔥 (10.5)
🫂	**thanks**, **hugs**, my, **hug**, **good**, you, **hugging**, **dear**, **friend**, **happy**	🥺 (11.2), 💙 (8.9), ✨ (8.3)

Table 2: Words with high attention scores and old emojis with similar inference scores for emoji inference of the new emojis from Emoji 13.0. High-attention words suggest words with similar semantics and the application scenario of emojis.

Table 2 demonstrates that attention scores are effective in identifying words semantically similar to specific emojis. These words not only mirror the primary meaning of the emojis but also reveal their application scenarios and extended interpretations. For entity-related emojis such as 🥷, 🪄, and 🪙, their high-attention words extend beyond their direct symbolism. For instance, the word "gaming" associated with 🥷 highlights its frequent use in the context of action video games. Similarly, the term "token" linked with 🪙 (coin) suggests its application in representing Bitcoin tokens, indicating a broader usage beyond its conventional meaning.

Regarding sentiment-related emojis, the high-attention score words encapsulate the sentiments embedded in the emojis. For instance, the word "perfect" associated with 🫰 implies positive sentiment, while "why" and "no" linked to 🥲 suggest negative sentiments. The qualitative results in Table 2 demonstrate the effectiveness of using the attention mechanism to probe the application setting and the extensive meaning of the new emojis.

5.2 Interpret Emojis with Old Emojis

Words with high attention scores do not fully capture the sentiments inherent in emojis. For example, the degree of negative sentiment conveyed by 🥲, or the specific sentiment associated with 🥷 and 🪙, remains ambiguous. Besides exploring words with similar semantics, we employ the rich and complex sentiments encoded in old emojis to interpret the newly created emojis.

We utilize LMs to conduct cross-version infer-

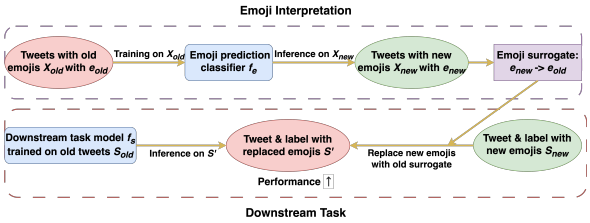


Figure 4: The upper half shows the framework of using old emojis to interpret new emojis and the lower half presents the pipeline of replacing new emojis in the downstream prediction task dataset to old similar emojis to enhance the results.

ence to explore old emojis with semantics similar to a new targeted emoji. If the semantics and syntactic features of the text containing two emojis are similar, the semantics of two emojis are also similar. Our method is to first fine-tune the LMs on an emoji classification dataset with old emojis as the labels, and we use the fine-tuned LMs to do the inference on tweets containing new emojis. If LMs predict tweets with a new emoji to contain another old emoji, it means that two emoji share a similar text context distribution, indicting similar semantics. The pipeline of the interpretation framework is shown in the upper half of Figure 4.

5.2.1 Cross-Version Analysis

We construct an emoji classification dataset with the input tweets $\mathcal{X}_{old}$ and the pre-defined emoji labels from old emojis $\{e^0_{old}, e^1_{old}, \cdots, e^n_{old}\}$. Pre-trained LM f_e is fine-tuned in tweet collection $\mathcal{X}_{old}$. We construct another tweet dataset $\mathcal{X}_{new}$, where each tweet contains the new emojis $\{e^0_{new}, e^1_{new}, \cdots, e^m_{new}\}$ and ask the fine-tuned LM f_e to do the inference on each tweet $x_{new} \in \mathcal{X}_{new}$. The prediction on the tweet x_{new} containing the emoji y_{new} is:

$$f_e(x_{new}) = \underset{e_{old}}{\operatorname{argmax}}\, p(e_{old}|x_{new}).$$

The semantic similarity between an old emoji e_{old} and a new emoji e_{new} can then be quantified as the proportion of predictions equal to e_{old}:

$$simi(e_{old}, e_{new}) = \frac{\sum_{x_{new}} \mathbb{1}(y_{new}=e_{new}, f_e(x_{new})=e_{old})}{\sum_{x_{new}} \mathbb{1}(y_{new}=e_{new})}. \quad (1)$$

We sort the old emojis for each new emoji by the calculated $simi(\cdot)$ values and extract the top three old emojis with the highest similarity values for each new emoji.

5.2.2 Experiment Setup and Results

We first constructed $\mathcal{X}_{old}$ using tweets from April 2020 to May 2020, before Emoji 13.0 appeared on Twitter. We extracted 10,000 tweets for each of the top 32 emojis from April to May, along with an additional 10,000 tweets without any emojis. This resulted in a training dataset, $\mathcal{X}_{old}$, comprising 330,000 tweets and 33 emoji labels (32 emojis plus a "no emoji" label). The pre-trained RoBERTa model, f_e, was fine-tuned on $\mathcal{X}_{old}$, and the fine-tuned model was then used to infer on the test dataset constructed with tweets from 2022 containing the emojis from Emoji 13.0, as described in Section 5.1. We calculated the semantic similarity values between the old and new emojis using Equation 1 and presented the top 3 similar ones for each new emoji, along with their similarity values, in the third column of Table 2.

Table 2 reveals that preexisting emojis effectively reflect the semantic content, particularly the sentiment dimension, of new emojis. For sentiment-related emojis such as 🥲, older emojis such as 😔 (pensive face), 🥺 (pleading face) and 😩 (weary face) encapsulate the blend of sadness and begging inherent in the new emoji. Furthermore, in the case of entity-related emojis, while it may be challenging to pinpoint analogous emojis that precisely capture an entity's characteristics, it is feasible to discern the underlying sentiments these emojis convey. For example, the 🥷 (ninja) emoji, in relation to existing emojis such as ✨ (sparkles), 👀 (eyes), and 😊 (smiling face with smiling eyes), reveals the positive sentiment in 🥷 in usage.

In the next section, we will demonstrate how this associated emoji analysis can be applied to downstream tasks. We find that simply substituting new emojis with their similar old emojis, as emoji surrogates, significantly increases the accuracy in the sentiment classification and irony detection task of fine-tuned language models.

5.3 Emoji Substitution in Downstream Tasks

Sentiment classification and irony detection are two well-studied natural language processing (NLP) tasks widely used in many deployed systems. Emojis, as non-verbal tokens rich in semantics, have been shown to be important for training effective machine learning models for sentiment classification or irony detection (Chen et al., 2019; Hayati et al., 2019).

The current state-of-the-art (SoTA) method for

text classification tasks on tweets involves fine-tuning pre-trained language models on a training dataset. However, given a fine-tuned language model on a tweet dataset from 2020, testing on tweets from 2022 containing the new emojis from Emoji 13.0 may lead to imprecise predictions due to the absence of these new emojis in the fine-tuning data.

To address this issue, instead of fine-tuning models on 2022 tweets again, which requires high computational cost, we propose a method to directly substitute the new emojis with old emoji surrogates known by the fine-tuned language models. We expect the semantics encoded in similar old emojis (emoji surrogates) to compensate for the loss of semantics in the unknown new emojis. We illustrate the emoji substitution process in the downstream classification tasks in the lower half of Figure 4.

5.3.1 Emoji Substitution

Suppose that we have a tweet classification dataset $\mathcal{S}_{new}$ that contains the new emojis and that each tweet is labeled with the label $y \in \{0, 1\}$. Given a classifier f_s fine-tuned in $\mathcal{S}_{old}$, where the new emojis are not in the vocabulary of the classifier f_s, for each tweet $s_i = [t_i^1, t_i^2, \cdots, e_{new}, \cdots, t_i^m] \in \mathcal{S}_{new}$, we can replace the new emoji e_{new} in tweet $s_i \in \mathcal{S}_{new}$ with the old emoji surrogates (top 3 old similar emojis sorted by the calculated similarity value in Section 5.2) and obtain a dataset $\mathcal{S}^{'}$ with $s_i^{'} = [t_i^1, t_i^2, \cdots, e_{old}^1, e_{old}^2, e_{old}^3, \cdots, t_i^m]$. Then we feed the tweet $s_i^{'}$ with the old emoji surrogates to the classifier f_s to get the model prediction, which is: $f_s(s_i) = \text{argmax}_{y \in \{0,1\}} p(y|s_i^{'})$.

5.3.2 Experiment Setup and Results

For the existing classification datasets for tweets, they were developed before 2018 and did not contain the newly created emojis for Emoji 13.0 or Emoji 14.0. Therefore, we collect tweets from 2020 and 2022, respectively, and ask LLM to annotate the labels. Due to the superiority of LLMs over human crowd-workers on the text annotation task (Gilardi et al., 2023; Zhou et al., 2024), we utilize ChatGPT to annotate the sentiment or irony label for tweets (Ouyang et al., 2022). Given a tweet input s, the ChatGPT model (GPT-3.5 for Emoji 13.0 and GPT-4o for Emoji 14.0), a sentiment label set {positive, neutral, negative}, an irony label set {irony, non-irony}, we ask ChatGPT to annotate the tweet with the given sentiment and irony label sets. We repeat the LLM annotation process twice with the temperature 0.7 and only keep the examples and labels agreed by two LLM annotators.

We first collect tweets from April 2020 to May 2020 and then use ChatGPT to label the sentiment of each tweet as $\mathcal{S}_{old}$. We first split $\mathcal{S}_{old}$ into training, validation, and test dataset by 8:1:1 and downsample the training set to make a balanced set with 35,388 tweets. We fine-tune the Roberta classifier (`twitter-roberta-base`) pre-trained on tweets before August 2019 as f_s (Barbieri et al., 2020). We repeat the collection and labeling process on tweets from April 2022 to May 2022 to form $\mathcal{S}_{new}$. In the randomly sampled test dataset, the fine-tuned model f_s can achieve an accuracy of 67.52% in tweets from 2020 and 67.65% in tweets from 2022 in the sentiment classification task, which means that the two-year language evolution is not significant to greatly affect the overall performance of the classifier.

To validate ChatGPT's reliability for tweet annotation, we evaluated its performance on the SemEval-2015 Task 10 sentiment classification test set (Rosenthal et al., 2015). ChatGPT achieved 77.29% agreement with the gold standard labels. This accuracy compares favorably to the reported average human agreement of 75.70% obtained during the dataset's construction. Given this comparable performance to human annotators on a standard benchmark, we consider ChatGPT sufficiently reliable for annotating tweets in our study.

To evaluate the effectiveness of our emoji replacement method for tweets containing new emojis, we randomly selected 16 popular emojis from Emoji 13.0 and Emoji 14.0, each appearing more than 300 times in $\mathcal{S}_{new}$. These included 8 sentiment-related emojis and 8 entity-related emojis. We collected a total of 91,066 tweets containing the selected emojis and used the fine-tuned language model f_s to make predictions for each tweet. Next, we applied the proposed emoji replacement method, substituting the new emojis with existing emoji surrogates, and asked f_s to re-predict. Additionally, we prepared two baseline methods for replacing new emojis with text: one replaced the new emoji with its corresponding emoji name (words representing the emoji's meaning), while the other replaced the emoji with a description generated by ChatGPT, leveraging its detailed understanding of emoji semantics (Zhou et al., 2024). The generated descriptions of ChatGPT for each emoji and the prompt are shown in Table 9 in the Appendix. We computed the prediction accuracy

emoji	surrogates	# test	original acc	replaced acc	name acc	description acc
Emoji 13.0 (sentimental)						
🥲	😔😅😄	23,080	64.4 ± 1.38	**66.4 ± 3.53**	60.9 ± 1.49	58.5 ± 7.77
🥸	🤔👀😊	1,908	55.8 ± 10.0	64.9 ± 1.42	**65.9 ± 1.97**	64.1 ± 1.07
🫂	😊💙✨	6,580	82.4 ± 2.01	**90.4 ± 1.22**	86.4 ± 1.63	85.2 ± 3.92
🤌	🤩🔥💯	6,972	80.8 ± 1.53	**83.6 ± 1.05**	65.7 ± 2.89	64.8 ± 2.85
Emoji 13.0 (entity)						
🩻	👍🔥💥	614	23.9 ± 2.34	20.7 ± 2.53	24.4 ± 1.32	**28.3 ± 2.06**
🥷	✨👀😊	6,649	56.1 ± 1.53	**58.4 ± 2.19**	56.0 ± 2.74	56.8 ± 2.50
🎱	💥🎉🔥	3,209	56.0 ± 5.22	**61.3 ± 0.88**	59.2 ± 0.49	56.3 ± 0.59
🪬	👀👀✨	922	66.5 ± 1.11	67.4 ± 0.82	**68.1 ± 0.85**	61.1 ± 5.56
Emoji 14.0 (sentimental)						
🫠	😊✨😀	9305	**78.7 ± 0.36**	77.2 ± 0.88	70.9 ± 5.22	70.5 ± 4.48
🥹	🥹✨💗	9328	80.0 ± 0.92	**82.5 ± 0.49**	77.5 ± 3.37	61.7 ± 0.84
🫶	✨💗😀	9374	76.7 ± 0.07	**81.9 ± 0.18**	80.2 ± 2.49	81.2 ± 1.15
🤏	💯✨👀	9333	**80.4 ± 0.53**	79.9 ± 1.31	76.7 ± 1.84	78.2 ± 0.49
Emoji 14.0 (entity)						
🪸	✨💥💯	1702	86.2 ± 0.16	**87.7 ± 0.25**	86.7 ± 0.75	81.4 ± 1.38
🪨	✨🎉🔥	1325	83.3 ± 0.32	**85.0 ± 0.48**	82.2 ± 1.28	79.0 ± 3.15
🪭	💥✨🔥	383	89.2 ± 1.67	**89.4 ± 1.66**	86.9 ± 1.10	88.4 ± 0.86
🪺	💥✨💯	382	**93.3 ± 0.18**	92.7 ± 0.74	90.8 ± 0.00	89.0 ± 1.41

Table 3: Sentiment prediction accuracy using emoji replacement for Emoji 13.0/14.0. Model pretrained pre-Aug 2019 (new emojis excluded). Accuracy columns compare: original text, replacement with similar old emojis, emoji names, or GPT descriptions.

for original tweets, emoji-replaced tweets, name-replaced tweets, and description-replaced tweets for each emoji. The results of the sentiment classification task are presented in Table 3, and those of the irony detection task are presented in Table 10 in the Appendix.

From Table 3, we observe a notable performance improvement for the model f_s fine-tuned on tweets without new emojis when replacing the new emojis with older emojis of similar semantics. Specifically, the model exhibits a relative improvement of 5.02% for sentimental emojis and 4.71% for entity-related emojis from Emoji 13.0, as well as a relative improvement of 1.78% and 1.51% for sentimental and entity-related emojis from Emoji 14.0, respectively. When compared to replacing new emojis with text (names or generated descriptions), our proposed method outperforms in 11/16 cases. This is likely because the words from emoji names or ChatGPT-generated descriptions fail to fully capture the complex sentiments conveyed by the emojis. Furthermore, as shown in Table 10 in the Appendix, a similar pattern emerges in the irony detection task, where our emoji replacement method achieves the best performance in 13/16 cases. These findings demonstrate the effectiveness and robustness of our proposed method in addressing out-of-vocabulary emoji tokens in downstream text classification tasks.

We also repeat the experiments on the Roberta model with new emojis in the pretrained corpus (`twitter-roberta-base-2022-154m`, pretrained on tweets before December 2022) (Loureiro et al., 2023), and present the results of sentiment classification in Table 11 in the Appendix. Compared with Roberta model pretrained on tweets before 2020 (`twitter-roberta-base`), the average relative improvement of replacing new emojis with old emojis becomes smaller or even negative, which follows our expectation that the fine-tuned model f_s has learned the semantics of new emojis in the pretraining data and the substitution of emojis causes loss of information.

6 Implications

Our research reveals the patterns of emoji diffusion and proposes a framework to understand the semantics of new emojis. We expect that our exploration of emoji diffusion can benefit the emoji designer by considering the influencing factors to design more attractive emojis in future works. For model developers, our emoji replacement framework can provide an effective way to increase the generalization of the fine-tuned LLMs on texts containing new emojis without further parameter updates, which can also be extended to other NLP tasks. For future emoji researchers, our work provides a pipeline to explore the pattern of new emoji diffusion and proposes a framework to utilize open source LMs to interpret the emoji semantics and usage scenarios.

Moreover, our work provides an insight into the diffusion patterns of a new language unit, which can inspire future researchers to generalize our diffusion exploration pipeline and interpretation framework on other new non-verbal tokens, such as memes and viral hashtags. Along with our work, we expect that the exploratory study on the language unit diffusion can reveal the evolution mechanism of the content in digital communities.

7 Conclusion

In this work, we analyzed the diffusion patterns of recently created emojis, finding external events influence short-term frequency and sentiment meanings can evolve during adoption. Early adopter community size and emoji semantics correlate with eventual popularity. We proposed an interpretation framework using associated words and similar existing emojis. Its effectiveness was demonstrated by substituting new emojis with interpreted surrogates, improving model performance on sentiment and irony detection tasks. This framework offers a way to enhance model generalization to new emojis in downstream applications.

8 Limitations

There are two limitations in our work. First, the sample size for our correlation measurement in Section 4 is small and such a small sample size is difficult to obtain the significant correlation (p-value< 0.05). We select 10 emojis from each version to analyze the factors that influence the emoji diffusion, since the emoji number is a long-tailed distribution, where the emoji numbers after the top 10 are close. We measure the association across three versions of emojis to make our conclusion more generalizable.

Second, when analyzing the characteristics of new emojis, in this paper, we only focus on the semantic features of emojis, also the focus of the previous work, but we find that the visual features could also be an important feature for emoji diffusion and may also influence the emoji semantics. For example, after one year of diffusion, the number of tweets with 🤍 (white heart) is 10 times higher than the tweets with 🤎 (brown heart). These visual features play an important role in exploring emoji usage and downstream emoji applications.

References

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *ICWSM 2017*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered representation learning for Cross-Lingual sentiment classification. In *WWW 2019*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint*.

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. *EMNLP*.

Yunhe Feng, Zheng Lu, Wenjun Zhou, Zhibo Wang, and Qing Cao. 2020. New emoji requests from twitter users: when, where, why, and what we can do about them. *ACM Transactions on Social Computing*, 3(2):1–25.

Jing Ge. 2019. Emoji sequence use in enacting personal identity. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 426–438, New York, NY, USA. Association for Computing Machinery.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46(4):293–319.

Colin ME Halverson, Claire E Donnelly, Michael Weiner, and Joy L Lee. 2023. Content analysis of emoji and emoticon use in clinical texting systems. *JAMA Network Open*, 6(6):e2318140–e2318140.

Shirley Anugrah Hayati, Aditi Chaudhary, Naoki Otani, and Alan W Black. 2019. What a sunny day: Toward emoji-sensitive irony detection. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 212–216.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Nguyen, and Jiebo Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM 2017*.

Xiaolei Huang and Michael Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM 2014*.

Michael Johann and Lars Bülow. 2019. One does not simply create a meme: Conditions for the diffusion of internet memes. *International Journal of Communication*, 13:23.

Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*.

Mayank Kejriwal, Qile Wang, Hongyu Li, and Lu Wang. 2021. An empirical study of emoji usage on twitter in linguistic and national contexts. *Online Social Networks and Media*, 24:100149.

