

Towards a Multi-modal Multi-Label Election-Context Repository for Classifying Misinformation

Prerana Khatiwada*, Qile Wang*, Kenneth E. Barner, Matthew Louis Mauriello

University of Delaware, USA
{preranak, kylewang, barner, mlm}@udel.edu

Abstract

The spread of multimodal election misinformation, where text and images jointly convey misleading narratives, seriously threatens democratic integrity. This emerging challenge demands better automated detection and deeper insight into how such narratives are formed and propagated. Yet, little work has systematically examined the characteristics of such content or evaluated how well Large Language Models (LLMs) can classify it across nuanced categories. We introduce a large-scale, annotated multimodal dataset of election-related social media posts from X.com (formerly Twitter), spanning the 2024 U.S. presidential election. Our dataset captures temporal trends, includes text-only and image-based posts, and is labeled across five nuanced misinformation categories: Conspiracy, Sensationalism, Hate Speech, Speculation, and Satire. Given the high cost and time demands of manual annotation, scalable solutions are essential. To address this, we explore automated labeling in our dataset using six LLMs of varying complexity. We compare three lighter-weight models against three full-scale models using a majority vote mechanism and human validation. Results show that lighter-weight models exhibit higher internal agreement and stability, particularly in classifying subjective categories such as Satire and Speculation. In contrast, larger models demonstrate more variance and lower inter-model reliability. We also inspect where models diverge and identify potential causes of disagreement, such as ambiguous tone, sarcasm, and exaggeration. Our preliminary findings indicate that LLM-driven annotation produces labels that are reliable enough to serve as a usable ground truth, especially for large-scale studies. We aim to facilitate future research in multimodal misinformation detection and annotation-efficient learning.

1 Introduction

Misinformation increasingly undermines modern elections by shaping public opinion, influencing voter behavior, and destabilizing democratic processes (Au, Ho, and Chiu 2022). As more people rely on social media platforms for political information, these networks accelerate the spread of misleading narratives (Moravec, Minas, and Dennis 2019). Although researchers have made significant progress

in detecting text-based misinformation (Su et al. 2020; Mridha et al. 2021), they have paid less attention to multimodal misinformation, where political actors deliberately combine misleading text and images. These actors often exploit visuals with misleading captions or fabricating imagery to distort reality (Botha and Pieterse 2020; Barari et al. 2021). This is particularly effective as people tend to interpret images more intuitively and trust them more readily than text (Nakamura, Levy, and Wang 2019).

Although misinformation continues to shape elections, researchers have produced few standardized datasets for evaluating multimodal misinformation in election-specific contexts, e.g., Raza et al. (2024). Most studies neglect the complex interplay between text and images in political deception. During the 2024 U.S. presidential election, political actors employed increasingly sophisticated tactics, relying heavily on fabricated or misleading visuals to influence public discourse (League 2024). These developments highlight the urgent need for robust, election-specific datasets to support the creation of more effective misinformation detection systems. However, developing such datasets poses significant challenges. Annotators must invest substantial time and effort to manually categorize content, a process that is costly and prone to inconsistency, especially when classifying ambiguous or subjective material (Said et al. 2017). Although some researchers have proposed automated annotation using LLMs (Tan et al. 2024; Su et al. 2022), few have systematically compared how different LLMs perform when labeling diverse misinformation categories with high accuracy. Additionally, the raw data is often messy and unstructured, showing the value of our curated and cleaned dataset for advancing misinformation research.

To help fill the gap in multimodal, election-specific misinformation datasets and explore new approaches for labeling large-scale data more efficiently, we created a new resource composed of text-image pairs collected from the 2024 U.S. presidential election. Our goal is twofold: to analyze misinformation patterns and evaluate LLMs' effectiveness in classifying such content. We tested six LLMs, varying in size and architecture, on tweets containing potential misinformation across five key categories: Conspiracy, Hate Speech, Satire, Sensationalism, and Speculation. We combined model outputs using a majority-vote approach to establish a reliable consensus label and verified a subset

*These authors contributed equally.

through human annotation. Our method draws inspiration from annotation-efficient frameworks like the vote-k strategy, which leverages selective annotation and prompt retrieval to improve performance on language tasks (Su et al. 2022). Our study centers on these research questions:

- **RQ1:** *What are the predominant patterns, themes, and characteristics observed in multimodal election-related misinformation?*
- **RQ2:** *How accurately can LLMs of varying sizes and architectures label multimodal misinformation content across diverse categories, and to what extent do their predictions agree?*
- **RQ3:** *What insights can we extract from the dataset regarding misinformation trends and model behavior?*

Through this work, we contribute foundational results in two key areas (i) the development of a recent and structured dataset of 2024 election-related multimodal posts and (ii) a comprehensive evaluation of six LLMs for multi-label classification grounded in a consensus-based annotation process reinforced by human review for a sample of tweets.

2 Related Work

Here, we review prior work on: (1) multimodal misinformation; (2) election-related narratives; (3) platform-specific misinformation trends; and (4) dataset and annotation gaps that motivate our scalable LLM-based approach.