Daniel James Kershaw. 2018. *Language change and evolution in online social networks*. Lancaster University (United Kingdom).

Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society*, 1(2):2056305115610141.

Victoria Lin, Louis-Philippe Morency, and Eli Ben-Michael. 2023. Text-transport: Toward learning causal effects of natural language. *arXiv preprint arXiv:2310.20697*.

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*.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2023. Tweet insights: A visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*.

Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. In *UbiComp 2016*.

Xuan Lu, Yanbin Cao, Zhenpeng Chen, and Xuanzhe Liu. 2018. A first look at emoji usage on github: An empirical study. *arXiv preprint*.

Long Ma, Chei Sian Lee, and Dion Hoe-Lian Goh. 2014. Understanding news sharing in social media: An explanation from the diffusion of innovations theory. *Online information review*, 38(5):598–615.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. 2014. Diffusion of innovations. In *An integrated approach to communication theory and research*, pages 432–448. Routledge.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*.

Shiyue Rong, Weisheng Wang, Umme Ayda Mannan, Eduardo Santana de Almeida, Shurui Zhou, and Iftekhar Ahmed. 2022. An empirical study of emoji use in software development communication. *Information and Software Technology*.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Paul Röttger and Janet B Pierrehumbert. 2021. Temporal adaptation of bert and performance on downstream document classification: Insights from social media. *arXiv preprint arXiv:2104.08116*.

Erin M Steffes and Lawrence E Burgee. 2009. Social ties and online word of mouth. *Internet research*, 19(1):42–59.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and Min Zhang. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change. *arXiv preprint arXiv:2210.17127*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what @you #tag: Does the dual role affect hashtag adoption? In *WWW 2012*.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Yuhang Zhou and Wei Ai. 2022. #emoji: A study on the association between emojis and hashtags on twitter. In *ICWSM 2022*.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2023. Explore spurious correlations at the concept level in language models for text classification. *arXiv preprint arXiv:2311.08648*.

Yuhang Zhou, Paiheng Xu, Xiyao Wang, Xuan Lu, Ge Gao, and Wei Ai. 2024. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. *arXiv preprint arXiv:2402.01681*.

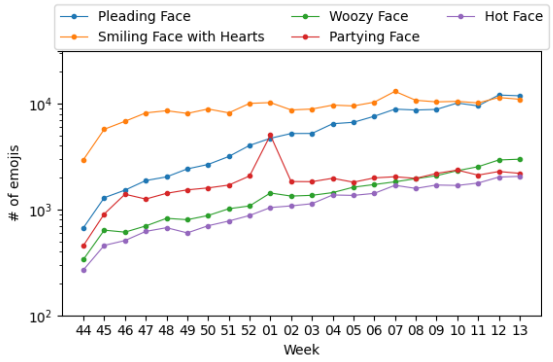


Figure 5: Frequency trends by week of the top 5 popular emojis in Emoji 11.0 from the end of 2018 to the start of 2019.

Emoji	Time period	Top 10 PMI words
🥳	51st week, 2018	birthday, happy, hope, great, days, we, day, **year**, amazing, christmas
🥳	1st week, 2019	happy, **new**, **year**, birthday, **2019**, **happynewyear**, may, everyone, **years**, **2018**
🥰	51st week, 2018	amazing, you, wait, love, thank, thanks, see, beautiful, heart, so
🥰	7th week, 2019	looking, **valentines**, tomorrow, **valentine**, ever, amazing, pretty, sweet, beautiful, you

Table 4: Highly-associated words (quantified by PMI value) of emoji 🥳 and 🥰 in different weeks. The semantics of the associated words coincide with the external events happened at that week.

Appendix

A Weekly Frequency Fluctuations and Event Correlation

Seeing the continuous growth of new emojis' popularity after release, we then examine the change in emoji frequency in a more fine-grained time period. We visualize the count of the 5 most popular emojis of Emoji 11.0 at each week from November 2018 to March 2019 in Figure 5. We observe two significant bumps on the lines of 🥳 (partying face) and 🥰 (smiling face with hearts): a bump for 🥳 in Week 1 of 2019 and the a bump for 🥰 in Week 7 of 2019.

The bumps coincide bursting external events (New Year in Week 1 and Valentine's day in Week 7), which have been discussed in literature as possible triggers for information cascade (Zhou et al., 2021; Crane and Sornette, 2008). We hypothesize that external events also influence the adoption of new emojis.

To verify, we examine whether the words associated with the new emoji in that period are related to external events. We collect the tweets with emojis in the first and seventh week of 2019 and calculate pointwise mutual information (PMI) between each emoji e and each word w. The PMI equation can be formulated as $\text{PMI}(e, w) = \log \frac{p(e,w)}{p(e)p(w)}$, where $p(w)$, $p(e)$ and $p(e, w)$ refer to the probability of a tweet containing the word w, the emoji e, and both of them, respectively. We present the top 10 associated emojis (based on PMI) for emoji 🥳 and 🥰 in different weeks in Table 4 and highlight the words related to external events, as recognized by the authors.

From Table 4, we observe that for emoji 🥳 from the 51st week of 2018 to the 1st week of 2019, the words about the New Year event such as "*happynewyear*" and "*2019*" appear in the associated words, and for emoji 🥰, the associated words about Valentine's days show in the 7th week of 2019. The observation verifies our hypothesis and suggests that external events can influence the adoption of new emojis.

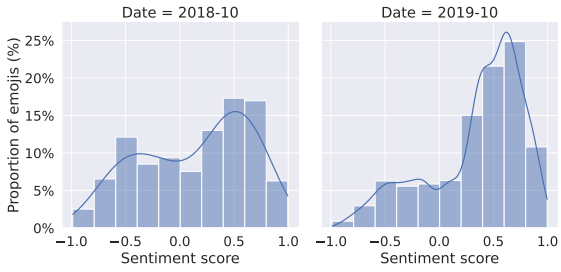


Figure 6: Distribution of Vader scores of tweets containing emoji 🥺 in October 2018 and October 2019. During one year period, 🥺 concentrate more on the positive tweets.

B Details of Influencing Factor Analysis

B.1 Influence of Community Size on Emoji Popularity

We first identify the early stages of Emoji 11.0, Emoji 12.0, and Emoji 13.0 as July 2018, September 2019, and November 2020, respectively, two months after their initial release, and set the late stage as one year after the early stage. We count the number of co-occurring hashtags for each emoji in the early stage and the number of emojis in the late stage. We present the top 10 most used emojis in the late stage from Emoji 13.0, along with their top 3 popular co-occurring hashtags, in Table 5 and the count of emojis and hashtags in the early and late stages of emoji diffusion. The co-occurring hash-

Emoji	Co-occurred hashtags	# hashtag (early)	# emoji (late)
😡	PLTPINKMONDAY, PLTPINKSUNDAY, PLTCyberMonday	79982.2	67,103
🤌	ATINYDAY, HeartbreakWeather, NiallHoran	18806.0	11,583
🗳️	EveryVoteCounts, StayHome, StaySafe	7172.9	7,829
🌑	Bitcoin, gold, DeFi	10267.2	5,900
🪄	MissguidedCyberTreat, PLTPINKMONDAY, ibelieveinfairys	11501.1	5,355
🤦‍♀️	syurabaikhati, DerApotheker, wolfpac	1128.8	4,510
🪴	GreenIndiaChallenge, SupportSmallStreamers, RRRMovie	46272.1	3,014
🤓	Election2020, PLTPINKMONDAY, TREASURE	83920.8	2,497
🫓	HappyThanksgivingEve, Design, Emoji	866.6	2,113
🪧	MerzmenschPresents, LatentVoices, JukeBox	3343.3	1,666

Table 5: Co-occurred hashtags for top 10 popular emojis in Emoji 13.0 in the late stage. # hashtag (early) represents the hashtag number in the early stage of emoji diffusion, two months after the first appearance of emojis. # emoji (late) shows the emoji number in the late stage, one year after the early stage.

tags for emojis in Emoji 11.0 and 12.0 are shown in Table 6.

B.2 Influence of Emoji Semantics on Emoji Popularity

Our prompt for querying GPT-4 is composed as follows: *Show me five common single words on Twitter with similar semantics to this emoji: emoji*, where *emoji* is a new emoji from Emoji 11.0, 12.0, or 13.0, such as 🥰 (smiling face with hearts). We count the average number of words in the early stage of new emojis (the same month as in Section B.1) and the emoji number in the late stage (one year after the early stage). Since we find that the emoji and word counts are on the same scale, we calculate the Spearman rank correlation coefficient and the Pearson correlation coefficient for the logarithm of these two numbers as the measurement of association (Zar, 2005; Cohen et al., 2009). We present the word and emoji numbers for the top 10 popular emojis from Emoji 13.0 in Table 7, and for the similar words for emojis in Emoji 11.0 and 12.0, we show them in Table 8 in the Appendix.

Emoji	Co-occurred hashtag	# hashtag (early)	# emoji (late)
🥹	SaveShadowhunters, EXO, BTSARMY,	37122.3	101,667
🥰	TeenChoice, SouhilaBenLachhab, Cover	70924.1	56,449
🥴	Prediction, Democrats, Obamacare	5121.8	29,549
🥵	Heatwave, SummerSkinSafety, WorldEmojiDay	693.8	14,650
🥳	Ethereum, cryptocurrency, eth	22075.3	13,070
🥶	TRuMP, WorldCup2018, UFC231	11416.0	2,137
🫂	Directive, ItsACelebration, TeamNGH	76.1	616
🪫	MAGA, CRO, Boxing	8831.0	539
🧁	Cakes, baking, weekendreads	1979.1	435
🦻	TMay, Chequers, UK	1736.25	339

(a) Co-occurred hashtags for emojis in Emoji 11.0

Emoji	Co-occurred hashtag	# hashtag (early)	# emoji (late)
🤍	J9, 3YearsWithCBX, StreamLYTLM	448.50	85,981
🧍	Tawan_V, TeamGalaxy, withGalaxy	144.20	13,473
😮‍💨	13ReasonsWhy3, 13ReasonsWhy, DeleteFacebook	291.60	13,127
🤎	thestripperoficial, camren, natiese	2.0	9,321
🪶	StarTrekDiscovery, Friends, RossGeller	390.0	4,433
🩸	PeriodEmoji, PeriodStigma, PeriodPoverty	12.0	3,648
🪶	onlyfans, camgirl, cammodel	793.0	2,729
🪄	AccessATE, InstructionalDesign, CADET	9.70	2,603
🟣	wolvsden	7.0	2,393
🥾	-	0	1,936

(b) Co-occurred hashtags for emojis in Emoji 12.0.

Table 6: Co-occurred hashtags for top 10 popular emojis in Emoji 11.0 and 12.0. # hashtag (early) represents the hashtag number in the early stage of emoji diffusion, two months after the first appearance of emojis. # emoji (late) shows the emoji number in the late stage, one year after the early stage.

Emoji	Words with similar semantics	# word (early)	# emoji (late)
🥲	bittersweet, emotional, touched, relieved, grateful	1401.8	67,103
🤌	gesture, expressive, Italian, emphasis, talkative	190.8	11,583
🫂	hug, comfort, support, embrace, togetherness	3714.6	7,829
🪙	money, currency, cash, change, gold	5999.8	5,900
🪄	magic, enchantment, spell, wizardry, mystical	609.6	5,355
🥷	stealthy, mysterious, skilled, warrior, covert	159.4	4,510
🪴	green, leafy, indoor, botanical, decorative	726.6	3,014
🥸	incognito, hidden, undercover, sneaky, disguised	403.8	2,497
🫓	bread, flat, pita, naan, food	674.0	2,113
🪧	sign, protest, message, board, banner	1311.0	1,666

Table 7: Words with similar semantics (from ChatGPT) for top 10 popular emojis in the late stage in Emoji 13.0. # word (early) represents the word number in the early stage of emoji diffusion, two months after the first appearance of emojis. # emoji (late) shows the emoji number in the late stage, one year after the early stage.

Emoji	Words with similar semantics	# word (early)	# emoji (late)
🥺	please, sorry, help, sad, desperate	12973.8	101,667
🥰	love, happy, adorable, blissful, sweet	33005.4	56,449
🥴	dizzy, confused, woozy, drunk, lightheaded	1242.0	29,549
🥵	hot, sweaty, exhausted, overwhelmed, burning	2578.2	14,650
🥳	celebrating, party, woohoo, ecstatic, jubilant	1470.0	13,070
🥶	cold, freezing, shivering, frosty, icy	591.6	2,137
🧸	cute, soft, cuddly, plush, adorable	4112.6	616
🧨	loud, explosive, bang, pop, fireworks	1273.2	539
🧁	sweet, delicious, cute, frosting, treat	5303.2	435
🦶	step, walk, run, sole, toe	2263.0	339

(a) Words with similar semantics for emojis in Emoji 11.0.

Emoji	Words with similar semantics	# word (early)	# emoji (late)
🤍	pure, love, clear, sincere, peace	21063.2	85,981
🧍	stand, upright, wait, solo, idle	4793.4	13,473
🥱	tired, sleepy, bored, exhausted, drowsy	1676.2	13,127
🤎	warmth, earthy, comfort, stable, rich	496.6	9,321
🪫	small, little, slight, precise, minimal	3843.8	4,433
🩸	blood, bleed, drip, red, donate	1376.2	3,648
🪐	saturn, space, cosmic, orbit, celestial	398.6	2,729
🦯	blind, aid, navigate, mobility, independence	266.2	2,603
🟣	purple, geometric, round, violet, circle	842.6	2,393
🦦	otter, playful, aquatic, furry, adorable	487.4	1,936

(b) Words with similar semantics for emojis in Emoji 12.0.

Table 8: Words with similar semantics (from ChatGPT) for top 10 popular emojis in Emoji 11.0 and 12.0. # word (early) represents the word number in the early stage of emoji diffusion, two months after the first appearance of emojis. # emoji (late) shows the emoji number in the late stage, one year after the early stage.

Emoji	ChatGPT Description
	Emoji 13.0
🥲	Bittersweet smile
🤓	Trying to act clever
🫂	Warm embrace
🤌	Perfectly done
🪜	Step up
🥷	Stealthy move
🪙	Shiny coin
🫐	Juicy blueberry
	Emoji 14.0
🫠	Feeling overwhelmed
🥹	Teary gratitude
🫶	Heartfelt gesture
🫡	Respectfully noted
🪷	Serene lotus
🪩	Disco vibe
🪸	Coral reef
🪺	Cozy nest

Table 9: ChatGPT-generated descriptions for emojis from Unicode Emoji 13.0 and 14.0. The prompt used for generating descriptions is: *Please use a few words to replace the emoji: {emoji}, preserving its meaning and ensuring the text reads naturally. Only give me the most suitable substitution.*

Emoji 13.0 (sentimental)						
emoji	surrogates	# test	original acc	replaced acc	name acc	description acc
🥲	😔🥺😩	23,080	72.0 ± 2.87	**74.6 ± 1.68**	74.3 ± 0.45	73.1 ± 0.15
🤓	🤔👀😔	1,908	56.7 ± 4.67	**69.7 ± 7.85**	67.0 ± 5.15	68.4 ± 4.96
🫂	🥺💙✨	6,580	95.0 ± 0.81	**95.8 ± 0.06**	93.9 ± 0.46	95.5 ± 0.08
🤌	😍🔥💯	6,972	85.7 ± 0.38	**86.8 ± 0.16**	85.8 ± 0.13	86.2 ± 0.32
	Emoji 13.0 (entity)					
🪜	👍🔥💥	614	96.8 ± 0.45	**97.1 ± 0.21**	96.9 ± 0.31	96.9 ± 0.22
🥷	✨👀😊	6,649	97.0 ± 0.05	97.1 ± 0.15	97.1 ± 0.02	**97.2 ± 0.02**
🪙	💥🎉🔥	3,209	97.0 ± 0.06	97.1 ± 0.09	97.2 ± 0.06	**97.3 ± 0.25**
🫐	💥👀✨	922	97.6 ± 0.08	**97.7 ± 0.07**	97.4 ± 0.12	97.5 ± 0.04
	Emoji 14.0 (sentimental)					
🫠	😔✨🥺	9305	78.2 ± 0.53	**78.8 ± 0.52**	76.2 ± 0.07	77.5 ± 0.30
🥹	🥺✨💗	9328	93.6 ± 0.46	**94.0 ± 0.03**	92.0 ± 0.88	93.4 ± 0.54
🫶	✨💗🥺	9374	96.2 ± 0.19	**96.5 ± 0.01**	96.1 ± 0.35	96.4 ± 0.13
🫡	💯✨👀	9333	84.4 ± 0.84	**84.5 ± 0.83**	76.2 ± 5.78	84.1 ± 0.37
	Emoji 14.0 (entity)					
🪷	✨💥💯	1702	97.1 ± 0.24	**97.2 ± 0.08**	96.7 ± 0.31	97.0 ± 0.39
🪩	✨🎉🔥	1325	95.5 ± 0.38	95.6 ± 0.24	95.1 ± 0.19	**95.7 ± 0.20**
🪸	💥✨🔥	383	98.3 ± 0.14	**98.7 ± 0.14**	97.9 ± 0.14	97.6 ± 0.00
🪺	💥✨💯	382	96.0 ± 0.41	**96.1 ± 0.28**	95.2 ± 0.27	95.3 ± 0.25

Table 10: Results of the emoji replacement method on irony detection tasks for selected emojis from Emoji 13.0 and Emoji 14.0, using models pretrained on data prior to August 2019 (excluding new emojis in the pre-training data). All values are reported in percentage.

Emoji 13.0 (sentimental)				
emoji	surrogates	# test	ori acc	replaced acc
🥲	😔🥺😩	23,080	**66.57 ± 0.03**	60.63 ± 0.06
🤓	🤔👀😔	1,908	58.22 ± 1.61	**68.10 ± 2.71**
🫂	🥺💙✨	6,580	83.86 ± 4.42	**91.33 ± 0.35**
🤌	😍🔥💯	6,972	83.14 ± 0.52	**83.90 ± 0.16**
	Emoji 13.0 (entity)			
🪜	👍🔥💥	614	**29.36 ± 9.85**	24.60 ± 9.94
🥷	✨👀😊	6,649	**58.69 ± 0.95**	57.45 ± 3.36
🪙	💥🎉🔥	3,209	63.05 ± 1.03	**66.55 ± 0.42**
🫐	💥👀✨	922	67.95 ± 2.83	**68.17 ± 1.62**

Table 11: Results of emoji replacement method on the sentiment prediction task for selected emojis from Emoji 13.0 on models pre-trained on data before December 2022, including new emojis in the pre-training data.

Social Construction of Urban Space: Using LLMs to Identify Neighborhood Boundaries From Craigslist Ads

Adam Visokay[1] **Ruth Bagley**[4] **Ian Kennedy**[2] **Chris Hess**[3]
Kyle Crowder[1] **Rob Voigt**[4] **Denis Peskoff**[1,5]

Sociology
[1]University of Washington
[2]University of Illinois Chicago
[3]Kennesaw State University

Linguistics
[4]Northwestern University
Information
[5]University of California, Berkeley

Abstract

Rental listings offer a window into how urban space is socially constructed through language. We analyze Chicago Craigslist rental advertisements from 2018 to 2024 to examine how listing agents characterize neighborhoods, identifying mismatches between institutional boundaries and neighborhood claims. Through manual and large language model annotation, we classify unstructured listings from Craigslist according to their neighborhood. Further geospatial analysis reveals three distinct patterns: properties with conflicting neighborhood designations due to competing spatial definitions, border properties with valid claims to adjacent neighborhoods, and "reputation laundering" where listings claim association with distant, desirable neighborhoods. Through topic modeling, we identify patterns that correlate with spatial positioning: listings further from neighborhood centers emphasize different amenities than centrally-located units. Natural language processing techniques reveal how definitions of urban spaces are contested in ways that traditional methods overlook.

1 Contested Neighborhood Boundaries

Neighborhood location matters for a wide range of individual and collective outcomes (Sampson et al., 2002; Sharkey and Faber, 2014; Minh et al., 2017; Chyn and Katz, 2021). Beyond objective demographic characteristics, the subjective features of a neighborhood—its reputation, status, or stigma—shape resident satisfaction, place attachment, and overall well-being (Tran et al., 2020; Kullberg et al., 2010; Permentier et al., 2011; Otero et al., 2024). Neighborhood reputation also structures the economic value of property, patterns of investment, and the residential mobility that drives neighborhood stratification (Krysan and Crowder, 2017; Evans and Lee, 2020; Kirk, 2024; Korver-Glenn and Mayorga, 2024).

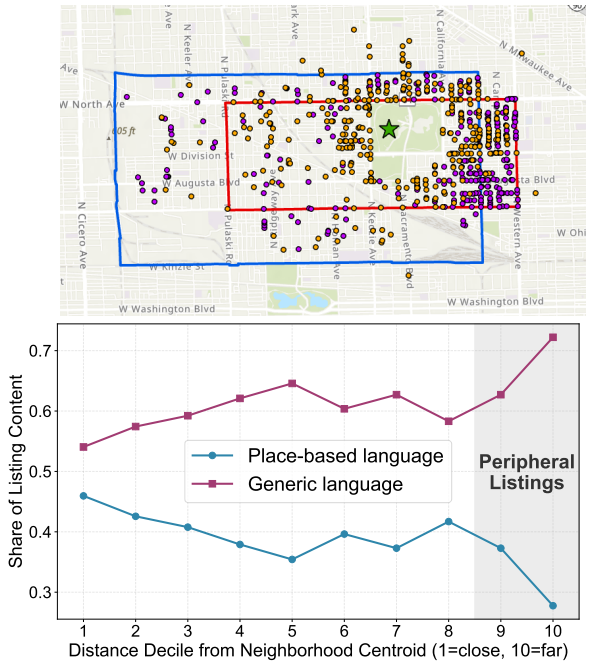


Figure 1: Extracting neighborhoods from unstructured rental listings with LLMs (RQ1, Section 3) provides insight into the social construction of space (*RQ2*, top)[1] and allows us to study how language changes relative to distance from neighborhood centers (*RQ3*, bottom).[2]

We define **neighborhood identity** not just as a boundary in space, but as the spatial area that corresponds with the patterns of social activity and perceptions of people living in the area. This conceptualization builds on a lineage of research utilizing user-generated content (UGC) to define urban space (Plangprasopchok and Lerman, 2009; Hollenstein and Purves, 2010; Hiippala et al., 2019;

[1] *Conflicting conceptions* of the Humboldt Park neighborhood according to the official City of Chicago limits (blue), Zillow's boundary (red) and rental advertisements (points). Orange points depict unit listings claiming to be Humboldt Park while purple points claim elsewhere. The green star denotes the LLM-defined social center of Humboldt park.

[2] Across Chicago, the share of place-based language decreases and generic "boilerplate" language increases in listings for units further from the social center of the neighborhood.

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 307–321

Brunila et al., 2023). Because this identity is socially constructed through interaction and language, it is inherently fluid and contested. Identifying a singular "true" neighborhood boundary is not our aim, but rather, mapping the contours of contestation: the systematic slippage between institutional maps and the spatial claims made by social actors.

Neighborhoods are socially constructed through interaction and language (Zelner, 2015; Hohle, 2023; Stuart et al., 2024), yet traditional urban research often relies on rigid administrative boundaries as a proxy for neighborhood identity. In practice, listing agents frequently navigate a tradeoff between geographic fidelity and reputational leverage, substituting symbolic identity for physical proximity when properties sit at the periphery of desirable areas. We characterize this behavior not as random noise, but as a systematic distortion of urban space. Still, an important question remains: do spatial claims that depart from institutional maps always represent strategic "reputation laundering" (Stuart et al., 2024), or might they reflect legitimate disagreements arising from the inherent ambiguity of socially constructed boundaries?

A related challenge for urban sociology has been the difficulty of observing and classifying such distortions at scale. Conventional data is blind to the strategic manipulation of spatial labels, and traditional NLP methods like string-matching struggle to distinguish between casual mentions and strong locational claims. Large Language Models (LLMs) provide a transformative opportunity to recover this latent social variable: claimed neighborhood identity. By evaluating LLMs on Chicago Craigslist rental advertisements (2018–2024), this work provides answers to the following Research Questions:

- **(RQ1) Measurement Viability:** Can zero-shot LLMs accurately identify specific neighborhood claims (vs. mere mentions) in unstructured rental listings compared to more traditional string-matching?
- **(RQ2) Social Location:** Where are neighborhoods actually located according to listing claims, and can we define a meaningful "social center" for each neighborhood?
- **(RQ3) Linguistic Substitution:** How does marketing language vary with spatial location? Specifically, does place-based language change as listings move farther from their claimed neighborhood center?

Section 2 describes our spatially-anchored corpus of Chicago listings. Section 3 evaluates LLM performance against string-matching baselines **(RQ1)**. Section 4 develops a framework for identifying neighborhood "social centers" **(RQ2)**. Section 5 and Section 5.2 analyze the semantic structure and statistical associations between unit language and spatial positioning **(RQ3)**. Finally, Section 6 contextualizes these findings within the broader field of computational social science.

2 Craigslist Housing Advertisements

We use data collected from Chicago Craigslist rental advertisements from 2018 to 2024 to identify listings with mismatches between the neighborhood containing the unit and the neighborhood claimed by the listing agent.[3]

2.1 Why Craigslist?

These rental listings offer a particularly valuable lens for examining how neighborhoods are socially constructed and contested because Craigslist's platform design creates an unusually unconstrained environment for spatial classification. Unlike many digital platforms that restrict users to predetermined administrative or commonly recognized neighborhood boundaries, Craigslist allows advertisers to freely designate any neighborhood label in their listings with unstructured text. This feature transforms rental advertisements into sites of boundary-making where the socio-spatial imaginary of the city becomes visible.

The neighborhood fields in these listings represent more than mere locational information—they reveal how real estate actors actively participate in constructing, reinforcing, or challenging existing spatial hierarchies. When landlords and property managers assign neighborhood labels to their listings, they engage in acts of spatial categorization that reflect both market strategies and internalized cognitive maps of urban space. These choices may align with officially recognized boundaries, reproducing understandings of place, or deliberately transgress established spatial categories to claim association with perceived higher-status areas.

[3]We have been actively collecting data in Chicago since 2018, providing a rich window into the discursive construction of urban space. It includes all available advertisements each day from December 2018 until June 2024 using the web-scraper `Helena` (Chasins et al., 2018; Hess and Chasins, 2022). Occasional changes to the architecture of the Craiglist website result in limited periods of data loss, the longest of which was from August 2019 to early October 2019.

Preceding computational analysis of Craigslist rental listings explores price distributions (Boeing and Waddell, 2017), neighborhood descriptions (Kennedy et al., 2020; Besbris et al., 2021), housing policy interventions (Boeing et al., 2021), exclusionary language (Stewart et al., 2023), affordability (Hess et al., 2023), and home security (Somashekhar et al., 2024). Holistically, research shows that rental listings on Craigslist align with and appear to reproduce social inequality. Recent work has begun to focus on the importance of neighborhood names and the places those names describe (Schachter et al., 2024), with specific focus on contested naming: when advertisements use neighborhood names that seem to diverge from the name most local residents would use for that space.

2.2 Why Chicago?

We focus on Chicago because it stands as a quintessential "city of neighborhoods," where locally recognized community areas hold exceptional cultural, economic, and social significance (Hwang and Sampson, 2014). Chicago is an ideal site for examining the social construction of urban space due to the high salience of its neighborhood boundaries. While the city maintains 77 officially recognized community areas, these rigid, non-overlapping boundaries often fail to represent the fluidity of neighborhood identity in reality. By using Chicago's stable institutional definitions as a point of comparison, we can more effectively identify and quantify how real estate actors use language to challenge or reinforce existing spatial hierarchies. This neighborhood orientation is so deeply embedded in Chicago's social fabric that it shapes how residents understand their place in the city, influences social networks, and structures daily mobility patterns (Kaysen, 2024).

By analyzing patterns in how these spatial designations align with or diverge from official boundaries, we can observe in real time the processes through which neighborhood reputations are reinforced or contested. The negotiated quality of these spatial boundaries becomes particularly visible when examining cases where advertisers claim association with neighborhoods other than those in which units are formally located, according to administrative boundaries. Such instances of spatial repositioning offer a window into the dynamics that shape how urban space is valued, categorized, and ultimately experienced by various stakeholders from listing agents to residents or potential tenants to municipal administrators.

3 Detecting Neighborhoods with LLMs

Online rental advertisements are generally unstructured and vary widely between listings. Distinguishing between which neighborhoods are mentioned in a listing from which neighborhood(s) the listing claims to be in is a nuanced task of great importance to social scientists interested in the social construction of urban space. This answer to "which neighborhood does this advertisement claim the unit to be in?" is not always obvious, even to a human annotator. For example,

> ...this fully restored **East Lakeview** property sits on a beautiful tree-lined street located in the heart of the popular **Wrigleyville** neighborhood ...

It is clear based on the language that this advertisement is not merely mentioning these neighborhoods, but staking a strong claim to being located in both. Wrigleyville is a Chicago neighborhood within the larger neighborhood of East Lakeview, so this claim is entirely coherent. However, reconciling such competing claims at scale is a principal challenge inherent to this particular task.

Furthermore, some listings contain mentions of several irrelevant neighborhoods, even dozens like this example:

> ...Disclaimer: Pricing, availability, and specials are subject to change at any time and without notice. HotSpot Rentals services the following neighborhoods: South Loop, Printers Row, Near North, River North, Gold Coast, West Loop, Fulton River District, West Loop Gate, The Loop, Streeterville, Lakeshore East, New East Side, Old Town, Medical District, University Village, Near North, River West, Lincoln Park South, Lakeview, Uptown, Ukrainian Village, Wicker Park, Edgewater, Ravenswood, Bronzeville, Logan Square.

Making the distinction between neighborhood mentions and strong claims that a unit is in a particular neighborhood is a nuance that large language models are particularly well-suited for compared to existing methods. To make this comparison, first we label the full corpus using a bespoke string-matching approach which serves as the baseline

"best practice" which we compare to the Language Model-based labeling. Both sets of labels are evaluated against a subset of 200 manually labeled neighborhood listings. These manual labels were produced by authors of this article.

3.1 Manual Labeling

The process for creating our validation set of "gold standard" labels considers three sources of information for each advertisement in the following order. First, if the title field includes a neighborhood name, that becomes the manual label for the neighborhood claim. Then if there is no claim in the title, we consider the body of the listing. This is the largest source of text in each advertisement, and also the most ambiguous with respect to identifying strong claims. When faced with multiple claims – as in the quote above – we take the first claim as the manual label. Finally, if neither the title nor body fields contain a neighborhood claim, we extract a neighborhood claim from the neighborhood field, if it exists. An advertisement only receives the 'unknown' label if there is not a strong claim in any of the three fields. We follow the same prioritization scheme in the string-matching and LM labeling procedures.

3.2 Data Pre-Processing

Before labeling we performed standard data preprocessing and de-duplication on the raw text. We removed common boilerplate text that appeared in virtually all Craigslist listings (such as "QR Code Link to This Post"), corrected Unicode translation errors to ensure consistent character rendering, and precisely mapped listings onto Zillow and City of Chicago neighborhood boundaries to confirm geographical validity. We also performed thorough de-duplication of listing entries by title and body text, retaining only the most recent version when multiple entries existed, as Craigslist saves edited listings as separate posts. This preparation distilled a clean dataset of 30,531 unique listings from an initial corpus of n=128,764 initial observations.

3.3 String-Match Labeling Neighborhoods

To determine which Chicago neighborhood a rental advertisement belongs to based on the text in the post we begin with a list of 192 distinct neighborhoods from Zillow, a real-estate marketplace company which provides commercial neighborhood lists for major US cities. Then, we manually construct a comprehensive list of regular expressions that can match the official name and its alternatives (e.g. *wriggleyville, wriglyville, wrigglyville, wrigley ville* for *wrigleyville*). These regex patterns are designed to be flexible with spaces and case-insensitive. Following the same protocol as the manual annotations, we use a function which tries to match neighborhoods in the listing title, body and dedicated neighborhood field. When multiple neighborhoods match, we select the label which appears earliest in the text.

3.4 Language Model Labeling Neighborhoods

We prompt GPT-4.1 mini as a high quality relative to cost option.[4] Table 1 contains our prompt.

```
BASE_PROMPT
Extract the Chicago neighborhood from the rental text.
Rules:
- NEVER use the address or zip code to determine the neighborhood
- Choose only explicitly stated neighborhood from possible responses
- If neighborhood is unclear mark it as [unknown]
- Format precision: "lakeview" but not "x, y, z" for "lakeview residence near x, y, z"
text: {body}
Possible responses: {zillow_list}
Return only:
label: [neighborhood]
```

Table 1: Prompt format for extracting Chicago neighborhood claims from rental listings.

We create three separate labeling workflows, one for each title, body and neighborhood field. We input the data independently to avoid leakage. A Zillow neighborhood list is provided to constrain possible responses to items in the list. In the case of unknowns, we prioritize the label from the title, then the body, then the neighborhood field. Only if all three are unknown is the final neighborhood claim labeled 'unknown'.

3.5 Label Post-Processing

While we engineer the prompt to provide structured output in the form `label:response`, the outputs still require post-processing. We implement a multi-stage post-processing pipeline to standardize the three LLM labels for each advertisement and assess model performance against manual validation. Through manual review we flag instances of minor spelling discrepancies (e.g. 'lakeview' instead of

[4] We compare a sample of regular vs mini vs nano, and other non-OpenAI options. Reasoning models are unnecessary for this extraction task.

Metric	String Match	GPT-4.1
Accuracy	0.79	**0.85**
Macro Average F1	0.62	**0.70**
Weighted Average F1	0.77	**0.85**

Table 2: GPT-4.1 outperforms the string match-based keyword search on the neighborhood claim labeling task. While the performance gain is marginal, the LLM labeling process was cheap, fast and can be scaled up.

'lake view', 'wrigglyville' instead of 'wrigleyville'). For text normalization, we construct a replacement dictionary to correct such instances. We implement the same priority-based claim-selection algorithm as is used for manual annotation and string matching. The post-processing system first examines the listing title label; if no neighborhood is identified, it analyzes the listing body label, and if still unsuccessful, it checks the neighborhood field label; and only then returns 'unknown' classification. This aligns with how potential renters process listing information, prioritizing the most prominent textual elements. When faced with compound neighborhood designations, we prioritize the first neighborhood when multiple were present, just as we do for manual labeling and string-matching. For outputs that contain separators (e.g. 'uptown, ravenswood' or 'avondale/logan square'), we extract the first neighborhood claim before a separator.

3.6 Evaluation of Labeling Task

We compare the string-match and LLM labels to the manual labels in the 200-item validation set. Performance metrics are shown in Table 2. While accuracy scores are comparable (GPT-4 mini's 85% is marginally better than the string matching's 79%), the disagreements largely reflect the inherent ambiguity of neighborhood classification – a challenge even human annotators face when listings claim multiple neighborhoods. A notable advantage of the LLM approach is its efficiency and scalability. Unlike string matching on keywords, which requires extensive manual configuration of locale specific patterns, the pre-trained LLM can work with a zero-shot prompt, making it adaptable.

4 Geospatial Analysis of Neighborhoods

Although Craiglists rental listings do not comprehensively show every available property in the Chicago area, there is still substantial coverage across the city, as seen in Figure 2. As a result, we take these rental listings to be a reasonably representative sampling, which we use to identify neighborhoods as they might be conceived of by the people, or at least the realtors, of Chicago.

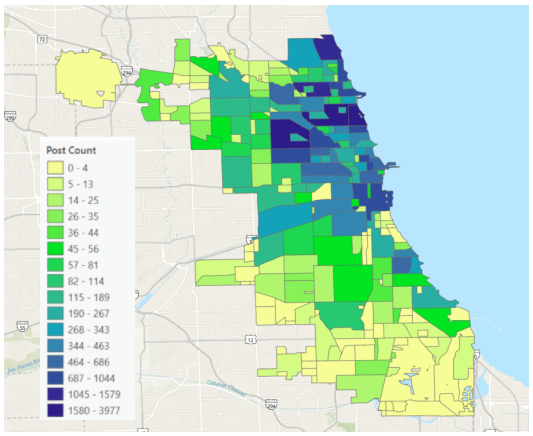


Figure 2: The city of Chicago and its constituent neighborhoods, as defined by Zillow, colored to represent the number of posts for properties located in each area

Neighborhood boundaries are nebulous and hard to define; even the City of Chicago and Zillow, both with access to a great deal of data/information, have developed quite different maps of neighborhood boundaries. Figure 1 shows that while the official Zillow and city boundaries of Humboldt Park include most of the posts claiming to be in that neighborhood, none of the borders are the same. In addition, borders between neighborhoods are likely less rigid than an official boundary might suggest; despite being located within the formal boundaries of Humboldt Park, there are a number of listings, mainly around the edges, claiming to be part of other nearby neighborhoods, primarily Ukrainian Village, Wicker Park, or West Town.

Using the rental listings, we conceptualize the "social center" of the neighborhood as the centroid of all listings that claim to be located within that neighborhood. We use `geopandas` to identify the centroid of all posts claiming to be in the same neighborhood using the latitude and longitude coordinates of all property locations. The map of Humboldt Park suggests this is an effective method, because the social center (represented by the green star) is in fact located in the eponymous park which is considered to be the heart of the neighborhood.

However, not all posts claiming to be in a given neighborhood may have the same strength to their claim; advertisements for apartments at the center of a popular neighborhood like Logan Square likely have different characteristics than posts for units on the fringes of the neighborhood. The posts on the

fringes might even be trying to seem more desirable by claiming to be in a more popular neighborhood, whereas it might be more of an objective description of location for a property actually located on Logan Square. We conceptualize this by identifying how far from the social center a post is, using three metrics for distance: 1) Raw Distance: Euclidean distance between the latitude and longitude coordinates of the post and that of the neighborhood centroid; 2) Relative Distance: raw distance for all posts in the neighborhood projected onto a [0,1] interval using min-max scaling; 3) Z-scored Distance: z-scored distance for posts claiming to be in the same neighborhood

We use these measures to distinguish a specific subset of the posts: those on the periphery of a neighborhood. We define peripheral posts as those that are in the furthest 20% of posts from the centroid for a given neighborhood (for any neighborhood labels with at least 5 posts).