2.1 Multimodal Misinformation Detection

Previous efforts have mainly focused on integrating text and images for fake news detection (Segura-Bedmar and Alonso-Bartolome 2022; Nakamura, Levy, and Wang 2019; Jindal et al. 2020). Nakamura, Levy, and Wang (2019) introduced Fakeddit, a large-scale multimodal dataset with over 1 million samples, enabling fine-grained fake news classification across multiple categories. Their work demonstrated the importance of multimodal approaches and distant supervision for dataset labeling. Segura-Bedmar and Alonso-Bartolome (2022) further improved multimodal fake news detection, showing that CNN-based architectures achieved 87% accuracy on the Fakeddit dataset, outperforming unimodal approaches. Their findings show that combining text and images is especially effective for categories like Manipulated Content and Satire, where visuals are essential. Work by Zeng et al. (2024) addresses the scarcity of large-scale real-world fact-checking data by using synthetic datasets generated by AI models. Their approach narrows the gap between synthetic and real-world data, demonstrating improved performance of a small multimodal LLM (13B) on fact-checking tasks, even outperforming GPT-4V in some cases. These developments highlight the growing importance of combining multiple modalities improve the effectiveness and reach of fake news detection systems.

2.2 Election-Specific Misinformation Detection

Efforts to detect misinformation related to elections are still emerging. FakeWatch (Raza et al. 2024), for example, focuses on identifying misinformation in North American

election-related news articles using both traditional machine learning models and state-of-the-art language models. While effective, the study highlights the trade-off between computational efficiency and model complexity, emphasizing the continued relevance of classical ML models alongside modern deep learning techniques. Although several datasets and models exist for misinformation detection, few provide reliable ground truth at scale. Many rely on heuristic-based or automated labeling schemes such as keyword matching, hashtags, or user metadata—to infer labels (Chen, Deb, and Ferrara 2022). These approaches often lack nuance and can lead to noisy or overly simplistic labels. For example, some Reddit-based datasets infer misinformation from subreddit themes or cherry-picked comments (Nakamura, Levy, and Wang 2019), which may not reflect the broader discourse. Similarly, labeling tweets based on subjective content or surface-level signals can miss contextual subtleties. Without considering the full content, these methods are often viewed as less reliable than LLM-based approaches. Our approach uses LLMs with consensus techniques to generate annotations that could be more scalable and consistent than traditional manual or heuristic methods.

This is especially important as the complexity and scale of misinformation continue to grow, particularly on dynamic platforms like social media (Aïmeur, Amri, and Brasard 2023), where more advanced techniques are becoming necessary. Misinformation on platforms like TikTok during the 2024 U.S. Presidential Election was studied using a dataset of 1.8 million videos, showing political clusters and misinformation trends through hashtag analysis (Pinto et al. 2024). However, datasets like this lack fine-grained, election-specific labels and rely heavily on weakly supervised methods, with no dedicated benchmark focusing on political election misinformation.

Recent advances in LLMs have enabled high-accuracy annotation across a range of tasks, often exceeding 85% accuracy¹. Building on this progress, we treat each LLM in our study as an independent AI annotator and aggregate their predictions using a majority-vote strategy. To improve reliability, we apply selective human verification to samples with conflicting outputs. This hybrid annotation approach not only reduces labeling costs but also achieves performance levels comparable to human annotations, as supported by recent findings (Goel et al. 2023; Mohta et al. 2023; Mirzakhmedova et al. 2024).

3 Methodology

Our methodology (Figure 1) includes three major steps: data collection, category definition for multi-label classification, and evaluation of LLM annotations using model comparison, majority voting, and human verification.

3.1 Dataset Collection

We collected a total of 76,765 posts from October 17, 2024, to March 11, 2025, covering the period of the U.S. presidential election. Since the academic API of X access was discontinued, we used a basic-tier subscription that only allows

¹<https://www.vellum.ai/llm-leaderboard>

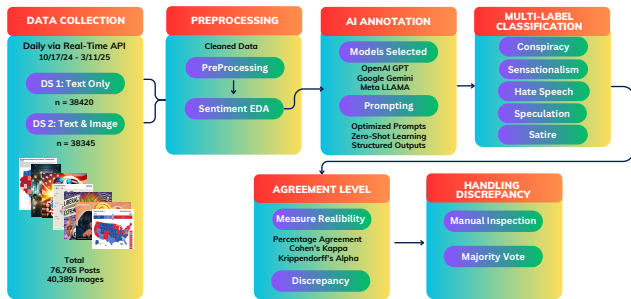


Figure 1: Overall Method

real-time collection of up to 10,000 posts per month and later increased to 15,000 posts monthly after the API subscription changes. To manage the collection rate effectively, we set a daily target to spread out the API usage evenly and collect about four times a day to capture context throughout the day. Full-archive API access is highly restricted and available only through an enterprise subscription, limiting accessibility and data completeness.