4.1 Comparison of Neighborhood Boundaries

In order to get a more concrete understanding of different conceptions of neighborhoods, we identify two more neighborhoods to explore more in depth: Logan Square and Pilsen.

Logan Square is a well-known neighborhood in Chicago, and also a neighborhood label where a large number (n=2495) of listings claim to be. We can see in Figure 4 that the City of Chicago boundaries for Logan Square contain primarily postings claiming to be in the neighborhood (orange points) without many posts claiming to be elsewhere (purple). The Zillow boundaries do encompass Logan Square claims that the Chicago boundaries do not, but the areas excluded by Chicago also have a higher concentration of claims of being in other neighborhoods. In addition, there are a number of listings claiming to be in Logan Square that are outside both formal boundaries–these posts would certainly be part of the 'peripheral' posts we defined earlier. This kind of posting merits further exploration to better understand what is happening in listings that claim to be part of a neighborhood when that might be more likely to be contested.

Although the blue City of Chicago boundaries appear to be a better approximation for Logan Square in Figure 4, this is not the case for every neighborhood. Pilsen, shown in Figure 3, is another well-known neighborhood within Chicago, but it is not included as a distinct neighborhood by the City of Chicago. However, even though Zillow

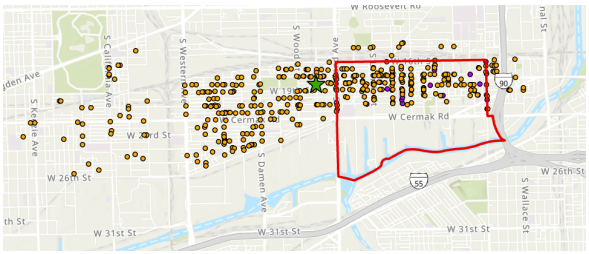


Figure 3: Contested boundaries for Pilsen neighborhood according to Zillow's definition (red). Orange points depict listings claiming to be in Pilsen, purple points are listings claiming to be elsewhere). The green star depicts the Pilsen neighborhood centroid. This is an example of what we call *border stretching*, demonstrating how static boundary systems may not capture the neighborhood identities as experienced by the people living in them.

does include Pilsen as an area, it also does not seem to define it in the same way as Craigslist advertisements. Many of the posts claiming to be in Pilsen appear in a clustered way outside of the Zillow boundaries, and even the centroid of the Craigslist-defined neighborhood is outside the Zillow bounds. This could represent a developing neighborhood identity, which may not have been as strong at the time of the map creation, and supports the need for a more dynamic model of neighborhoods.

5 Content Analysis of Rental Listing Text

We use the `tomotopy` Python package to identify latent topics in the content posted in Craigslist rental advertisements. `tomotopy` uses Gibbs-sampling and is based on the LDA approach described in (Blei et al., 2003) and (Newman et al., 2009). We prepared the text corpus by combining listing titles and body text from the Craigslist dataset. Standard preprocessing was applied using NLTK – lowercasing, removing alphanumeric characters, tokenization, custom stopword filtering, and lemmatization. We remove common real estate jargon that would otherwise dominate the topic distributions without providing meaningful differentiation between the content in different listing types. For the LDA implementation we selected hyperparameters that allow for moderate document sparsity ($\alpha = 0.1$) and greater topic-word concentration ($\eta = 0.01$). We trained for 100 iterations and tried $k = 5, 7$ and 10 topics which returned coherence scores of `0.7344`, `0.7425` and `0.7509`, respectively. After manual review we decided to focus on the $k = 7$ results for our remaining analysis as these had clearer

Topics	Common Words
1. Furnished Short-Term *3.6% of corpus*	lease, home, month, furnished, mo, amenity, community, view, apartment, offer, blueground, access, neighborhood, enjoy, stay, loop
2. Rental Terms *25.9% of corpus*	fee, lease, included, tenant, credit, street, pay, deposit, application, large, move, heat, per, gas, security, pet, utility
3. Property Search *1.1% of corpus*	apartment, rental, property, place, hill, cheap, grove, find, apt, height, center, agency, search, local, credit
4. Spanish Language *2.1% of corpus*	apartment, rental, property, place, hill, cheap, alquilar, propiedades, buscar, alquileres, height, search, agency, google
5. Amenities *35.0% of corpus*	floor, central, new, appliance, dishwasher, heat, large, space, air, stainless, feature, modern, steel, living, cat, closet
6. Listing Conditions *13.3% of corpus*	subject, unit, change, price, center, property, onsite, hour, availability, special, dog, pricing, studio, fitness, housing, amenity
7. Contact Information *18.9% of corpus*	contact, info, show, click, feature, view, call, id, renovated, , included, closet

Table 3: Topic interpretations based LDA topic modeling on Chicago Craigslist rental listings.

separation than the $k = 5$ topics and are more interpretable than the $k = 10$ results.

5.1 Topic Interpretations

We identify seven distinct topics in the Craigslist advertisements, shown in Table 3. These topics reveal distinct patterns in how rental listings are framed, which have important implications for understanding the conception of neighborhood reputation and representation in the online rental market.

Topic 1 focuses on furnished short-term rentals, highlighting amenities and comfort with words like “furnished," “month," and “stay," suggesting a market segment catering to temporary residents seeking turnkey living situations. Topic 2 centers on rental requirements and financial considerations, with terms like “fee," “credit," “deposit," and “application," reflecting the administrative and financial aspects of renting. Topics 3 and 4 both relate to property search, with Topic 4 specifically including Spanish-language terms like “alquilar" and “propiedades," indicating efforts to reach Spanish-speaking audiences in Chicago’s rental market. Topic 5, the most prevalent across the corpus at approximately 35% of document content, focuses on apartment features and amenities such as “appliance," “stainless," “modern," and “dishwasher," underscoring the prevalence of interior quality in marketing rental properties. Topic 6 addresses listing conditions and availability, featuring terms related to pricing, special offers, and property policies. Finally, Topic 7 concentrates on contact information and viewing arrangements, with words like “contact," “show," “click," and “schedule," facilitating the connection between potential renters and property managers. The distribution of these topics across advertisements reveals how Chicago’s rental market is presented online, with physical features and financial terms dominating the discourse.

5.2 Regression Analysis

We estimate the relationship between listing characteristics and proximity to the social center of an associated neighborhood using OLS regression with relative distance to neighborhood center as our primary dependent variable. Our model includes unit characteristics (bedrooms, bathrooms, rent, and square footage) and the topic proportions identified in our LDA analysis. For a full table of regression outputs see Table 5 in the Appendix.

This analysis reveals several patterns in spatial representations. Advertisements with more bedrooms/bathrooms are associated with being further from neighborhood centers, while higher-priced listings are closer to their respective social centers. Most notably, listings with higher proportions of Topic 3 (Property Search) content exhibit increased relative distance from the center of the neighborhood (+0.20, $p<0.001$). Conversely, listings emphasizing apartment amenities (Topic 5) are associated with being closer to the centroid (-0.03, $p<0.01$). These findings indicate that misrepresentation is not random but correlates with specific marketing approaches in rental listings.

6 Results

Our analysis of over 30,000 unique Chicago rental listings reveals that neighborhood boundaries are actively renegotiated through strategic linguistic claims. By establishing a “social center” for neighborhoods based on the density of textual claims rather than rigid municipal boundaries, we demonstrate how agents navigate the tradeoff between geographic reality and reputational leverage. The following sections detail our findings in relation to our primary research questions.

6.1 Labeling Viability (RQ1)

We find that Large Language Models are highly effective for identifying neighborhood claims within unstructured text, outperforming traditional string-matching techniques. While the accuracy gain is incremental (85% for GPT-4.1 mini vs. 79% for string-matching), the LLM approach can scale beyond Chicago to other urban contexts without requiring reconfiguration or the same level of local real estate knowledge. Our labeling system prioritizes precision by selecting the single strongest neighborhood claim, addressing the inherent ambiguity found in listings that mention multiple areas.

6.2 Defining Social Centers (RQ2)

Our geospatial analysis reveals that the "social center" of a neighborhood—defined as the centroid of all property listings claiming that identity—often aligns with local landmarks, such as the eponymous park in Humboldt Park. However, these Craigslist-defined centers frequently diverge from institutional boundaries. We identify significant variation in representation: neighborhoods like Lake View are heavily overrepresented in claims relative to their Zillow-defined footprints, while areas such as Englewood and North Austin are significantly underrepresented, suggesting a lack of strong neighborhood identity or the presence of territorial stigma in the rental market.

6.3 Linguistic Substitution (RQ3)

Our analysis characterizes spatial misrepresentation not as random market noise, but as a predictable distortion of urban space. We identify three distinct patterns of spatial claim discrepancies: (1) *Conflicting Conceptions*, where stakeholders disagree on boundaries (e.g. Humboldt Park in Figure 1); (2) *Border Stretching*, where listings claim adjacent, plausible neighborhoods (e.g. Pilsen in Figure 3); and (3) *Reputation Laundering*, where properties or peripheral areas claim association with distant, desirable neighborhoods.

To identify these patterns systematically, we operationalize "peripheral claims" as those located in the furthest 20% from a neighborhood's social centroid. By comparing these LLM-labeled claims against institutional Zillow boundaries, we find evidence of systematic over and underrepresentation. Specifically, Lake View is significantly overrepresented–claimed more frequently than geographic distributions would predict–while neighborhoods such as Englewood, North Austin, and Hanson Park are significantly underrepresented. This pattern is consistent with reputation laundering: agents appear to distance properties from stigmatized neighborhood names while strategically claiming higher-status alternatives.

This substitution is statistically observable through a compensatory language pattern. Regression results show that as properties move further from the neighborhood social center, listing agents substitute symbolic identity for physical proximity. For instance, generic property search language (Topic 3) is positively associated with relative distance from centroid ($+0.20, p < 0.001$), while specific interior amenity language (Topic 5) is negatively associated ($-0.03, p < 0.01$). Further, compared to central listings, the language used in peripheral listings shifts from location specific, non-portable amenities (Topics 1,5) toward more abstract and generic property-search language (Topics 2,3,4,6,7), a pattern illustrated in Figure 1.

7 Implications Beyond Sociology

The use of LLMs has exploded in recent years, and they can be seen by the general public as a simple, reliable solution to many routine problems. However, it remains an open question how powerful they may be in interdisciplinary research (Ziems et al., 2024). In order to better understand the impact of NLP on a broader scale and help address a specific question in the field of Urban Sociology, we demonstrate the effectiveness of LLMs at a notoriously difficult task: identifying where rental listings claim to be located on a large scale, in order to inform our understanding of processes of social construction of urban space.

However, although the impact of language models on this task may be transformative in the ability to quickly expand the scope of analysis, the quantifiable improvement on simple algorithms designed by experts is perhaps more incremental. In addition, the LLM labeling was certainly not perfectly accurate, suggesting that there is still room for improvement in large language models that might not be captured by tests and benchmarks developed solely within the field of NLP, and that interdisciplinary collaboration could lead to improvements both in NLP methodology and in making research questions and analyses in a broad range of fields more tractable and scalable.

8 Limitations

Our analysis of Craigslist rental listings provides valuable insights into neighborhood claims and social construction, but several important limitations should be acknowledged. The process of collecting data from Craigslist presents inherent challenges regarding post uniqueness and identification. Despite our deduplication efforts, the platform's structure makes it difficult to definitively determine which posts represent truly unique listings versus slightly modified versions of the same property.

While our dataset offers substantial coverage across Chicago neighborhoods, it cannot claim to be fully representative of the entire rental market. Craigslist represents just one segment of available rental properties, potentially skewing toward certain price points or property types. Additionally, our data may over represent certain management companies and realtors who post frequently on the platform, as opposed to "mom and pop" owners. These high-volume posters are more likely to use standardized language across multiple listings, which may introduce uniformity in how neighborhoods are described that doesn't reflect broader market patterns.

A fundamental challenge in this research is the absence of an authoritative catalog of Chicago's neighborhoods. As we argue, such a catalog is conceptually impossible. For practical purposes, we relied on Zillow's neighborhood boundaries as our primary reference, but these designations do conflict with local understandings. For example, questions arise about whether "West Loop" constitutes its own neighborhood distinct from "West Loop Gate," or whether "East Rogers Park" should be considered separate from "Rogers Park." These ambiguities reflect the socially constructed nature of neighborhoods themselves. Also, many units along Lake Michigan have a view of the water, and therefore advertise "views of the lake" or "lake views" which can be impossible to distinguish from listings claiming to be a "lake view" without considering the corresponding latitude and longitude coordinates.

Furthermore, assigning a single neighborhood label to each listing proved challenging when advertisements contained multiple, sometimes competing neighborhood claims. Our hierarchical labeling approach (prioritizing title, then body, then neighborhood field) necessarily simplifies what can be complex spatial positioning strategies employed by listing agents. A rental advertisement might strategically claim association with multiple neighborhoods simultaneously, a nuance our single-label framework cannot fully capture. For instance, a listing located on the border between Logan Square and Humboldt Park might leverage both neighborhood identities depending on the perceived audience and market conditions.

These limitations highlight the inherent complexity of studying socially constructed spatial boundaries through digital traces and suggest opportunities for future research employing more nuanced approaches to neighborhood identification and classification.

Ethics Statement

Deciding how to approach this analysis is a non-trivial decision and following the extensive work in sociology and economics is important. We spoke experts in both fields in scoping and executing this project.

We collect data from Craigslist which, in some cases, contains specific information about individual posters. Craigslist is a public forum whose housing section should not contain much information irrelevant to the housing ads themselves. We do not release these data publicly per the non-commercial use terms of service and ensure no PII appear in any of our writing, results or figures.

Another limitation when working with pre-trained foundation models like GPT-4 is a lack of reproducibility, as we do not have access to the training data or the weights. In the interest of reproducibility, we keep annotation costs under $100.

We utilized multiple Generative AI tools (OpenAI's GPT-4/5 and Anthropic's Claude 4.5 Opus/Claude Code) in the production of this manuscript, in the following ways: 1) producing computer code for data cleaning and analysis and 2) iteratively improving the concision and clarity of the writing. We have carefully reviewed all aspects of the manuscript for accuracy and coherence. All scientific insights, analysis and interpretation of data and scientific conclusions are made solely by the authors.

9 Acknowledgements

We thank Drew Messamore, Diag Davenport, and Michael Reher for providing valuable suggestions on an earlier version of this manuscript. Adam gratefully acknowledges the resources provided

by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). We are also grateful to the University of Washington's Department of Sociology writing workshop for their comments and feedback and to the University of California, Berkeley Bellwether Postdoctoral program.

References

Kenan Alkiek, Anna Wegmann, Jian Zhu, and David Jurgens. 2025. Neurobiber: Fast and interpretable stylistic feature extraction. *arXiv preprint arXiv:2502.18590*.

Abdulkareem Alsudais. 2020. Incorrect data in the widely used inside airbnb dataset. *Decision Support Systems*, 141:113453.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Max Besbris, Ariela Schachter, and John Kuk. 2021. The unequal availability of rental housing information across neighborhoods. *Demography*, 58(4):1197–1221.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, pages 1–4.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Geoff Boeing and Paul Waddell. 2017. New insights into rental housing markets across the united states: Web scraping and analyzing craigslist rental listings. *Journal of Planning Education and Research*, 37(4):457–476.

Geoff Boeing, Max Besbris, Ariela Schachter, and John Kuk. 2021. Housing search in the age of big data: smarter cities or the same old blind spots? *Housing Policy Debate*, 31(1):112–126.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Sarah E. Chasins, Maria Mueller, and Rastislav Bodik. 2018. Rousillon: Scraping distributed hierarchical web data. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 963–975.

Eric Chyn and Lawrence F. Katz. 2021. Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives*, 35(4):197–222.

Megan Evans and Barrett A. Lee. 2020. Neighborhood reputations as symbolic and stratifying mechanisms in the urban hierarchy. *Sociology Compass*, 14(10):1–15.

George Galster and Erin Godfrey. 2005. By words and deeds: Racial steering by real estate agents in the us in 2000. *Journal of the American Planning Association*, 71(3):251–268.

Edward L. Glaeser, Michael Luca, and Erica Moszkowski. 2020. Gentrification and neighborhood change: Evidence from yelp. *National Bureau of Economic Research*.

Igal Hendel, Aviv Nevo, and François Ortalo-Magné. 2009. The relative performance of real estate marketing platforms: Mls versus fsbomadison. com. *American Economic Review*, 99(5):1878–1898.

Chris Hess and Sarah E. Chasins. 2022. Informing housing policy through web automation: Lessons for designing programming tools for domain experts. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9.

Chris Hess, Rebecca J. Walter, Ian Kennedy, Arthur Acolin, Alex Ramiller, and Kyle Crowder. 2023. Segmented information, segregated outcomes: Housing affordability and neighborhood representation on a voucher-focused online housing platform and three mainstream alternatives. *Housing Policy Debate*, 33(6):1511–1535.

Randolph Hohle. 2023. Rusty gardens: stigma and the making of a new place reputation in buffalo, new york. *American Journal of Cultural Sociology*, 11(2):193–219.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Biqing Huang and Ronald Rutherford. 2007. Who you going to call? performance of realtors and non-realtors in a mls setting. *The Journal of Real Estate Finance and Economics*, 35:77–93.

Jackelyn Hwang and Robert J. Sampson. 2014. Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods. *American Sociological Review*, 79(4):726–751.

Ronda Kaysen. 2024. Apartment rent renewal rates are rising. *The New York Times*.

Ian Kennedy, Chris Hess, Amandalynne Paullada, and Sarah Chasins. 2020. Racialized discourse in seattle rental ad texts. *Social Forces*, 99(4):1432–1456.

Richard Kirk. 2024. Legitimising displacement: Academic discourse, territorial stigmatisation and gentrification. *Urban Studies*, 61(13):2492–2512.

Elizabeth Korver-Glenn and Sarah Mayorga. 2024. *A good reputation: how residents fight for an American barrio*. University of Chicago Press.

Maria Krysan and Kyle Crowder. 2017. *Cycle of Segregation: Social Processes and Residential Stratification*. Russell Sage Foundation, New York.

Agneta Kullberg, Toomas Timpka, Tommy Svensson, Nadine Karlsson, and Kent Lindqvist. 2010. Does the perceived neighborhood reputation contribute to neighborhood differences in social trust and residential wellbeing? *Journal of Community Psychology*, 38(5):591–606.

Anita Minh, Nazeem Muhajarine, Magdalena Janus, Marni Brownell, and Martin Guhn. 2017. A review of neighborhood effects and early child development: How, where, and for whom, do neighborhoods matter? *Health & Place*, 46:155–174.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828.

NYC Office of Special Enforcement. 2023. Short-term rental registration and verification by booking services.

Gabriel Otero, Quentin Ramond, María Luisa Méndez, Rafael Carranza, Felipe Link, and Javier Ruiz-Tagle. 2024. The damages of stigma, the benefits of prestige: Examining the consequences of perceived residential reputations on neighbourhood attachment. *Urban Studies*, 61(3):462–494.

Jeffrey Nathaniel Parker. 2019. *That Kind of Neighborhood: Creating, Contesting, and Commodifying Place Reputation*. Ph.D. thesis, The University of Chicago.