Our search operators finds posts that contain both “*election*” and “*2024*” The query is also filtered for original posts (excluding retweets). The dataset includes posts from users worldwide and is not limited to any specific geographic region, as place-tagged posts are under 1% and geo coordinates-tagged posts make up less than 0.1% (Kruspe et al. 2021). To approximate user location, we included the “user location” field provided in the tweet metadata; however, this location information is user-reported and, therefore, serves only as an estimate rather than a verified geographic indicator. We collected two types of datasets concurrently: one consisting of text-only posts and the other including posts with images. Both datasets include additional metadata provided by the API, such as public engagement metrics and content sensitivity. We perform post-processing (e.g., removing URLs, mentions, hashtags, punctuation, and converting text to lowercase) and append additional columns such as `Word_count`, `Clean_text`, and sentiment analysis features generated by VADER².

For tweets containing images, we retrieved the corresponding media files using the media keys and URLs provided in the API responses. The cost of our data collection was approximately \$150 - \$200 per month. We acknowledge that this collection limit is a small percentage of the available data. Due to limited funding, we do not have access to the higher-tier API subscription, which costs \$5,000 per month. A detail of dataset statistics can be seen in Table 1.

3.2 Annotation Categories

We defined five categories for multi-label binary classification based on theoretical and empirical research in misinformation, political discourse, and harmful online content. Focusing on the text modality from both datasets (DS1 and DS2), we aimed to capture distinct but overlapping modes of manipulation and distortion. Two researchers conducted

Metric	Text Only (DS1)	Text + Images (DS2)
Number of Posts	38,420	38,345
Sensitive	404 (1.05%)	497 (1.30%)
Verified	1,036 (2.70%)	1,654 (4.31%)
Retweet Count	0.30 ± 5.37	4.56 ± 58.35
Like Count	1.44 ± 24.63	19.25 ± 232.34
Impression	85.51 ± 951.15	898.93 ± 1604.11
Word Count	28.79 ± 13.96	32.87 ± 13.47
Images Collected	-	40,389

Table 1: Dataset Statistics (Avg ± SD)

a literature review to consolidate related labels into broader categories. For example, we grouped “Hoaxes” and “False Political Claims” under “Speculative or Unverified Claims,” and merged “Inflammatory Rhetoric” with “Derogatory Speech and Defamation” to form “Hate Speech”. The final annotation framework consists of five independent binary labels, each described below in terms of its definition and what it captures. A post may not belong to any category, or it could belong to multiple.

- **Conspiracy** — Misinformation that promotes exaggerated claims such as stolen elections, secret political plots, or suppressed candidates. These narratives often spread rapidly and can distort political perceptions or encourage radicalization (Douglas, Sutton, and Cichocka 2017; Bartlett and Miller 2010).
- **Sensationalism** — Alarmist content featuring dramatic or misleading statements, warnings about political or geopolitical events. Sensationalism is known to drive engagement and contribute to the virality of misinformation (Mourão and Robertson 2019).
- **Hate Speech** — Posts with personal attacks, slurs, or inflammatory rhetoric against political figures or marginalized groups, potentially inciting violence or discouraging civic engagement (Rasaq et al. 2017; Solovev and Pröllochs 2022). Our rationale for including it stems from the recognition that, in some contexts, hate speech can overlap with discussions around polarized or questionable content, which is often associated with misinformation content (Cinelli et al. 2021), for example, when it propagates false stereotypes and targeting specific groups.
- **Speculation** — Bold assertions lacking credible evidence, often involving predictions or allegations about political outcomes or global events (Blom et al. 2021; Rojecki and Meraz 2016).
- **Satire** — Political satire that uses humor, irony, or exaggeration to critique public figures, ideologies, or policies. While often protected as free expression, satire can blur the line between commentary and misinformation (Kulkarni 2017).

3.3 Experiments

Our experimental setup treats each LLM as an independent annotator and compares their outputs to simulate the work of human annotators. For each tweet, we collect predictions

²<https://pypi.org/project/vaderSentiment/>

from six models and use a majority vote among three diverse LLMs to generate high-agreement labels.

- **Lighter-weight Models:** “gpt-4o-mini-2024-07-18”, “Gemini-2.0-Flash”, “Llama-3.1-8B-Instruct”
- **Full Models:** “gpt-4o-2024-08-06”, “gemini-1.5-pro”, “Llama-3.3-70B-Instruct”

These selected models are categorized based on their number of parameter scales and from their benchmark performance on the Vellum.ai LLM Leaderboard³. The selection focused on general performance (MMLU), coding efficiency (HumanEval), reasoning abilities (GPQA), scalability, efficiency, and tool-use/multilingual capabilities. Larger models like GPT-4o and LLAMA 3.3 excel in MMLU, suitable for complex tasks, while models like GPT-4o (90.2%) perform exceptionally well on HumanEval for accurate zero-shot code generation. Lighter-weight models offer efficient alternatives for resource-limited scenarios, especially in reasoning tasks, without significant performance trade-offs.

We adopted a zero-shot learning approach where each LLM was prompted to classify tweets into predefined categories based on detailed definitions. We set the model *Temperature* to 0 to improve reproducibility and minimize variability in responses. For each tweet, the model returned a boolean value (*True* or *False*) for each category, indicating its presence or absence. We enforced structured responses in the output to minimize formatting errors. We designed the following prompt for our experiment:

LLM Prompt

Your task is to accurately classify social media posts related to the U.S. Presidential Election. Determine whether the given post falls into one or more of the following categories: **Conspiracy, Sensationalism, Hate Speech, Speculation, and Satire**. Use the detailed definitions provided for each category and respond with **True** or **False** for each category only.

- **Conspiracy:** *Simplifies complex events by attributing them to secret plots, rejects mainstream information, forms closed belief communities, replaces science with alternative explanations, or frames events as elite deception.*
- **Sensationalism:** *Uses exaggerated or dramatic language, shock and fear appeal, oversimplifies issues, or employs clickbait-style framing to increase engagement.*
- **Hate Speech:** *Contains incitement of discrimination, defamation, hostility, or violence based on identity, or makes false statements that damage a person’s reputation (libel/slander).*
- **Speculation:** *Circulates unverified claims for political advantage, driven by partisan interests, amplified in ideological echo chambers, and sustains political controversy.*
- **Satire:** *Uses humor, political satire, and internet memes to criticize or comment on politics, often spreading through viral online platforms.*

Post: "{post}"

Once we obtain the predicted labels from each model, we begin by identifying agreement or disagreement across categories and compared the consistency of predictions across multiple models. Focusing on model reliability between two

LLMs, we assessed it using two commonly established measures: the percent agreement score, defined as:

$$\text{Percent agreement} = \frac{\text{Number of Matches}}{\text{Total Predictions}} \quad (1)$$

and Cohen’s Kappa, given by $\kappa = \frac{P_o - P_e}{1 - P_e}$. For agreement between three models, we used Krippendorff’s Alpha, given by $\alpha = 1 - \frac{D_o}{D_e}$, which generalizes reliability for multiple raters and categories by quantifying observed versus expected disagreement. We computed Krippendorff’s Alpha by combining predictions from multiple models (e.g., GPT-4o, Gemini 1.5 Pro, Llama 3.3) into a shared matrix and removing rows with missing values. We assign final labels using a majority vote across models. We flag disagreements and manually review a subset to check reliability. Additionally, we conducted a qualitative analysis to examine patterns of disagreement, particularly in categories like hate speech and satire, where agreement scores tended to be significantly lower.

4 Results

Here, we discuss exploratory data analysis and evaluate model predictions across categories to assess their performance and classification tendencies.

4.1 Exploratory Data Analysis

Combining all data and features from DS1 and DS2 in Table 1, we obtained a total of 76,765 entries. It includes 45,787 unique users. 98.83% of posts are not marked as sensitive, and 1.17% are. Verified user accounts make up 3.50% of the data. The overall sentiment distribution consists of 43.14% positive, 35.40% negative, and 21.46% neutral entries.

Hashtags. We began our dataset exploration by examining the most frequent hashtags, which primarily relate to the 2024 U.S. presidential election, political figures (e.g., #trump, #donaldtrump, #kamalaharris), election events (e.g., #election2024, #electionday), and broader political topics (e.g., #politics, #news).

Co-occurrence. To better understand the thematic landscape of our dataset, we constructed a hashtag co-occurrence graph (Figure 2). Each node represents a hashtag, and edges indicate co-occurrence within the same post. The visualization shows several distinct clusters, including a tightly connected group of social movement hashtags (e.g., #blm, #metoo, #feeltheben, #jinjiyanazadi), a global news cluster (#ukraine, #rwanda, #srilanka, #breakingnews), and a politically oriented group centered around #trump, and #uselection2024. Interestingly, we observe strong links between #elonmusk and economic-political terms such as #megabanks and #oligarchs, indicating discussions that critique power structures. Another interesting pattern is the association of #donaldtrump with #immigration and #kamalaharris, reflecting polarized political discourse. Surprisingly, hashtags like #ai, #chatgpt appear within political clusters, implying that AI is being discussed in socio-political contexts beyond its technical scope. These overlaps highlight how different themes ranging from technology to activism are entangled in online discourse.

³<https://www.vellum.ai/llm-leaderboard>

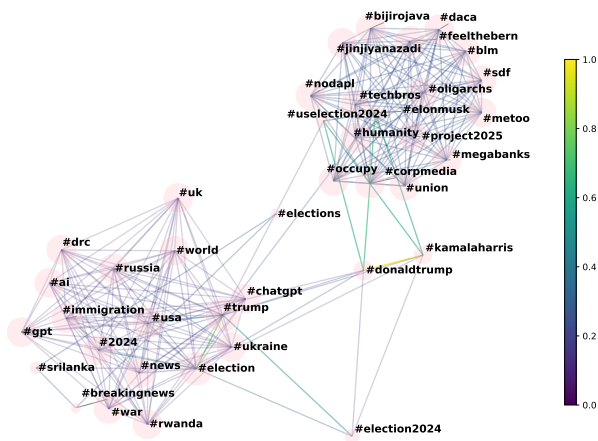


Figure 2: Top 40 hashtag pairs from the 500 most frequent co-occurrences. Node size reflects hashtag frequency; edge color indicates co-occurrence strength.