M. Permentier, G. Bolt, and M. Van Ham. 2011. Determinants of neighbourhood satisfaction and perception of neighbourhood reputation. *Urban Studies*, 48(5):977–996.

Kirsten Robertson and Antony Doig. 2010a. An empirical investigation of variations in real-estate marketing language over a market cycle. *Housing, Theory and Society*, 27(2):178–189.

Kirsten Robertson and Antony Doig. 2010b. An empirical investigation of variations in real-estate marketing language over a market cycle. *Housing, Theory and Society*, 27(2):178–189.

Robert J. Sampson, Jeffrey D. Morenoff, and Thomas Gannon-Rowley. 2002. Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, 28(1):443–478.

Ariela Schachter, John Kuk, Max Besbris, and Lydia Ho. 2024. What's in a name? place misrepresentation and neighbourhood stigma in the online rental market. *Urban Studies*, 61(16):3050–3068.

Youngme Seo, JongHo Im, and Brian Mikelbank. 2020. Does the written word matter? the role of uncovering and utilizing information from written comments in housing ads. *Journal of Housing Research*, 29(2):133–155.

Patrick Sharkey and Jacob W. Faber. 2014. Where, when, why, and for whom do residential contexts matter? moving away from the dichotomous understanding of neighborhood effects. *Annual Review of Sociology*, 40(1):559–579.

Mahesh Somashekhar, Chris Hess, Ian Kennedy, and Kyle Crowder. 2024. How do real estate actors advertise in mixed-income neighborhoods? the importance of home security. *Socius*, 10:23780231241260253.

Remy Stewart, Chris Hess, Ian Kennedy, and Kyle Crowder. 2023. Move-in fees as a residential sorting mechanism within online rental markets. *Cityscape (Washington, DC)*, 25(1):239.

Forrest Stuart, Charles R. Collins, Bocar Wade, Rebecca D. Gleit, and Caylin Louis Moore. 2024. Where do neighbourhood reputations come from? analysing chicago community areas using a systematic neighbourhood reputation score, 1985–2020. *Urban Studies*, pages 00420980241297088.

Emma Tran, Kim Blankenship, Shannon Whittaker, Alana Rosenberg, Penelope Schlesinger, Trace Kershaw, and Danya Keene. 2020. My neighborhood has a good reputation: Associations between spatial stigma and health. *Health & Place*, 64:102392.

United States Census Bureau, the. 2023. American community survey 5-year data (2009-2021).

Mohammadzaman Zamani and H. Andrew Schwartz. 2017. Using Twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 28–33, Valencia, Spain. Association for Computational Linguistics.

Sarah Zelner. 2015. The perpetuation of neighborhood reputation: An interactionist approach. *Symbolic Interaction*, 38(4):575–593.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Livia Hollenstein and Ross Purves. 2010. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1:1–18.

Anon Plangprasopchok and Kristina Lerman. 2009. Constructing folksonomies from user-specified relations on Flickr. In *Proceedings of the 18th International Conference on World Wide Web*, pages 781–790.

Mikael Brunila, Jack LaViolette, Sky CH-Wang, Priyanka Verma, Clara Féré, and Grant McKenzie. 2023. Toward a Critical Toponymy Framework for Named Entity Recognition: A Case Study of Airbnb in New York City. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4676–4695, Singapore. Association for Computational Linguistics.

Tuomo Hiippala, Anna Hausmann, Henrikki Tenkanen, and Tuuli Toivonen. 2019. Exploring the Linguistic Landscape of Geotagged Social Media Content in Urban Environments. *Digital Scholarship in the Humanities*, 34(2):290–309.

A Full Performance Metrics for Neighborhood Claim Validation

A.1 Topic Modeling Results

Working with more than 35,000 unique and often verbose advertisements, we fit a topic model on $k = 25$ topics to identify main themes in the rental listings. We do not provide interpretations of the topic, but show the common words in Table 6.

	String Match			GPT-4 Mini			
Neighborhood	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Support**
the loop	0.12	0.50	0.20	1.00	1.00	1.00	2
rogers park	1.00	1.00	1.00	0.75	1.00	0.86	3
lake view	0.63	0.86	0.73	0.94	0.73	0.82	22
ranch triangle	–	0.00	0.00	–	0.00	0.00	1
lincoln park	0.90	1.00	0.95	0.94	0.89	0.91	18
fulton river district	–	0.00	0.00	1.00	1.00	1.00	1
south loop	1.00	1.00	1.00	1.00	1.00	1.00	3
park west	–	0.00	0.00	–	0.00	0.00	1
west town	1.00	1.00	1.00	1.00	1.00	1.00	1
logan square	0.88	0.88	0.88	0.89	1.00	0.94	8
lake view east	–	0.00	0.00	0.90	0.90	0.90	10
university village - little italy	–	0.00	0.00	–	0.00	0.00	2
uptown	1.00	0.83	0.91	1.00	0.83	0.91	6
wicker park	1.00	1.00	1.00	1.00	1.00	1.00	4
portage park	1.00	1.00	1.00	1.00	1.00	1.00	2
old town	1.00	0.83	0.91	1.00	0.83	0.91	6
avondale	–	0.00	0.00	–	0.00	0.00	1
west loop gate	–	0.00	0.00	1.00	1.00	1.00	4
streeterville	1.00	1.00	1.00	1.00	1.00	1.00	2
ravenswood	1.00	0.80	0.89	1.00	0.80	0.89	5
wrigleyville	1.00	0.67	0.80	1.00	1.00	1.00	3
river north	0.83	0.71	0.77	0.86	0.86	0.86	7
buena park	1.00	1.00	1.00	0.75	1.00	0.86	3
bucktown	1.00	1.00	1.00	1.00	1.00	1.00	7
kenwood	–	0.00	0.00	1.00	1.00	1.00	1
old irving park	1.00	1.00	1.00	1.00	1.00	1.00	1
edgewater	1.00	1.00	1.00	1.00	1.00	1.00	5
pilsen	1.00	1.00	1.00	1.00	1.00	1.00	7
garfield ridge	1.00	1.00	1.00	1.00	1.00	1.00	1
humboldt park	1.00	1.00	1.00	1.00	1.00	1.00	1
andersonville	1.00	1.00	1.00	1.00	0.67	0.80	3
west rogers park	1.00	1.00	1.00	1.00	1.00	1.00	1
ukrainian village	1.00	1.00	1.00	1.00	1.00	1.00	1
gold coast	0.75	1.00	0.86	0.75	1.00	0.86	3
hyde park	1.00	1.00	1.00	1.00	1.00	1.00	1
roscoe village	1.00	1.00	1.00	0.33	1.00	0.50	2
east garfield park	1.00	1.00	1.00	1.00	1.00	1.00	1
albany park	1.00	1.00	1.00	1.00	1.00	1.00	1
lincoln square	0.33	1.00	0.50	0.00	0.00	0.00	1
south shore	1.00	1.00	1.00	1.00	1.00	1.00	1
hermosa	1.00	1.00	1.00	0.50	1.00	0.67	1
ravenswood manor	–	0.00	0.00	–	0.00	0.00	1
unknown	0.78	0.88	0.82	0.50	0.50	0.50	8
Average Metrics							**Total: 163**
Accuracy		0.79			0.85		
Macro Avg	0.76	0.74	0.62	0.78	0.81	0.70	
Weighted Avg	0.88	0.79	0.77	0.91	0.85	0.85	

Table 4: Detailed performance metrics by neighborhood for string matching and GPT-4 Mini classification methods. The Support column indicates the number of test samples for each neighborhood. Dashes indicate cases where precision could not be calculated due to zero predictions.

Variable	Coef.	Std. Err.	t	P>|t|
Intercept	0.10	0.01	10.40	0.00
Bedrooms	0.02	0.00	8.83	0.00
Bathrooms	0.02	0.00	6.25	0.00
Rent	0.00	0.00	-14.56	0.00
Square Footage	0.00	0.00	2.17	0.03
Topic 2 Rental	0.03	0.01	3.40	0.00
Topic 3 Property	0.20	0.01	14.99	0.00
Topic 4 Spanish	0.10	0.01	9.42	0.00
Topic 5 Amenities	-0.03	0.01	-2.98	0.00
Topic 6 Condition	0.02	0.01	1.86	0.06
Topic 7 Contact	0.05	0.01	5.56	0.00

Table 5: Regression results demonstrating how relative distance from the center of the neighborhood associates with rental unit characteristics.

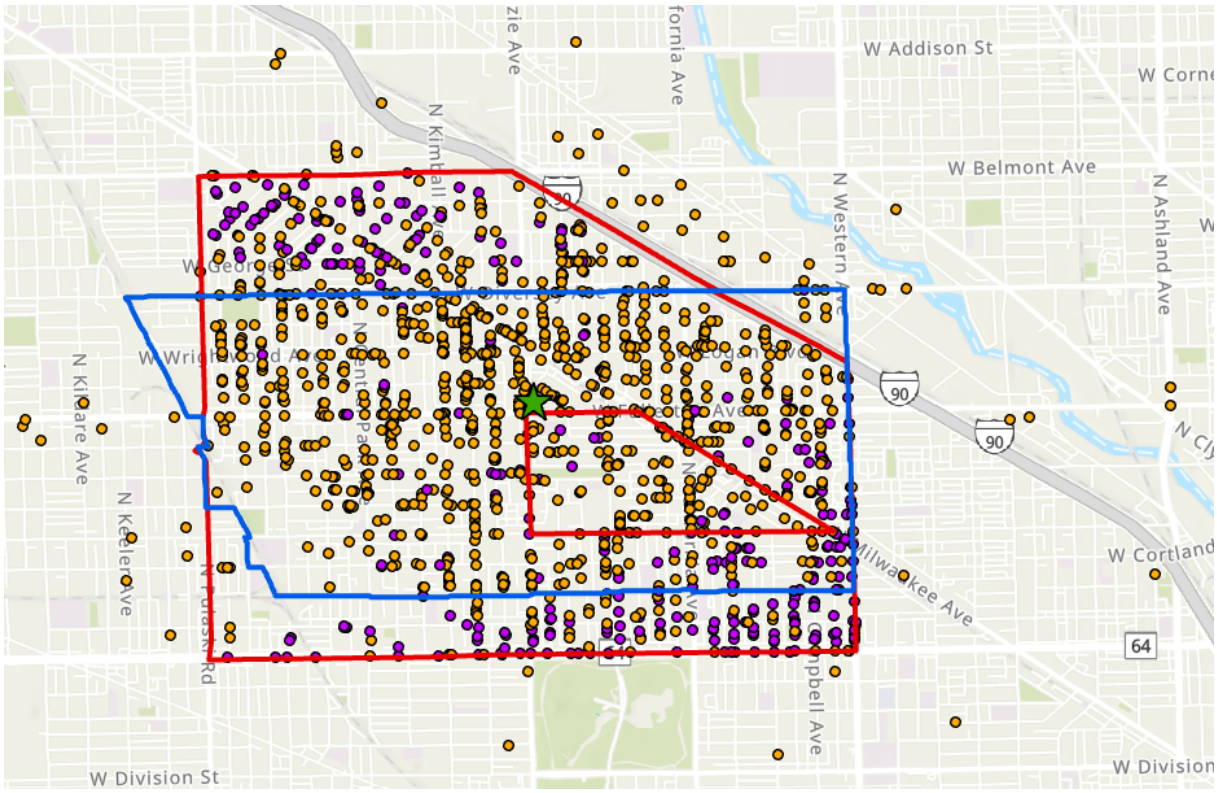


Figure 4: Contested boundaries for Logan Square neighborhood according to the official City of Chicago limits (blue), Zillow's definition (red), and claims in Craigslist rental listings (orange points for listings claiming to be in the neighborhood, purple if claiming to be elsewhere); the center of our Craigslist-defined neighborhood is represented by a green star. Logan Square represents a popular neighborhood with many postings and a number of nearby listings claiming to be within the neighborhood.

Table 2: Topics for $k = 25$

Topic	Common Words
Topic 1	hyde, property, si, terrace, la, estos, con, mac, elli, street
Topic 2	large, space, floor, living, dining, closet, bedroom, heat, storage, central
Topic 3	contact, call, schedule, photo, tour, please, unit, info, show, actual
Topic 4	space, walk, home, lease, great, living, street, one, loft, large
Topic 5	real, estate, star, text, tour, community, view, lounge, apartment, finish
Topic 6	lease, mo, blueground, month, furnished, amenity, stay, offer, view, access
Topic 7	cat, feature, floor, call, dishwasher, one, fee, lakeview, n, dog
Topic 8	modern, heat, property, central, bus, appliance, call, gas, cat, air
Topic 9	apartment, rental, lake, property, place, hill, alquilar, cheap, new, grove
Topic 10	unit, special, availability, subject, dog, weight, onsite, please, price, various
Topic 11	view, center, lake, amenity, free, onsite, community, michigan, call, window
Topic 12	fee, credit, deposit, tenant, new, security, included, move, pay, pet
Topic 13	hyde, apartment, regent, horas, onsite, market, la, est, e, museum
Topic 14	apartment, rental, lake, property, place, hill, cheap, find, new, grove
Topic 15	studio, property, bjb, internet, complimentary, e, change, picture, subject, fitness
Topic 16	fee, mile, white, home, tile, neighborhood, make, company, new, tenant
Topic 17	housing, water, included, heat, opportunity, landstar, equal, group, act, feature
Topic 18	contact, show, info, price, availability, email, renovated, included, click, web
Topic 19	lake, bus, block, cta, walk, red, minute, away, stop, distance
Topic 20	info, contact, show, click, il, id, feature, friendly, text, n
Topic 21	amenity, view, center, pool, fitness, luxury, outdoor, lounge, garage, window
Topic 22	logan, banker, coldwell, square, blue, real, estate, please, opportunity, equal
Topic 23	hyde, village, height, center, drexel, grand, nuestros, river, maintenance, m
Topic 24	fee, per, cat, lease, dog, application, deposit, one, heat, gas
Topic 25	new, appliance, stainless, steel, floor, central, large, feature, granite, dishwasher

Table 6: Topic clusters from LDA topic modeling on Chicago Craigslist rental listings, $k = 25$.

The Hidden Language of Harm: Examining the Role of Emojis in Harmful Online Communication and Content Moderation

Yuhang Zhou **Yimin Xiao** **Wei Ai** **Ge Gao**
University of Maryland, College Park
{tonyzhou, yxiao, aiwei, gegao}@umd.edu

Abstract

Social media platforms have become central to modern communication, yet they also harbor offensive content that challenges platform safety and inclusivity. While prior research has primarily focused on textual indicators of offense, the role of emojis, ubiquitous visual elements in online discourse, remains underexplored. Emojis, despite being rarely offensive in isolation, can acquire harmful meanings through symbolic associations, sarcasm, and contextual misuse. In this work, we systematically examine emoji contributions to offensive Twitter messages, analyzing their distribution across offense categories and how users exploit emoji ambiguity. To address this, we propose an LLM-powered, multi-step moderation pipeline that selectively replaces harmful emojis while preserving the tweet's semantic intent. Human evaluations demonstrate that our approach effectively reduces offensiveness while preserving semantic integrity. Our analysis also reveals heterogeneous effects across offense types, offering nuanced insights for online communication and emoji moderation.

1 Introduction

Social media platforms host an incredibly diverse range of content, which is central to how people communicate online. However, due to varying degrees of censorship policies, platforms like Twitter often become repositories for offensive language, threatening the cohesion and safety of online communities (Davidson et al., 2019). When analyzing offensive tweets, most research has focused on textual elements—explicit slurs, abusive phrases, or implicit language that reflects social biases (Caselli et al., 2020; Zampieri et al., 2019). In response, scholars have developed various approaches, many leveraging Large Language Models (LLMs), to detect offensive content across cultural and linguistic contexts (Zhou et al., 2023a; Deng et al., 2022a). Yet, despite these efforts, one critical aspect of online communication has been largely overlooked: the role of emojis in conveying offensive messages.

Emojis, as visual symbols, are embedded in the context of communication and carry more complex semantics than individual words. On the one hand, very few emojis directly convey offensive meanings: exceptions include emojis like 🖕 (middle finger) and 💩 (pile of poop), as emojis are generally not designed with the intent to offend. On the other hand, the widespread use of emojis leads to varying interpretations. Emojis, with their symbolic representation of objects or ideas through similar shapes, can convey offensive meanings. For example, users often use the 🍑 (peach) emoji to symbolize buttocks and the 💦 (droplets) emoji to symbolize sperm. Moreover, for sentiment-related emojis, one of the key characteristics of emojis is their ability to express irony or sarcasm (Hu et al., 2017). Emojis such as 🙃 (upside-down face) and 🤣 (rolling on the floor laughing) are often used to intensify offense by conveying a sarcastic tone. Even emojis typically associated with positive sentiment, such as 😍 (smiling face with heart-eyes), can take on an offensive meaning when used in inappropriate contexts, such as sexual harassment.

Given the subtle yet potent ways in which emojis contribute to offensive communication, it is essential to systematically examine their roles within online discourse. We begin by identifying emojis frequently found in offensive tweets and analyzing how they relate to different types of offensive content. To deepen our understanding, we classify offensive tweets by category and investigate which emojis are commonly used within each category.

While content moderation has traditionally focused on text (Zampieri et al., 2019; Pitsilis et al., 2018; Husain and Uzuner, 2021), we argue that emojis present a common jailbreaking way that users exploit, either deliberately or unintentionally, to convey offensive meaning through stereotypi-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 322–340

cal associations. To empower users to navigate this complex landscape, we propose an audience-oriented, mitigation-focused pipeline powered by LLMs. Rather than resorting to full-text rewriting, which can eliminate linguistic nuance, our approach performs a targeted emoji replacement. This lightweight intervention is designed to reduce the perceived offense for the viewer while preserving the semantic intent. The pipeline is implemented to identify emojis that have the potential to evoke offense, and recommend emoji surrogates that preserve the tweet's semantics. Human evaluations show that this pipeline effectively reduces offensiveness while maintaining the tweet's meaning. We also analyze its heterogeneous effects across different tweet types and examine the relationship between emoji functionality and offensiveness.

We summarize our contributions as follows:

- We explore the relationship between emojis and offensive content in online communication, examining the roles emojis play under different offensive types.
- We design and implement a multi-step LLM pipeline to better moderate offensive emojis in tweets and recommend emoji surrogates.
- We conduct a human evaluation to demonstrate the effectiveness of our pipeline and analyze its heterogeneous effects across offensive types.

2 Related Work

Our work is based on two lines of existing work: emoji functionality and offensive content detection.

Emoji Functionality and Interpretation Emojis, as prevalent visual elements, have attracted the interest of researchers. The semantics embedded in emojis extend beyond a single word token, giving their various functionalities such as expressing sentiments and irony, softening tones, and enhancing communication (Ai et al., 2017; Ge, 2019; Hu et al., 2017; Miller et al., 2016; Cramer et al., 2016). The rich meanings and diverse functionalities of emojis make them useful in various tasks, including sentiment analysis, predicting user behavior, and increasing communication (Felbo et al., 2017; Chen et al., 2018, 2019; Zhou et al., 2023b; Zhou and Ai, 2022).