TweetMap. We also analyzed the tweet distribution across states to understand regional patterns and trends. Figure 3 shows Indiana has the highest representation at 7.67%, followed by Oregon (6.99%), North Dakota (5.54%), Louisiana (5.45%), and California (5.35%). The data has a diverse participation of users from different states, with relatively high activity from the Midwest and Southern states.

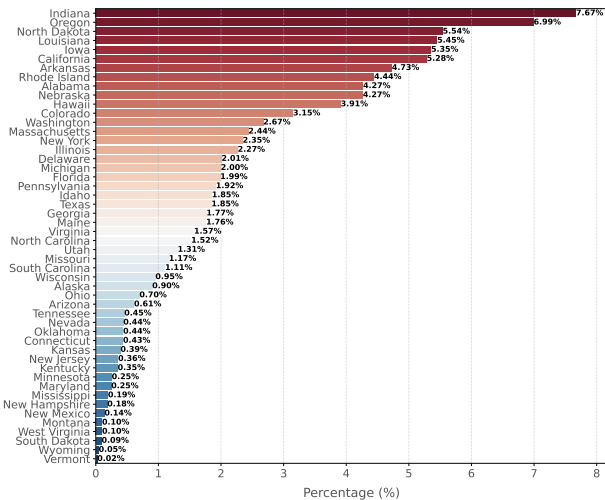


Figure 3: Distribution of Tweets Across U.S. States

Term Polarity. Next, we were interested in understanding how people emotionally respond to certain topics on social media even when those topics aren't mentioned directly in the text. Figure 4 illustrates the sentiment distribution of tweets related to various selected keywords. To remove sentiment bias from keywords like "violence," we recompute the sentiment score after excluding them from the original text. Each bar reflects the proportion of positive, neu-

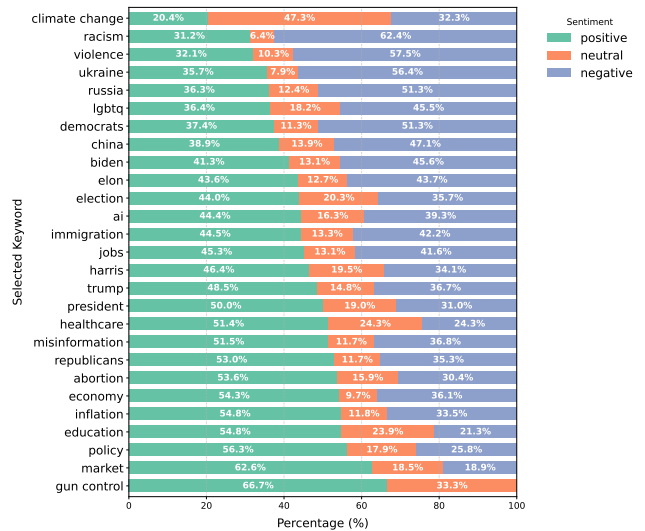


Figure 4: Sentiment distribution by selected keywords.

tral, and negative sentiments for a given term. "gun control" shows the highest positive sentiment at 66.7%, followed by "market" (62.6%) and "policy" (56.3%). On the other hand, "racism" has the highest negative sentiment (62.4%), alongside "violence" (57.5%) and "ukraine" (56.4%). Keywords like "climate change" and "healthcare" display more balanced sentiment distributions. Terms such as "trump," "biden," and "republicans" show mixed sentiment. The average sentiment scores also vary across U.S. states, with New Mexico, Wyoming, and New Hampshire displaying the highest positive sentiment, while West Virginia, Kansas, and Oklahoma show the lowest.

Temporal Sentiment. We were also interested in examining how positive sentiment toward "Trump" and "Harris" changed over time. Figure 5 illustrates the percentage change, either increasing or decreasing, relative to the total discussion about each candidate on a daily basis. Peaks and fluctuations reflect changes in public perception driven by significant events like the Election and Inauguration Day. Overall, Trump shows larger variations, ranging from a high of over 80% to a low of less 15%.

4.2 Model Comparison and Classification

Table 2 provides a comparative analysis of AI models based on their parameters, knowledge cutoffs, computing times, and associated costs, highlighting the trade-offs between model size, performance, and operational expenses. The models used in our experiment had a knowledge cutoff date prior to when the dataset was curated from X.com, to avoid data contamination. This allows us to fairly evaluate the true performance of the LLMs without limiting the inclusion of other political events. The lighter models tend to run faster and are less expensive, while the larger models are more costly and take more time. The human annotation estimate assumes one Amazon Mechanical Turk (AMT) worker spending 10 seconds per post (5 tasks) at \$15/hour (USD).

Model	Parameters (n)	Knowledge Cut-off	Computation Time	Experiment Cost
GPT-4o Mini	8 B*	Oct, 2023	128 min	\$7
GPT-4o	200 B *	Oct, 2023	112min	\$110
Gemini 2.0 Flash	20 B *	Aug, 2024	76 min	\$4
Gemini 1.5 Pro	200 B *	Sept 2024 *	117 min	\$47
Llama 3.1	8 B	Dec, 2023	166 min	\$13 (4x Nvidia L4)
Llama 3.3	70 B	Dec, 2023	240 min	\$75 (4x Nvidia A100)
One Human Annotator (Hypothetical)			213 hrs *	\$3,199 *

Table 2: Comparison of Various LLM Models. * indicate estimated value

Model	Conspiracy	Sensationalism	Hate Speech	Speculation	Satire
GPT-4o Mini	14204 (18.50%)	24069 (31.35%)	3680 (4.79%)	42339 (55.15%)	3419 (4.45%)
GPT-4o	14410 (18.77%)	23321 (30.38%)	3808 (4.96%)	31841 (41.48%)	4096 (5.34%)
Gemini 2.0 Flash	19225 (25.04%)	14773 (19.24%)	5140 (6.70%)	42385 (55.21%)	4786 (6.23%)
Gemini 1.5 Pro	18699 (24.36%)	6617 (8.62%)	4580 (5.97%)	35989 (46.88%)	46824 (61.00%)
Llama 3.1 (8B)	14061 (18.32%)	22147 (28.85%)	5891 (7.67%)	22327 (29.08%)	5668 (7.38%)
Llama 3.3 (70B)	19689 (25.65%)	30040 (39.13%)	7211 (9.39%)	41801 (54.45%)	6703 (8.73%)

Table 3: Model predictions by category: Counts show total detections; percentages (in parentheses) reflect category proportion. Bolded values indicate category-wise minimum and maximum.

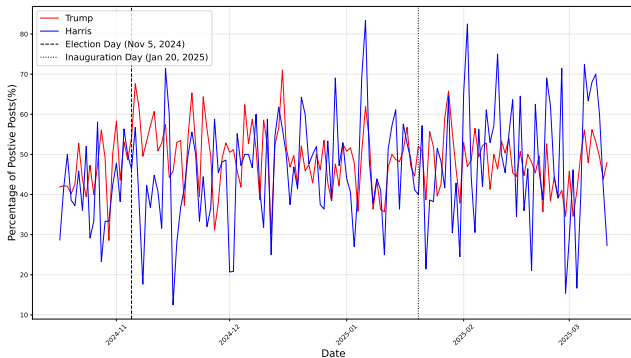


Figure 5: Temporal Percentage Change in Positive Sentiment for Posts Mentioning “Trump” and “Harris”

Table 3 compares six models based on classification counts across five categories and their distributions. Results indicate that models generally detect a high proportion of tweet content as Speculation, especially for models like GPT-4o Mini (55.15%) and Gemini 2.0 Flash (55.21%). The Llama 3.3 model identifies a higher proportion of Conspiracy (25.65%) and Sensationalism (39.13%) compared to other models. The lowest category detected by most models is generally Hate Speech, with detection percentages ranging from 4.79% (GPT-4o mini) to 9.39% (Llama 3.3 70B). This suggests that Hate Speech is the least common classification across the datasets being analyzed. The most striking anomaly is the extremely high Satire detection rate by Gemini 1.5 Pro (61.00%), which is much higher compared to all other models where the Satire detection usually falls between 4.45% and 8.73%. This indicates a strong bias or possibly an overfitting issue towards the Satire category in the Gemini 1.5 Pro model. Faster models like Gemini 2.0

Flash sacrifice accuracy in certain categories but are computationally efficient. Llama 3.3 (70B) offers broader category coverage but at the cost of slower processing, highlighting a trade-off between accuracy and speed.

In addition, 26.04% ($n = 18,274$) of posts were labeled false in all categories by all three full LLM models, meaning none of the models marked any category as true. This indicates a consistent absence of misinformation. For example: “I love helping companies become successful, which is why this election is so important to me and to all of us. This election is a battle for entrepreneurs. #Election2024 #FreeSpeechMatters.” If a majority vote approach is used across the three LLMs, the percentage of posts labeled as not containing any misinformation is likely to be higher. However, the topic of “Election 2024” is expected to contain a greater concentration of misinformation compared to other events. Conversely, we also observed posts labeled as true across all five categories. For example, GPT-4o labeled the following post as true in every category: “Sure JoePedo did, the 5Eyes Governments would NEVER lie to US! Those 10 million missing voters just stayed home for the 2024 election! Probably have Covid-24!”

4.3 Agreement Metrics for Model Evaluation

We compare model families rather than just size, as lighter models (e.g., GPT-4o Mini) perform similarly to full versions (e.g., GPT-4o), with minimal gaps ($\kappa = 0.82 - 0.97$). Results in Table 4 indicate that the larger models are not consistently aligned in their predictions across categories, particularly when compared with each other, with Cohen’s Kappa scores often near zero, particularly for categories like Sensationalism, Hate Speech, and Satire. This indicates that the models are making almost entirely independent predictions, which is confirmed by the matched count and percentage agreement showing near-zero overlap or

even negative (perfect disagreement). Krippendorff’s Alpha scores, although fair for categories like Hate speech, are considered poor for more subjective categories like Speculation and Satire, suggesting difficulty in consistent classification.

In contrast, the smaller models demonstrate comparatively better agreement among themselves particularly in categories like Conspiracy, Hate Speech, and Satire, where Cohen’s Kappa scores exceed 0.6 and percentage agreements are mostly above 90%, although their agreement with the larger models remains poor. This indicates a more standardized approach, likely due to their simpler architectures or training scopes. Results suggest larger models may be optimized differently, leading to divergence in nuanced cases, while high agreement on categories like Hate Speech indicates some are easier to classify consistently.