Offensive Content Detection The prevalence of social networks has encouraged users to develop more flexible forms of offensive behavior. Researchers have examined patterns of offense in online communication and developed various methods to detect and mitigate offensive content in text (Davidson et al., 2019, 2017; Pitsilis et al., 2018; Poletto et al., 2021). There is increasing interest in leveraging them for the effective detection of hate speech and other hidden performance bias (Huang et al., 2023; Li et al., 2023; Zhu et al., 2023; Zhou et al., 2025). Furthermore, due to their text generation capabilities, some studies have used LLMs to augment collected datasets, thus enhancing the robustness of hate speech detection models (Xiao et al., 2024; Nghiem and Daumé III, 2024).

Beyond general offense, researchers have explored offenses of different types (Vandenbosch et al., 2015; Davidson et al., 2019; Zhong et al., 2019), as each offense type tends to target different groups. Given the strong link between offense and culture, researchers have also explored offensive content across multiple languages (Pitsilis et al., 2018; Deng et al., 2022b; Husain and Uzuner, 2021; Battistelli et al., 2020) and more increasingly diverse datasets have emerged to enhance the detection of offensive content. A few recent studies study the value of emojis as signals for offensive language detection (Kirk et al., 2021; Mubarak et al., 2023; Wiegand and Ruppenhofer, 2021). Furthermore, our research contributes to the understanding emoji-based offense in two unique ways. First, we conduct a systematic, bottom-up analysis of how a wide range of emojis contribute to offensive language. Second, building on this systematic understanding, we propose a pipeline that aims to reduce offensiveness through targeted emoji replacement, beyond the task of detection.

3 Emojis in Offensive Contexts

We begin our exploration of emoji functionality in offensive contexts by analyzing their roles and distributions. To identify offensive content, we first collected a broad, random sample of public tweets geolocated to the U.S. from January 1 to December 31, 2019, via the Twitter API[1]. We then employed a two-step approach to create a high-quality dataset of offensive tweets. Given the prohibitive cost of applying LLMs to our full one-year dataset, we first used the efficient finetuned RoBERTa model

[1] https://developer.twitter.com/en/docs/twitter-api

[2] to perform a broad initial filtering (Liu et al., 2019; Barbieri et al., 2020). The goal of this step was high recall to capture a wide range of potentially offensive content. Tweets with a predicted probability greater than 0.5 were selected, and this smaller subset was then processed by GPT-4 for a high-fidelity classification of whether the tweet contained offensive content (OpenAI, 2023). Detailed prompts can be found in Appendix A.2. Our choice of GPT-4 for this large-scale annotation task was informed by prior work demonstrating that LLMs show strong agreement with human judgments on nuanced social computing tasks, including emoji interpretation (Zhou et al., 2024b; Lyu et al., 2024).

This process resulted in 9,285 annotated offensive tweets. To ensure precise annotation, we define offensive content as posts containing unacceptable language (profanity) or targeted offenses, whether direct or veiled, including insults, threats, profane language, or swear words, following the definition used in previous work (Poletto et al., 2021; Zampieri et al., 2019). Moreover, to validate the annotation quality, one author manually reviewed a random sample of 100 tweets and confirmed 91 out of 100 as offensive tweets.

3.1 Emoji Role in Offensive Tweets

To understand how emojis function in offensive tweets, we developed a taxonomy grounded in the literature on emoji functions (Section 2) and their interaction with offensive language. We categorize emojis into four roles:

- **Offensive in itself.** The emoji alone constitutes an offense, such as 🖕 (middle finger).
- **Intensify offense.** Emojis can enhance the intensity of an offensive tweet by expressing irony or sarcasm (Weissman and Tanner, 2018), thereby amplifying its offensive nature.
- **Mitigate offense.** Emojis can also soften or adjust the tone of a tweet, reducing its offensive impact (Cramer et al., 2016; Ge, 2019).
- **Unrelated to offense.** The emoji is not directly connected to the offensive content of the tweet.

Based on the proposed taxonomy, we use GPT-4 to annotate the role of all emojis present in the collected offensive tweets. Furthermore, to validate the annotation quality, one author annotated a random sample of 100 tweets from the dataset and confirmed an agreement of 83%. The distribution of emoji roles is illustrated in Figure 1. Detailed prompts can be found in Appendix A.2.

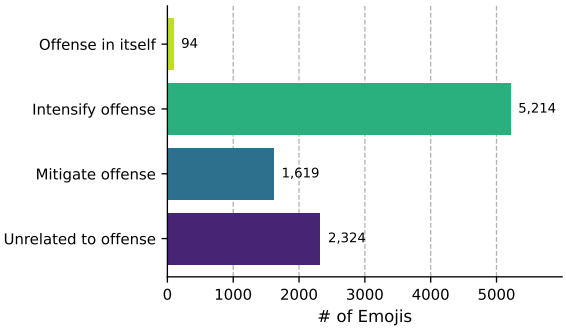


Figure 1: Distribution of emoji role in offensive tweets

Role	Top 10 Frequent Emojis
Offensive in itself	🤡💩🤢🤬🖕😈🍆🤮🍑😡
Intensify offense	😂😭🤣🙄💀😩😫😒💯😍
Mitigate offense	😂😭🥺😔🤣😅😫😩❤️😢
Unrelated to offense	😂😭😍🤣🤔🔥😫😩❤️👀

Table 1: Top-10 emojis under each emoji role in offensive tweets

The distribution (Figure 1) reveals that most annotated emojis intensify, mitigate, or are unrelated to the offense, with few being offensive in themselves, as expected given that few emojis carry inherently offensive meanings. The prevalence of intensification suggests a notable link between emojis and offensive expression. Table 1 lists the top 10 emojis for each role. Emojis categorized as "Offensive in itself" (e.g., 💩, 🖕) show no overlap with other categories and are often used for direct insults. Conversely, significant overlap exists among the top emojis for the other roles. Emojis like 😂 and 😭 appear across these categories, highlighting their context-dependent functions. Notably, even positive emojis like 😍 can intensify offense, particularly in contexts like sexual harassment.

Given that specific emojis (e.g., 🍑) seem linked to particular offensive themes, we next apply topic modeling to further explore emoji usage across different types of offensive content.

3.2 Emojis Associated with Different Offensive Topics

To better understand the offensive context in which each emoji appears, we identify the latent offensive types embedded in each tweet and explore

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

the emojis associated with each specific type. To summarize the offensive types across tweets, we first clustered tweets into distinct topics using unsupervised topic modeling (BerTopic (Grootendorst, 2022)). We then extracted representative words (ranked by tf-idf) for each topic and used GPT-4 to generate topic descriptions (Aizawa, 2003). We set a minimum threshold of 20 documents per cluster for our dataset, resulting in the identification of 14 distinct topics. Using the topic descriptions and representative keywords, we employed GPT-4 to summarize the offensive types and align each topic with its corresponding offense category. The types and their associated topic descriptions are presented in Table 8 in Appendix B.

To validate our GPT-4-assisted thematic grouping, we performed two checks. First, our derived categories align well with established offense types like sexual and violent offenses from related work (Vandenbosch et al., 2015; Davidson et al., 2019). Second, to quantitatively assess the rigor of the GPT-4 annotations, one author annotated a random sample of 100 tweets from the dataset and confirmed an agreement of 84%.

Based on the associated topics under each type, we further summarize the taxonomy of each offensive type, as outlined below:

- **Sexual Content and Gender Issues**: This offensive type includes sexual harassment, gender discrimination, body shaming, and objectification. Gender-based insults and derogation also fall into this category.
- **Personal Attacks and Disrespect**: This includes direct insults, disrespect, or derogation targeting individuals based on personal characteristics.
- **Racial and Ethnic Offense**: This includes racial slurs, ethnic stereotyping, and various forms of discrimination based on race or ethnicity.
- **Political and Social Issues**: This includes political attacks and harassment against individuals or groups over their political views.
- **Violence and Abuse**: This includes topics related to physical or verbal abuse and violence. This can be related to threats, aggressive behaviors, and other forms of violence as forms of offensive content.

We aggregate the emojis within each topic and use the matching relationship between topics and offensive types to assign emojis to each offensive type. In Table 2, we present the top 10 most frequent emojis for each offensive type. We note that for the type of "political and social issues" of offense, only 8 emojis are present.

Offense Type	Top 10 Frequent Emojis
Sexual Content	💦🍆👅😍🍑😈😋😏🔥💋
Personal Attacks and Disrespect	💩😂🤡🙄🤬🤮🐀🤣🖕😡
Racial and Ethnic Offense	😂💯😭🥴🤣💀🙄🗣️🚮😏
Political and Social Issues	🐀💩🤬😡🤡😂🖕🇺🇸
Violence and Abuse	🤬😡😤😂😠🗣️💯💀😭😏

Table 2: Top-10 emojis under different offense types. Note that there are 8 emojis for the offense type: political and social issues, in our dataset.

From Table 2, we observe that different offensive types are associated with distinct sets of frequently used emojis. The emojis used often reflect the offensive nature of the tweet. For tweets classified as "Sexual Content," we find that users frequently employ emojis such as 💦 (droplets), 🍆 (eggplant), and 👅 (tongue) to symbolize body parts. Emojis like 😍 (smiling face with heart-eyes) and 💋 (kiss), which typically convey positive sentiment, are used in these contexts to amplify the offensiveness when combined with sexual content. For the "Personal Attacks," "Racial Offense," and "Political Issues" categories, item-related emojis such as 💩 (pile of poo), 🚮 (trash), and 🐀 (rat) are commonly used to dehumanize the target and intensify the offensive content. Moreover, for the "Violence and Abuse" category, the most frequent emojis, such as 😡 (rage) and 🤬 (cursing face), reflect users' aggressive emotions and sentiments. These findings demonstrate that specific emojis are closely related to the offensive context of the tweet, amplifying the underlying offensive content.

Now that we have explored the prevalent offensive types and their associated emojis, we are also interested in understanding whether these emojis are predominantly used in offensive content or appear more frequently in unoffensive content. In the next section, we will address this question by quantifying the distribution of emojis across unoffensive and offensive tweets.

3.3 Emoji Distribution: Usage in Offensive vs. Non-Offensive Tweets

To quantify this distribution and determine whether the emojis listed in Table 1 are predominantly as-

Frequency in offensive tweets	Emojis
High frequency	🖕💩🍆🍗🤬🚮🤮👅🚿🤡
Moderate frequency	🤢😠🍎😡😏😈😤💀🗣️
Low frequency	🙄😉🤣😂😌💯😋😡💋😒😭😩🤔🔥😘😍

Table 3: Categorization of Emojis Based on Frequency in Offensive vs. Non-Offensive Tweets.

sociated with offensive content, we analyzed usage patterns: For each target emoji, we randomly collected a sample of 1,000 tweets containing it and calculated the proportion of offensive tweets within the sample for that specific emoji.

This percentage reflects how frequently an emoji co-occurred with offensive content within our collected samples. We then categorized the emojis based on this calculated offensive content rate:

- **High Frequency:** Emojis where over 30% of the tweets in their respective sample were offensive.
- **Moderate Frequency:** Emojis where 20% to 30% of the tweets in their respective sample were offensive.
- **Low Frequency:** Emojis where less than 20% of tweets in their respective sample were offensive.

Table 3 presents the emojis categorized according to their association level with offensive content based on this analysis.

Table 3 indicates that most emojis appearing in offensive tweets with a high frequency are item-related, used to dehumanize others or suggest body parts. For moderately frequent emojis, we find that they predominantly convey negative sentiment, such as 😤 and 😡. In contrast, emojis with positive or neutral sentiment are used less frequently in offensive tweets. This suggests that offensive tweets are more likely to be accompanied by negative or dehumanizing emojis.

4 Reduce Offense with Language Models

The analysis so far emphasize that emojis play a significant role in conveying offense, with their function closely tied to the context of the tweet.

Our next question shifts from content creation to content consumption: how can we empower users to curate a less offensive online experience? Rather than relying on platform moderation, we explore an audience-oriented intervention. This approach is motivated by recent work showing that audience desire more personalized control over their content consumption (Jhaver et al., 2023), effectively framing the goal from censoring authors to empowering viewers to manage their online experience.

While audience-side tools are promising, heavy-handed solutions like a full-text rewrite of an offensive tweet can be counterproductive. Such rewriting can drastically reduce content diversity (Padmakumar and He, 2023; Guo et al., 2023), and audience may still wish to understand the gist of the original message. Therefore, we propose a more nuanced, audience-side pipeline that acts as a lightweight filter, selectively replacing the offensive emojis. Functioning much like a browser extension, this targeted approach does not alter an author's original writing. Instead, it operates on the viewer's end to reduce their exposure to offensive content, thereby enhancing their subjective experience.

4.1 Task Setup

Our goal is to design a tool that, from the audience's perspective, replaces potentially offensive emojis with surrogates that mitigate perceived offense while preserving the semantic meaning. We design a multi-step pipeline that leverages LLMs to iteratively select appropriate emoji surrogates. This process is guided by our findings from Section 3, which we incorporate into the prompts to ensure that the replacements align with the original tweet. The input is any tweet a user encounters on their feed. The output of our pipeline is a locally rendered version of that tweet, where potentially offensive emojis have been substituted with alternatives to enhance the viewing experience.

4.2 Pipeline to Identify Emoji Surrogate

Our pipeline consists of four steps. Initially, we classify whether the tweet contains offensive content. If offensive content is detected, we then categorize the roles of each emoji into one of four categories, as described in Section 3.1. Once we identify emojis that either serve as the offense themselves or intensify the offensive tone, we leverage LLMs to suggest appropriate emoji surrogates. We present the visualization in Figure 2. Below, we provide a detailed explanation of each step and the details of prompts are shown in Appendix B.2.

Offensive Content Classification Our first step is to determine whether the tweet contains offensive content. If offensive content is identified, we

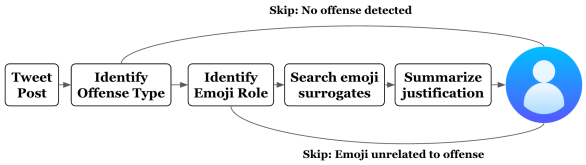


Figure 2: Multi-step pipeline to suggest the emoji surrogates for selected tweets. The blue person icon represents the end-user of our system.

prompt LLMs to classify it into one of the predefined offensive types (sexual, personal attacks, racial offense, political issues, or violence). To guide the model, we include the taxonomy of offensive types within the prompt, along with two demonstrations: one featuring an offensive tweet and the other a non-offensive tweet, to serve as examples for accurate classification. For non-offensive tweets, we leave them as is, while offensive tweets are passed to the next stage.

Determining the Role of Each Emoji We provide the LLM with the identified offense type (from step 1), our four-dimensional emoji role taxonomy and exemplars for each role. Crucially, we incorporate findings from our analysis (Section 3), such as common emojis for each offense type and role, and their general offensive frequency, into the prompt to guide the LLM towards more contextually accurate role identification.

Recommending Emoji Surrogates We ask the LLMs to suggest emoji replacements that remain consistent with the original content and sentiment of the tweet. We also include two demonstrations within the prompt. Emojis classified as "Mitigate offense" or "Unrelated to offense" are intentionally preserved to prevent losing the original semantics.

Summarize the justification Finally, we ask LLMs to summarize the reasoning of emoji replacement. The output presented a summarized justification explaining how the emoji substitutions reduce the offense level.

4.3 Experiment and Setup

For the experiment, we use GPT-4 as the LLM to recommend emoji surrogates and run the multi-step pipeline on our collected 9,285 offensive tweets (Section 3). To demonstrate effectiveness, we compare it with a **direct prompting** baseline, where the LLM is simply asked to replace emojis in offensive tweets to mitigate offense while maintaining tone (prompt details are shown in Appendix B.2).

After running our proposed pipeline, we generated emoji surrogates for a total of 7,142 tweets. For the remaining offensive tweets, no emojis were identified as playing a role in intensifying offense or directly representing offensive content.

4.4 Qualitative Evaluation

Emoji Distribution after Substitution We first examine whether offensive emojis were effectively eliminated by comparing the emoji distributions for each offensive type before and after running our pipeline and the baseline. Our analysis shows the multi-step pipeline successfully eliminates most item-based emojis frequently used for offense (e.g., 💩, 💦, 🍆) and reduces the frequency of negative sentiment emojis (😡, 🤬) and those with sarcastic tones (🙄). In contrast, the direct prompting method retained many problematic emojis (e.g., 💦, 🍑 in sexual content). This suggests our multi-step approach, informed by prior analysis, is more effective at filtering offensive emojis. (The full emoji comparison in Table 9 in Appendix B.1).

Case Studies To illustrate the pipeline's effectiveness, we present several case studies covering different offensive types. These examples showcase how the pipeline identifies the role of emojis in context and recommends appropriate, less offensive surrogates while providing step-by-step justifications. These detailed examples and the LLM's reasoning are provided in Appendix 4.6.

4.5 Human Evaluation

This section details the human-centered study to assess the pipeline's effectiveness for reducing emoji-related offensiveness from a viewer's perspective.

Evaluation Design The goal of our pipeline is to reduce offensiveness while preserving semantics. A straightforward evaluation approach is to present the original and processed tweets side by side and ask audience to assess whether the pipeline effectively reduces offensive content and whether semantic meaning is preserved. However, this method may introduce cognitive bias, as audience might be inclined to perceive the two versions as inherently similar (Haselton et al., 2015).

To mitigate this bias, we conduct a within-subject user experiment in which annotators evaluate each tweet independently by answering a series of questions related to semantics and offensiveness. We recruited native English-speaking annotators with a >98% approval rate from Prolific (details in

Category	Measured Variables (*Scale / Type*)
Offensiveness	Offensiveness Score *(1-5)*, Sarcasm *(% Yes)*, Body Symbol Emojis *(% Yes)*, Dehumanizing Emojis *(% Yes)*
Semantics	Sentiment *(1-3)*, Arousal *(1-3)*, Extra Meaning *(% Yes)*, Clarity *(% Yes)*, Fluency *(% Yes)*

Table 4: Measured variables for tweet annotation. Detailed variable meanings are shown in Appendix A.3

Measured Variables	Original	Direct	Multi-step
Offensiveness Score (1-5)	3.00	2.94	2.58*
Sarcastic (% Yes)	34.0%	38.0%	37.5%
Body Symbol (% Yes)	49.5%	51.0%	52.0%
Dehumanization (% Yes)	20.0%	12.5%	16.5%
Sentiment Score (1-3)	1.48	1.59	1.64
Arousal Score (1-3)	2.14	2.06	2.01
Extra Meaning (% Yes)	25.0%	24.5%	19.0%
Clarity (% Yes)	74.0%	70.5%	72.5%
Fluency (% Yes)	77.0%	76.0%	77.0%

Table 5: Human evaluation results comparing our proposed multi-step pipeline with the direct prompting method. Statistical significance is indicated as follows: *: $p < 0.05$ (paired t-test).

Appendix A.1). Each annotator assesses 60 tweets presented in a randomized order, consisting of 20 original tweets, 20 versions of these tweets processed by our pipeline, and 20 versions processed by the baseline method. We then compare the differences in responses to the tweets before and after processing to assess the pipeline's impact. This methodology allows us to measure the change in perception for each annotator individually. This focus on the delta is powerful because it controls for the inherent subjectivity of perceiving offensiveness, providing a clearer signal of our intervention's effect. Consequently, our primary analysis relies on the statistical significance of this within-subject change, making traditional inter-annotator agreement scores a less critical measure for evaluation robustness. Ideally, annotators should perceive the rewritten tweets as less offensive.