As a final step, we qualitatively analyzed the instances of disagreement for each category to work towards establishing consensus label using a majority-vote approach, where the most frequently predicted label was considered the correct classification. When a tie or disagreement occurred, we conducted manual verification to determine the final label.

4.4 Manual Assessment of Model Divergence and Consensus Labels

Because lighter-weight models tended to struggle with speculation, while the full models often had difficulty with satire, we chose to focus our evaluation on cases where their predictions diverged. Specifically, we analyzed unmatched outputs to understand the nature of these disagreements. We reviewed 100 text samples, focusing on categories where models demonstrated moderate to high agreement, to determine whether the consensus labels aligned with human judgment.

Speculation. Lighter-weight models often disagreed on tweets related to speculation, particularly those involving election conspiracy theories, foreign interference, or distrust in the political process. These tweets frequently reference hypothetical scenarios or secret plans, making them prone to speculative interpretation. Examples include: “*Gen Michael Flynn: Will The November Election Happen or Is Disaster Coming*” and “*They are 100% gonna try to pull some WW3 Shit to cancel the 2024 Election*”—both reflecting predictions rooted in personal belief rather than evidence.

Some tweets blend exaggeration or sarcasm, blurring the line between speculation and satire. For example, “*The Media are now talking about FORCING ‘kamala’ OUT OF THE PRESIDENTIAL ELECTION AND BRINGING BACK ‘biden’*” may read as satirical but can be misclassified as speculative depending on model interpretation. These inconsistencies highlight the challenge models face in distinguishing speculative content from satire or criticism, emphasizing the need for improved contextual and tonal understanding. We also evaluated effectiveness of majority voting for labeling; lighter models reliably detected sensationalism through common provocative framing patterns.

Satire. The tweets where full models disagree on the classification of satire are heavily characterized by exaggerated, hyperbolic statements and mocking language aimed at political figures and processes. For example, tweets like “*The*

Redcoats Are Coming! U.K. Labour Party to Campaign for Kamala (Isn’t This Election Interference?)” and “*Reported: BIDEN said he wants ‘Trump in jail’. That’s rich. Two weeks before the 2024 Election...*” both exaggerate scenarios and present them in a tone intended to ridicule political actors. Similarly, phrases like “*All votes for the executed Kameltoe Harris Must & Shall not be counted!*” are likely intended to be sarcastic or hyperbolic rather than a genuine call to action. This distinction between satire and actual hate speech or speculation requires a more sophisticated understanding of intent and tone, which models struggle to achieve consistently. The disagreements here indicate their ability to detect satirical intent is limited, particularly when sarcasm overlaps with politically charged topics.

Similarly, in the conspiracy category, full models showed strong sensitivity to content involving election interference, manipulation, and political corruption. They consistently identified conspiracy narratives featuring themes like fraud, foreign influence, and biased rulings. These results suggest that majority voting among full models is effective for accurately labeling conspiracy-related content.

5 Planned Evaluations and Project Tracks

The planned evaluations for this project aim to improve how we analyze tweets related to election content by focusing on i) image annotation, ii) visual impact assessment, and iii) improving Natural Language Understanding (NLU) by combining text and images.

The first track, **Image Annotation**, investigates whether ML models can categorize images in the dataset with minimal preprocessing. This includes grouping images by political alignment (*e.g.*, left-leaning vs. right-leaning), textual content (*e.g.*, dense infographics vs. short slogans like “Make America Great Again”), image type (*e.g.*, memes, cartoons, photographs), and propaganda strategy (*e.g.*, fear-mongering, emotional appeal). For instance, a meme portraying a politician negatively may be labeled both as “Right-Wing” and “Ridicule” under propaganda strategy.

Second, the **Visual Impact and Content Analysis** track examines how well image classifications predict user engagement. We construct similarity networks based on shared visual features, such as facial expressions, symbols, or design styles, and analyze their spread. For example, images of political figures like Donald Trump making dramatic gestures may generate more engagement (*e.g.*, retweets, comments, impressions) than informational posts without human faces. We also track how these posts circulate over time to understand how specific visual styles influence virality.

Finally, we introduce the **NLU Enhancement** track, which proposes a unified multimodal framework that integrates textual content, visual elements, propaganda strategies, and user engagement metrics. While prior studies have shown the value of combining text and images for classification accuracy (Qi et al. 2019), our approach models the virality of themes and analyzes how language and visuals drive engagement and spread. For example, rather than simply labeling an infographic titled “How to Vote” as “Informational,” our model evaluates how design elements (*e.g.*, icons, colors, typography) and captions