Measures We collected measures that aim to assess the offensiveness and semantic aspect of each tweet. For offensiveness, we firstly measured the perceived level of offensiveness by asking annotators to rate how offensive they find each tweet on a scale from 1 to 5, ranging from 'not offensive' to 'extremely offensive.' Additionally, based on the findings in Sections 3.1 and 3.2, we observe that emojis can enhance offensiveness by expressing irony, symbolizing body parts, or dehumanizing the target. We ask annotators whether the emojis in each tweet exhibit these functionalities.

We also consider the influence of emoji replacement on the semantic aspect, given emojis' role, such as conveying sentiment. For each tweet, we ask annotators to assess the sentiment, emotion arousal, whether the emojis contribute external meaning, clarity, and fluency. These annotations allow us to quantify the semantic integrity of each tweet. We present the collected variables in Table 4 and the questionnaire in Appendix A.3.

Average Evaluation Result For our evaluation, we randomly sampled 600 tweets for evaluation. This sample size was determined to be sufficient for our analytical goals, as it yielded a diverse set of examples covering all identified offense types. We then present annotators with 600 tweets: 200 original tweets, 200 tweets from our multi-step pipeline, and 200 tweets via direct prompting. Each tweet is annotated by two annotators with the predefined questions. After annotation, we compute the average scores for overall offensiveness, sentiment, and arousal, as well as the percentage of 'Yes' responses for other variables. The results for the original, pipeline-processed, and direct-prompting-processed tweets are shown in Table 5.

As shown in Table 5, our proposed multi-step pipeline significantly reduces the offensiveness scores assigned by annotators. In terms of semantic preservation, tweets processed by our pipeline exhibit no notable changes in meaning. Compared to our pipeline, the direct prompting baseline achieves only a minor and statistically insignificant reduction in offensiveness. We suspect this is because, without prior knowledge of the relationship between offensiveness and emojis, LLMs struggle to identify suitable emoji surrogates.

4.6 Qualitative Evaluation: Case Studies

We present four random examples of the original tweet and the revised tweet after processing through our pipeline, covering different offensive types in Figure 3. In addition, we include the justifications summarized by the LLMs in the final step of our pipeline for each emoji substitution.

The examples and justifications presented in Figure 3 demonstrate that our pipeline effectively identifies offensive content, provides the reasoning behind its offensiveness, and captures the role of emojis within the tweet. For instance, in the first

Measurement Variables (Δ)	Personal Attacks		Political/Social		Racial/Ethnic		Sexual/Gender		Violence/Abuse	
	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step	Direct	Multi-step
Offensiveness (1-5)	−0.18	−0.05	+0.15	−0.23	−0.09	−0.94*	−0.38*	−0.38*	+0.12	−0.60*
Sarcastic (% Yes)	−2.6%	−2.6%	+12.5%*	+2.5%	+2.9%	+5.9%	+5.0%	+7.5%	0.0%	+4.3%
Body Symbol (% Yes)	+2.6%	+2.6%	+2.5%	+2.5%	+2.9%	+4.7%	0.0%	−15.0%*	+4.8%	0.0%
Dehumanization (% Yes)	−5.1%	−12.8%*	−7.5%	−22.5%*	−2.9%	0.0%	0.0%	+2.5%	0.0%	−2.4%
Sentiment (1-3)	+0.03	+0.00	+0.10	+0.25	+0.03	+0.24	+0.23	+0.13	+0.12	+0.17
Arousal (1-3)	−0.31	−0.13	−0.20	−0.38	+0.24	−0.15	−0.20	+0.00	+0.10	−0.02
Extra Meaning (% Yes)	−2.6%	−20.5%*	−2.5%	0.0%	0.0%	−2.9%	+2.5%	−7.5%	−2.4%	0.0%
Clarity (% Yes)	−2.6%	−2.6%	−10.0%	0.0%	−5.9%	0.0%	−2.5%	−2.5%	+4.8%	−2.4%
Fluency (% Yes)	+5.1%	+7.7%	−15.0%*	−5.0%	−8.8%	−8.8%	+5.0%	+2.5%	+9.5%	+2.4%

Table 6: Mean differences in human evaluation metrics after processing tweets using Direct Prompting or Multi-step Pipeline (Δ = Processed Score - Original Score), across different offense types. *: $p < 0.05$ (paired t-test).

example, the LLM accurately interprets 🍑 as a reference to a body part, which intensifies the offense. It suggests replacing it with the flower emoji 🌸 to keep the positive sentiment while reducing the offensive nature of the tweet. Moreover, in the second example, where the post includes the word "tacos" and the 😡 emoji, our pipeline detects the implicit racial offense toward Hispanic or Latino culture. It recommends replacing 😡 with 😕 to reduce the offense while maintaining the semantics of the post. In conclusion, our pipeline effectively identifies offensive content within tweets, uncovers the relationship between emojis and offensive material, and precisely recommends emoji surrogates to mitigate the offense while preserving the tweet's overall semantics.

While our case study demonstrates the effectiveness of our multi-step pipeline, the next step is to quantitatively assess its impact on offensiveness reduction for each tweet. In the following section, we leverage human annotations to evaluate the offensiveness of tweets before and after LLM rewriting.

4.7 Heterogeneous Effects by Offensive Types

While our pipeline reduces overall offensiveness, users may post offensive tweets of varying types. This raises the question of whether our pipeline's effectiveness remains consistent across offense types. We re-calculate the variables within each category and the results are presented in Table 6.

Table 6 shows our multi-step method effectively removed contextually problematic emojis. It reduced dehumanizing symbols by 12.8% in personal attacks and body-part symbols by 15.0% in sexual offenses. Corresponding semantic shifts were observed: e.g., emojis conveying 'extra meaning' decreased by 20.5% alongside dehumanizing ones in personal attacks. These aligned changes suggest the pipeline correctly targeted emojis based on their function. Sarcasm levels were generally unaffected, although the direct baseline notably increased perceived sarcasm (+12.5%) while decreasing fluency (-15.0%) for political tweets, likely reflecting contextually poor emoji choices.

Crucially, removing problematic emojis did not consistently lower overall offensiveness scores, especially for highly offensive content. Despite reducing dehumanizing emojis in personal (12.8%) and political (22.5%) attacks, the change in offensiveness scores for these categories was statistically insignificant. Examining individual cases revealed that offensiveness reduction primarily occurred in mildly offensive tweets (original score < 3). A similar pattern held for sexual offenses: removing body symbols reduced offensiveness in some cases but had no impact on tweets already rated maximally offensive (score=5). This indicates that while emoji moderation effectively removes specific offensive elements, its impact on overall perceived offensiveness is limited when strong verbal attacks dominate. Emoji replacement appears most effective for mitigating milder forms of offense.

We present a case study in Table 7 in Appendix A. In the first example, replacing 💩 with (🤔, 😒) does not reduce offensiveness, as strong verbal attacks remain. In contrast, the second example shows that removing 💩 in a mildly offensive tweet lowers its perceived offensiveness. This suggests that emoji moderation is more effective in less severe cases, while text plays a dominant role in highly offensive tweets. Our analysis defines the operational boundaries of our pipeline, showing its effectiveness on mild but not severe offenses where text dominates.

5 Implications

This work examines emojis' often-overlooked role in online offensiveness. Our proposed targeted

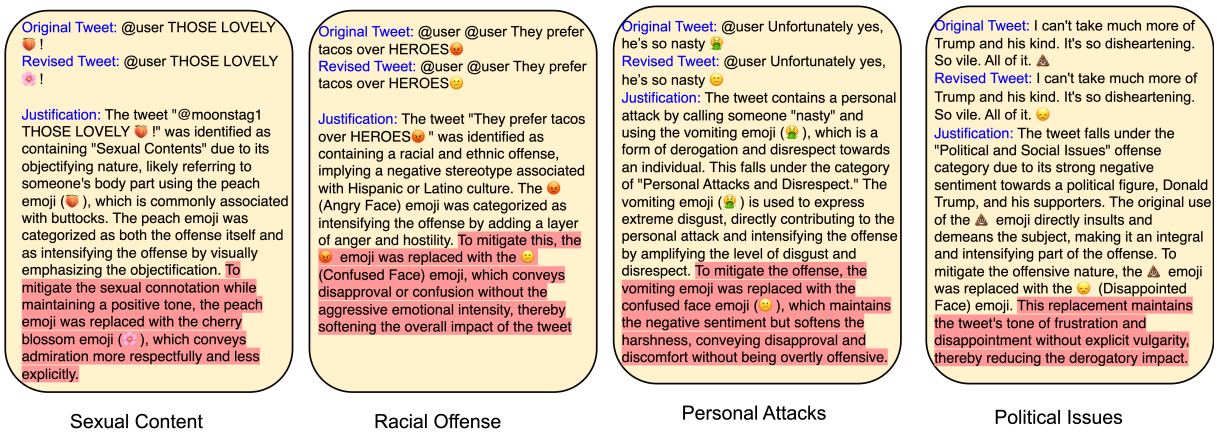


Figure 3: Justification from our multi-step pipeline of emoji replacement. The red color highlights the reason of choosing that emoji surrogate. The offense type of each tweet is labeled below.

moderation pipeline offers practical tools and conceptual insights for improving online discourse. For social media users and platforms, recognizing how emojis intensify, mitigate, or reframe offense can enhance communication clarity and moderation transparency. It underscores that offense can be conveyed beyond explicit text.

For content moderation researchers, our findings emphasize the need to analyze non-textual cues. Emojis implicitly carry offense through symbolism, sarcasm, or stereotypes. We demonstrate that LLMs can selectively substitute offensive emojis while preserving semantic content.

For developers of moderation tools, this highlights the importance of incorporating non-verbal signals like emojis, moving beyond current text-centric models. It calls for broader multimodal understanding and opens opportunities for fine-tuning or prompting techniques that address text-emoji interactions in offensive communication.

6 Conclusion

In this paper, we investigate the role of emojis in offensive social media content and propose a multi-step LLM pipeline to mitigate offensiveness while preserving tweet semantics. Our analysis reveals that emojis can amplify, mitigate, or subtly alter offensive content, emphasizing the need for moderation beyond textual cues. Through human evaluation, we demonstrate that our approach effectively reduces offensiveness compared to direct prompting. Our findings highlight the importance of integrating emoji semantics into content moderation and encourage future work to explore adaptive, user-aware moderation strategies.

7 Limitations

Despite the promising results, our approach has several key limitations. First, emoji interpretation varies across individuals and cultural backgrounds (Lu et al., 2016; Zhou et al., 2024b,a). In this study, our focus on English-language tweets from U.S.-based users was a deliberate methodological choice for this foundational study. This scope allowed us to first establish that a systematic relationship exists between emoji use and offensive content and to test our pipeline's feasibility in a large, data-rich environment while minimizing cross-cultural confounding variables. Consequently, while the specific semantic mappings we found are English-centric, we argue that our core methodology, the framework for analyzing emoji roles and performing targeted replacement, is generalizable. We therefore position our work as a proof-of-concept that offers an adaptable blueprint for future work, where applying this framework across different languages and cultures remains a crucial next step.

Second, our dataset's scope is intentionally focused on a specific region (U.S.) and time period (2019). This was a deliberate choice to establish a foundational proof-of-concept using a large-scale, stable, pre-COVID baseline dataset. However, we acknowledge this limits the direct generalizability of our findings regarding specific emoji trends, as online communication evolves. While the fundamental behavioral patterns we identify (e.g., using emojis for sarcasm or dehumanization) are likely durable, the specific emojis used to express these patterns may change over time and across regions. Future work should validate these patterns on more

contemporary and geographically diverse datasets.

Third, our study was designed to first answer a critical prerequisite question: can our intervention effectively reduce perceived offensiveness in a controlled environment? Our work provides this foundational validation, demonstrating that the pipeline is successful at its primary task. We acknowledge that this does not measure real-world user acceptance or behavioral responses, which is a vital next step. However, this constitutes a different type of research question that requires a distinct experimental setup (e.g., a custom user interface and a longitudinal study), and we frame our current work as an essential precursor to such future user-centered studies.

Fourth, the effectiveness of our pipeline is inherently tied to the capabilities and potential biases of the LLMs it employs (e.g., GPT-4, RoBERTa). These models, despite their advancements, can reflect biases present in their training data, potentially leading to skewed interpretations of emoji offensiveness or unfair targeting of certain emoji uses or user expressions. Furthermore, the generalization of the pipeline to novel, rapidly evolving emoji slang or newly introduced emojis is a continuous challenge. LLMs may not immediately grasp the nuanced offensive uses of emojis that emerge after their last training update, requiring ongoing monitoring and model fine-tuning. Therefore, while we justify our use of LLMs for their scalable analytical power, we emphasize that our human evaluation (Section 4.5) serves as the final arbiter of our method's success, providing a crucial check against these potential model-centric biases.

8 Ethical Consideration

Ethical considerations for the annotation process were carefully observed. The privacy of annotators was protected as no personally identifiable information was collected. The task involved evaluating tweet content and did not entail extensive or intrusive tool usage. In line with our Institutional Review Board's (IRB) protocols for research not involving the collection of identifiable private information about human subjects, this portion of the study was deemed exempt from formal IRB review. We also note that AI assistants were employed to support coding tasks during the implementation of our experiments.

References

Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *ICWSM 2017*.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. 2020. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6193–6202.

Zhenpeng Chen, Xuan Lu, Wei Ai, Huoran Li, Qiaozhu Mei, and Xuanzhe Liu. 2018. Through a gender lens. *WWW 2018*.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered representation learning for Cross-Lingual sentiment classification. In *WWW 2019*.

Henriette Cramer, Paloma de Juan, and Joel Tetreault. 2016. Sender-intended functions of emojis in us messaging. In *MobileHCI 2016*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022a. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Yong Deng, Chenxiao Dou, Liangyu Chen, Deqiang Miao, Xianghui Sun, Baochang Ma, and Xiangang Li. 2022b. Beike nlp at semeval-2022 task 4: prompt-based paragraph classification for patronizing and condescending language detection. *arXiv preprint arXiv:2208.01312*.

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. *EMNLP*.

Jing Ge. 2019. Emoji sequence use in enacting personal identity. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 426–438, New York, NY, USA. Association for Computing Machinery.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Martie G Haselton, Daniel Nettle, and Paul W Andrews. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 724–746.

Tianran Hu, Han Guo, Hao Sun, Thuy-vy Nguyen, and Jiebo Luo. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM 2017*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A Hale. 2021. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *arXiv preprint arXiv:2108.05921*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

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*.

Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. In *UbiComp 2016*.

Hanjia Lyu, Weihong Qi, Zhongyu Wei, and Jiebo Luo. 2024. Human vs. lmms: Exploring the discrepancy in emoji interpretation and usage in digital communication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2104–2110.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "blissfully happy" or "ready tofight": Varying interpretations of emoji. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 259–268.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457.

Huy Nghiem and Hal Daumé III. 2024. Hatecot: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. *arXiv preprint arXiv:2403.11456*.

OpenAI. 2023. Gpt-4 technical report.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Laura Vandenbosch, Johanna MF van Oosten, and Jochen Peter. 2015. The relationship between sexual content on mass media and social media: A longitudinal study. *Cyberpsychology, Behavior, and Social Networking*, 18(12):697–703.

Benjamin Weissman and Darren Tanner. 2018. A strong wink between verbal and emoji-based irony: How the brain processes ironic emojis during language comprehension. *PloS one*, 13(8):e0201727.

Michael Wiegand and Josef Ruppenhofer. 2021. Exploiting emojis for abusive language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 369–380.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations. *arXiv preprint arXiv:2406.12223*.

Case	Original Tweet	Revised Tweet
No Change in Offensiveness (Score: 4)	You all are dog 💩 💩Trump is a disgrace, a traitor and must be removed from office.	You all are dog 🤔😒 Trump is a disgrace,a traitor and must be removed from office.
Offensiveness Reduced (Score: 2 → 1)	#TRUMP2020 Landside stable genius 💩	#TRUMP2020 Landside stable genius 😒

Table 7: Case study of tweets before and after dehumanizing emoji removal.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Ruiqi Zhong, Yanda Chen, Desmond Patton, Charlotte Selous, and Kathy McKeown. 2019. Detecting and reducing bias in a high stakes domain. *arXiv preprint arXiv:1908.11474*.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

Yuhang Zhou and Wei Ai. 2022. #emoji: A study on the association between emojis and hashtags on twitter. In *ICWSM 2022*.

Yuhang Zhou, Giannis Karamanolakis, Victor Soto, Anna Rumshisky, Mayank Kulkarni, Furong Huang, Wei Ai, and Jianhua Lu. 2025. Mergeme: Model merging techniques for homogeneous and heterogeneous moes. *arXiv preprint arXiv:2502.00997*.

Yuhang Zhou, Xuan Lu, and Wei Ai. 2024a. From adoption to adaption: Tracing the diffusion of new emojis on twitter. *arXiv preprint arXiv:2402.14187*.

Yuhang Zhou, Xuan Lu, Ge Gao, Qiaozhu Mei, and Wei Ai. 2023b. Emoji promotes developer participation and issue resolution on github. *arXiv preprint arXiv:2308.16360*.

Yuhang Zhou, Paiheng Xu, Xiyao Wang, Xuan Lu, Ge Gao, and Wei Ai. 2024b. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. *arXiv preprint arXiv:2402.01681*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Appendix

A Supplementary Evaluation Materials

A.1 Detailed Recruitment Process

Participant recruitment was conducted via the Prolific online platform, yielding a cohort of 20 annotators. To ensure high-quality data, we used strict filtering criteria: participants were required to be adults, identify as native English speakers, and have a high platform approval rate (>98%) with extensive prior task experience. All annotators were provided with detailed instructions and examples (see Appendix A.3) to calibrate their understanding of the offensiveness scale, and we conducted a pilot study to confirm these instructions were clear regarding emoji use and interpretation. The annotation protocol involved each individual assessing 60 tweets, a task projected to require approximately 90 minutes of engagement, with compensation provided at a rate of $18 USD per hour.

A.2 Detailed Prompts for Data Annotation

For all data annotation tasks described in Section 3, we used GPT-4 with a temperature setting of 0. This was done to ensure deterministic and reproducible outputs. The following zero-shot prompts rely on clear definitions to guide the model.

Prompt for Offensive Content Classification

This prompt was used for the final, high-fidelity classification of tweets after the initial RoBERTa filtering.

Instruction:

Analyze the following tweet and determine if it contains offensive content based on the definition provided below.

Definition of Offensive Content:

Offensive content is defined as posts containing unacceptable language (profanity) or targeted offenses, whether direct or veiled, including insults, threats, profane language, or swear words.

Output Format:

Your response must strictly follow this format:

Offensive: [Yes/No]

Justification: [A brief reason for your classification based on the definition.]

Tweet to Classify: {tweet}

Prompt for Emoji Role Annotation

This prompt was used to annotate the function of each emoji within tweets that were already identified as offensive.

Instruction: You will be given an offensive tweet containing one or more emojis. For EACH emoji in the tweet, determine its function based on the definitions of the four categories provided below.

Functionality Category Definitions:
Offensive in itself: The emoji alone constitutes an offense.
Intensify offense: The emoji enhances the intensity of the offensive tweet, for example by expressing irony or sarcasm.
Mitigate offense: The emoji softens or adjusts the tone of the tweet, reducing its offensive impact.
Unrelated to offense: The emoji is not directly connected to the offensive content of the tweet.

Output Format: For each emoji, provide your response in the following format. Repeat for all emojis present in the tweet.
Emoji: [The emoji character]
Functionality: [The name of the category from the definitions above]
Justification: [A brief reason for the assigned functionality based on the definition and the tweet's context.]

A.3 Detailed Questionnaire of Human Evaluation

Annotator Background Questions

1. **What is/are your native language(s)?**
   (Please specify in the text box below.)
   [Fill in the blank]

2. **In the past month, how often do you use social media platforms (e.g., Twitter, Reddit, Instagram, TikTok, Messenger)? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

3. **In the past month, how often do you see content that includes emojis on social media platforms? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

4. **In the past month, how often do you use emojis yourself when engaging on social media platforms? Please select the option that best describes the frequency.**
   - Never: Never in the past month
   - Rarely: Fewer than once a week
   - Sometimes: Two or three times a week
   - Often: More than three times a week, but not everyday
   - Always: Almost everyday

Guidelines for Assessing Offensiveness

Offensiveness in the context of tweets refers to content that conveys disrespect, hostility, or harm toward an individual or group. This may occur through negative intent, harmful stereotypes, or inappropriate language. Offensive content is typically characterized by its potential to cause emotional distress or harm to the dignity of those targeted.

When assessing the offensiveness of a tweet, consider the following factors:

Intent Evaluation

- Does the tweet imply a negative or derogatory attitude toward any individual or group?
- Assess whether the speaker's intent appears harmful, disrespectful, or mocking.

Outcome Evaluation

- Could the tweet cause emotional harm or be perceived as offensive by the targeted individual or group?
- Reflect on how the individuals or groups mentioned in the tweet might feel upon reading it.

Content Evaluation

- Does the tweet include harmful stereotypes, slurs, or degrading language?
- Look for both explicit and implicit elements that might convey disrespect.

Acceptability Assessment

Based on the content and context, determine whether the tweet is socially acceptable:

- **Acceptable**: It is reasonable for the speaker to make this statement.
- **Unacceptable**: The speaker should not have made this statement.

Overall Offensiveness Rating

Rate the tweet's offensiveness on a scale of 1 to 5:

- 1 (Not Offensive At All): No evidence of harmful intent, outcome, or content.
- 2 (Slightly Offensive): Minimal or ambiguous evidence of harmful content or intent.
- 3 (Moderately Offensive): Clearly problematic in content or intent but without strong emotional harm.
- 4 (Highly Offensive): Contains explicit disrespect, harm, or stereotypes; likely to cause emotional harm.
- 5 (Extremely Offensive): Extreme disrespect or hostility; causes significant harm or distress.

Examples of Assessing Offensiveness

To provide more clarity, here are a few examples to illustrate how to evaluate the offensiveness of tweets:

Example 1

Tweet: "People from [Group X] are lazy and don't want to work."

- **Intent**: Likely derogatory due to stereotyping.
- **Outcome**: Likely to cause emotional harm to Group X.
- **Content**: Contains harmful stereotypes.
- **Acceptability**: Unacceptable.
- **Overall Offensiveness Rating**: 4 (Highly Offensive)

Example 2

Tweet: "I had a terrible experience at [Restaurant Y]."

- **Intent**: Negative but not targeting an individual or group.
- **Outcome**: Unlikely to cause emotional harm.
- **Content**: No harmful language.
- **Acceptability**: Acceptable.
- **Overall Offensiveness Rating**: 1 (Not Offensive At All)

Example 3

Tweet: "This looks like something only a [Group Z] person would wear."

- **Intent**: Ambiguous; may or may not be derogatory.
- **Outcome**: Could cause mild emotional harm to Group Z.
- **Content**: Implicitly reinforces stereotypes about Group Z.
- **Acceptability**: Borderline unacceptable.
- **Overall Offensiveness Rating**: 3 (Moderately Offensive)

Tweet Evaluation Form

Instructions: Please read the tweet carefully and answer the following questions based on your interpretation. Your responses will help us understand perceptions of the tweet's content.

Tweet: [Display the tweet here]

Offensiveness Assessment

a. **Overall Offensiveness**
On a scale from 1 to 5, where 1 means "Not offensive at all" and 5 means "Extremely offensive," how offensive do you find this tweet?

- 1 - Not offensive at all
- 2 - Slightly offensive
- 3 - Moderately offensive
- 4 - Very offensive
- 5 - Extremely offensive

b. **Offensive Elements (Optional)**
If you found the tweet offensive, please briefly mention which elements contributed to this perception (e.g., language, emojis, insinuations).

c. **Is the tweet attempting to be humorous or sarcastic?**

- Yes
- No

d. **Does the tweet contain any emoji to symbolize human body parts?**

- Yes
- No

e. **Does the tweet contain any emoji to dehumanize a target?**

- Yes
- No

Sentiment Analysis

a. **Overall Sentiment**
How would you rate the overall sentiment of the tweet?

- 1 - Negative
- 2 - Neutral
- 3 - Positive

b. **Emotional Intensity**
How emotionally intense or stimulating is the tweet? Consider the energy, excitement, or agitation it conveys, regardless of whether the sentiment is positive or negative.

- 1 - Low arousal
- 2 - Moderate arousal
- 3 - High arousal

Extra Emoji Meaning

a. **Do you agree with this statement: "The meanings of all emojis in the tweet are disclosed by the text of the tweet."**

- 1 - Disagree
- 2 - Agree

Tweet Attributes

a. **Clarity**
Is the tweet clearly written? Does this tweet provide enough information for the reader to understand its meaning?

- Yes
- No

b. **Fluency**
Does the tweet sound fluent or natural? Consider whether the tweet is easy to read and flows smoothly.

- Yes
- No

B Supplementary Results

B.1 Qualitative Evaluation: Emoji Distribution Comparison

We present the emoji distribution for each offensive type before and after running our pipeline in Table 9.

B.2 LLM Pipeline Prompting Details

This appendix provides the detailed prompts used in each step of the LLM pipeline described in Section 4.2.

Step 1: Offensive Content Classification Prompt
You are tasked with analyzing a tweet for offensive content. Determine if the tweet contains any offensive language or sentiments. Offensive content includes any form of non-acceptable language (profanity) or a targeted offense (veiled or direct), such as insults, threats, profane language, or swear words.

If offensive content is detected, identify the type of offense based on the following categories:

1. **Sexual Content and Gender Issues**: Includes sexual harassment, gender discrimination, body shaming, objectification, gender-based insults, and derogation.

2. **Personal Attacks and Disrespect**: Ranges from direct insults to subtle disrespect/derogation targeting individuals/groups based on personal characteristics.

3. **Racial and Ethnic Offense**: Includes racial slurs, ethnic stereotyping, and discrimination/prejudice based on race or ethnicity.

Offensive Type	Topics
Personal Attacks and Disrespect	0_Personal Confrontations and Profanity, 1_Explicit Content Solicitation, 3_Offensive Language Usage, 6_Casual Slang and Swearing, 9_Offensive Language and Slang Usage, 10_Casual Profanity Usage
Sexual Content and Gender Issues	4_Demeaning Language Toward Women
Racial and Ethnic Offense	2_Racial Slur Usage in Conversation, 11_Racial Discrimination and Stereotyping
Political and Social Issues	5_Criticism of Trump's Statements, 13_Offensive Political and Religious Comments
Violence and Abuse	7_Child Abuse Concerns, 8_Sleep and Fatigue, 12_Student Violence

Table 8: Categories of offensive types and the belonged topics. Numbers before each topic represents the topic number, ranked by the number of tweets in this topic.

Offense Type	Pipeline	Emoji Distribution
Sexual Content	Original	🍆👅😍🍑😈😊😏🔥💋
	Direct prompting	😍💦😊😋🔥🍑😂🍆😈💋
	Multi-step	😊😅🚤⭐😜😋🤔😐😍🌸
Personal Attacks	Original	💩😂🤡😒🤬🤮🖕🤣🖕😡
	Direct prompting	😂😏😡😅😤🤡🤔😠🙄🤣
	Multi-step	😅😐😒🤔😐😤😡😡👎😊
Racial Offense	Original	😂💯😠🥴🤣💀🧐📊😏
	Direct prompting	😂💯😭😅😏😐😡🤔🤣🙄
	Multi-step	😅😐🤔👍😊😒😐😡😏😆
Political Issues	Original	🖕💩🤬😡🤡😂🖕🇺🇸
	Direct prompting	🐭🤡🤔😡😏🤥👎🇺🇸💩😏
	Multi-step	🤔😐😒😡🤡😡🇺🇸😅🙄😐
Violence and Abuse	Original	🤬😡😤😂😠🗣️💯💀😏
	Direct prompting	😡😤😒😅🗣️😭🔥💯😂😏
	Multi-step	😐😤😡😐😒😅🚫😒👍😏

Table 9: Top 10 emojis by offense type, comparing the emojis in the "Original" tweets against the outputs of the "Direct prompting" baseline and our "Multi-step" pipeline after substitution.

4. **Political and Social Issues**: Encompasses political attacks, social discrimination, harassment, and aggression over political views or social status.

5. **Violence and Abuse**: Includes topics related to physical or verbal abuse, violence, threats, and aggressive behaviors.

Your Response Should Include:

- Whether the tweet is offensive (*Yes/No*).
- If offensive, the offense category (*e.g., Sexual Content and Gender Issues*).
- A brief justification for the chosen category.

Examples:

- **Tweet:** `@sinnersworldxxx` 🚀😒💋💗 `RT to f*p with a surprise in DM #sex #horny https://t.co/VtTjihrWSP`
- **Offensive:** Yes
- **Offense Category:** Sexual Content and Gender Issues
- **Justification:** This tweet promotes sexually suggestive behavior (*e.g., "RT to f*p"*) and references adult content (*e.g., #sex, #horny*), violating standards around explicit material.
- **Tweet:** `My throat hurts. Can god give me a break` 🥰.
- **Offensive:** No
- **Justification:** This tweet expresses personal discomfort casually and contains no inappropriate or offensive language.

Tweet to classify:

- **Tweet:** `{tweet}`

Step 2: Emoji Role Determination Prompt You will be given an offensive tweet identified as type: `{offense_type}`. Analyze each emoji (e.g., 😂) within this tweet to determine its functionality. An emoji may fit multiple categories.

The functionality categories are:

1. Emoji represents the **offense itself**
2. Emoji **intensifies** the offense
3. Emoji **mitigates** the offense

4. Emoji is **not directly related** to the offense

Use the following data (emojis based on provided list) to inform your analysis:

Emojis Often Associated with Specific Offense Types:

- *Sexual Content:* 😂, 🤣, 😭, 😍, 😉, 😋, 😌, 😩, 😈, 🙄, 😒, 💦, 🥺, 👀, 👅, 🔥, 🤔, 😒, 💀, ❤️, 💋, 🥵
- *Personal Attacks:* 😂, 😭, 🤣, 🙄, 🤡, 😉, 💀, 💩, 🤮, 😒, 🤔, 😡, 🤬, 😳, 😔, 🤢, 😤, 😐, 😠, 😑, 👀, 🥺, 😅, 😩, 😉, 🗣️
- *Racial/Ethnic:* 😂, 😭, 🤣, 😉, 😒, 💀, 🙄, 💯, 🗣️, 🤔, 🤢
- *Political/Social:* 😂, 😡, 🙄, 😉, 🤢, 😮, 😌, 😭
- *Violence/Abuse:* 😂, 😭, 🤣, 💀, 😡, 🙄, 🤢, 💯, 🤬, 😉, 😒, 😍, 🔥

Popular Emojis by Typical Functionality:

- *Offense Itself (1):* 🤡, 💩, 🖕, 😂, 🖕, 😡, 🤣, 🖕, 😈, 🙄, 🤮, 💫, 🖕, 🐀, 😭, 😩, 😔, 🥵, 💯
- *Intensify Offense (2):* 😂, 😭, 🤣, 🙄, 💀, 😉, 😩, 😒, 💯, 😍, 😋, 🔥, 😤, 🤢, 😡, 🤡, 😈, 🤮, 😌, 🤬, 🤔, 😏, 💦, 🗣️, 😛
- *Mitigate Offense (3):* 😂, 😭, 🥺, 😔, 🤣, 😩, 😉, ❤️, 😢, 😳, 🙃, 🥰, 💀, 💕, 😊, 😐, 😐, 😴, 😍, 😌, 😘, 🤗, 😜
- *Unrelated (4):* 😂, 😭, 😍, 🤣, 🤔, 🔥, 😩, 😉, ❤️, 👀, 🥺, 💀, 🥰, 😎, 😳, 🗣️, 💙, 😌, 💯, 🖤

Emoji Frequency in Offensive Tweets:

- *High Freq.:* 🖕, 💩, 🍆, 🐀, 🥵, 🚮
- *Medium Freq.:* 🤮, 👅, 💦, 🤡, 🤢, 😡, 🍑, 😡, 😒, 😈, 😤, 💀, 🗣️
- *Low Freq.:* 🙄, 😉, 🤣, 😂, 😌, 💯, 😋, 😛, 💋, 😏, 😭, 😩, 🤔, 🔥, 😘, 😍

Your Task: For each emoji in the provided tweet:

- State the emoji (e.g., 😭).
- Assign functionality category/categories (1-4).
- Provide a brief justification based on context and the data provided.

Examples:

- **Tweet:** `Y'all seriously why do her boobs look like analog sticks`😭🤣
- **Emojis:** 😭, 🤣
- **Analysis for 😭:**
  - **Functionality:** 2 (Intensifies), 4 (Unrelated)
  - **Justification:** Amplifies mockery, making the comment seem excessively funny, reinforcing disrespect. Also unrelated to the core offensive content itself.
- **Analysis for 🤣:**
  - **Functionality:** 2 (Intensifies), 4 (Unrelated)
  - **Justification:** Emphasizes the offensive joke as humorous, diminishing seriousness. Also unrelated to the core offensive content.
- **Tweet:** `just remembered some of the stupid sh*t I did last night, how do I get myself in these situations` 😂😂
- **Emojis:** 😂, 😂
- **Analysis for 😂 (first):**
  - **Functionality:** 3 (Mitigates), 4 (Unrelated)
  - **Justification:** Downplays potential offense with humor/self-deprecation, framing it as non-serious.
- **Analysis for 😂 (second):**
  - **Functionality:** 3 (Mitigates), 4 (Unrelated)
  - **Justification:** Repetition further emphasizes humor, reducing seriousness.
- **Tweet:** `F**k de calor`🥵
- **Emojis:** 🥵
- **Analysis for 🥵:**
  - **Functionality:** 2 (Intensifies)
  - **Justification:** Intensifies the expressed frustration/discomfort, aligning with the profanity to amplify sentiment intensity.

- **Tweet:** `Like and RT for more clip💦#cum [...] #nuses... https://t.co/gSNIDclo6s`
- **Emojis:** 💦
- **Analysis for 💦:**
  - **Functionality:** 1 (Offense Itself), 2 (Intensifies)
  - **Justification:** Often implies sexual acts/arousal, directly contributing to and intensifying the explicit content.

Tweet to analyze:

- **Tweet:** `{tweet}`
- **Emojis in Tweet:** `{emoji}`

Step 3: Emoji Surrogate Recommendation Prompt You are given an offensive tweet (type: `{offense_type}`) where emojis have been classified by functionality:

1. Emoji represents the **offense itself**
2. Emoji **intensifies** the offense
3. Emoji **mitigates** the offense
4. Emoji is **not directly related** to the offense

Your task is to recommend replacements only for emojis categorized as **1 (Offense Itself)** or **2 (Intensifies Offense)**. Emojis categorized as **3 (Mitigates)** or **4 (Unrelated)** should be kept.
For emojis needing replacement:

- Suggest a replacement emoji (e.g., 🤔) that maintains the tweet's general tone but mitigates the specific offense.
- Choose replacements consistent with the tweet's original content context.

Your Response Should Include (for each replaced emoji):

- Emojis to Replace: Emoji
- Replacement Emojis: Emoji
- Justification: Brief explanation for the replacement choice.
- (Finally) Revised Tweet: The full tweet after all necessary replacements.

Examples:

- **Tweet:** `Y'all seriously why do her boobs look like analog sticks`😭🤣
- **Emojis/Functionality:** 😭 (2, 4), 🤣 (2, 4)
- *Replacement 1:*
  - Emoji to Replace: 😭
  - Replacement Emoji: 🤔
  - Justification: Maintains playful tone but shifts from ridicule to curiosity/confusion.
- *Replacement 2:*
  - Emoji to Replace: 🤣
  - Replacement Emoji: 😊
  - Justification: Keeps tweet lighthearted without amplifying offense; signals amusement softly.
- **Revised Tweet:** `Y'all seriously why do her boobs look like analog sticks`🤔😊
- **Tweet:** `Omg yes!` ✊😩 `sex havers roll call` 😈
- **Emojis/Functionality:** ✊ (4), 😩 (2), 😈 (1, 2)
- *Replacement 1:*
  - Emojis to Replace: 😩
  - Replacement Emojis: 😅
  - Justification: Retains excitement but softens intensity; conveys playful enthusiasm without explicit connotations.
- *Replacement 2:*
  - Emojis to Replace: 😈
  - Replacement Emojis: 😜
  - Justification: Maintains mischievous tone but mitigates offense; more lighthearted/less suggestive.
- **Revised Tweet:** `Omg yes!` ✊😅 `sex havers roll call` 😜

Tweet to process:

- **Tweet:** `{tweet}`
- **Emojis in Tweet (with functionalities):** `{emoji}`

Step 4: Justification Summary Generation Prompt You will be given an original offensive tweet, its revised version where some emoji were replaced to mitigate offense, and justifications for the offense type, original emoji functionalities, and emoji replacements.

Your task is to **summarize** these justifications into a single, concise paragraph explaining *why* specific emoji were replaced.

Recall the emoji functionality categories:

1. Offense Itself
2. Intensifies Offense
3. Mitigates Offense
4. Unrelated to Offense

We only replaced emojis categorized as **1** or **2**.

Your Response Should Include:

- A summary paragraph integrating the offense type, the functionality of the replaced emojis, and the reason for their replacements.

Inputs Provided:

- Original Tweet: `{tweet}`
- Emojis in Original Tweet: `{emoji}`
- Revised Tweet: `{revised_tweet}`
- Justification of Tweet Offense Type: `{offense_type}`
- Justification of Emoji Functionality: `{emoji_func}`
- Justification of Emoji Replacement: `{emoji_replace}`

Generate Justification Summary:

Prompt of Direct Prompting The prompt for the direct prompting baseline is:
You will be given a tweet with emojis. If this tweet is offensive, try to only replace the emojis with ones that maintain the tweet's tone but mitigate the offense. If the tweet is non-offensive, provide the original tweet as the revised tweet.

Author Index