Models		Conspiracy	Sensationalism	Hate Speech	Speculation	Satire
Light	GPT-4o Mini vs Llama3.1(8B)	0.78 (93.19%)	0.63 (84.14%)	0.65 (95.78%)	0.46 (71.72%)	0.56 (94.89%)
	GPT-4o Mini vs Gemini2.0 Flash	0.75 (91.25%)	0.53 (81.80%)	0.71 (96.56%)	0.68 (84.46%)	0.57 (95.20%)
	Llama3.1(8B) vs Gemini2.0 Flash	0.70 (89.24%)	0.54 (82.45%)	0.68 (95.38%)	0.44 (70.82%)	0.52 (93.41%)
	Average Pairwise	0.74 (91.23%)	0.57 (82.80%)	0.68 (95.91%)	0.53 (75.67%)	0.55 (94.50%)
	Krippendorff’s Alpha	0.74	0.57	0.68	0.51	0.55
Full	GPT-4o vs Gemini1.5 Pro	0.76 (91.65%)	0.24 (73.63%)	0.02 (89.28%)	0.41 (70.67%)	-0.03 (37.41%)
	GPT-4o vs Llama 3.3(70B)	0.74 (90.83%)	0.73 (87.41%)	0.63 (94.92%)	0.63 (80.93%)	0.61 (94.69%)
	Gemini1.5 Pro vs Llama3.3(70B)	0.80 (92.37%)	0.22 (67.42%)	-0.003 (85.25%)	0.42 (70.85%)	-0.05 (37.50%)
	Average Pairwise	0.77 (91.62%)	0.40 (76.15%)	0.22 (89.82%)	0.49 (74.15%)	0.18 (56.53%)
	Krippendorff’s Alpha	0.77	0.39	0.22	0.48	-0.15

Table 4: Reliability Measurements: Cohen’s Kappa scores (agreement percentage) for lighter and full models, along with average pairwise scores. Krippendorff’s alpha is computed across three same-size LLMs; bold values indicate highest values.

affect perception and sharing. It also tracks how politically charged memes (*e.g.*, exaggerated depictions of politicians) spread more widely than factual content, revealing visual propaganda’s reach.

The novelty of this work lies in its ability to quantify and model how propaganda strategies work at scale by integrating textual cues and visual characteristics. By creating a network of similar posts and evaluating their spread, the framework can detect biased narratives, even when the text alone is neutral but the image carries provocative implications. We will incorporate hierarchical classification models, where content is first classified by its general type (*e.g.*, meme, infographic, photo), then by its theme (*e.g.*, political alignment, propaganda strategy), and finally by its potential impact (*e.g.*, high engagement vs. low engagement).

To deepen our understanding of model behavior, we plan to explore how stochasticity, such as temperature settings, affects annotation consistency through ablation studies. Although GPT models showed the lowest error rates, our dataset contains some missing predictions, which we aim to address through detailed classification metrics, error analysis, and case studies of misclassified examples. While high-agreement labels were generated using a majority vote from three diverse LLMs, handling tie cases remains ongoing work. Additionally, we aim to more clearly disentangle hate speech from misinformation, recognizing them as related but distinct phenomena that should be analyzed in parallel rather than grouped under a single taxonomy.

6 Limitations

The presented study faces some limitations, primarily due to the subjectivity involved in labeling nuanced categories like satire. Even human annotators face challenges in maintaining consistency, making it unsurprising that LLMs exhibit similar complexities, particularly when compared to more clearly defined categories, as evidenced by the lower agreement scores for satire. Additionally, our annotation process diverges from traditional fact-checking labels such as “true” or “fake,” limiting compatibility with existing datasets and pre-trained models. This opens up space to develop richer, more context-aware labeling systems that better capture the complexity of online content. Our current framework does

not differentiate between bot and human-generated content, which presents another potential limitation, as the behavior and language patterns of bots may differ significantly from those of real users. While not addressed in this study, integrating bot detection methods in future work could help refine our analysis and offer a better understanding of how misinformation spreads across user types.

7 Conclusion

This study introduces a large-scale annotated dataset of election-related social media posts and evaluates LLMs for multi-label misinformation classification. Through consensus techniques and human oversight, we provide scalable, consistent labels that improve on traditional manual or heuristic approaches. While labels are not verified individually, our approach provides a valuable baseline and reflects a realistic compromise between quality and scalability, particularly when applied to large corpora such as millions of posts from online social networks. This facilitates exploring co-occurring misinformation signals like hate speech appearing alongside satire, which can be challenging to disentangle manually. The majority voting across these models mimics common human annotation protocols while drastically reducing the time and cost burdens of full manual labeling. Results indicate that lighter-weight models achieve higher inter-model agreement, particularly in subjective categories like *Satire* and *Speculation*, while larger models exhibit greater variance. These findings underscore the complexity of subjective classification and the trade-offs between model size and reliability. Future work will focus on three directions: improving image annotation through semi-supervised learning and prompt optimization, incorporating visual features to improve multimodal analysis, and modeling engagement dynamics to better understand how misinformation spreads. To encourage further research in this area, we plan to release our dataset in the future. Please contact the first author for early access.

Acknowledgments

We thank all the members of the SMIDGen team from UD’s Sensify Lab for their help with data scraping and tool development and Dr. Benjamin Bagozzi for his insights.

References

- Aïmeur, E.; Amri, S.; and Brassard, G. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1): 30.
- Au, C. H.; Ho, K. K.; and Chiu, D. K. 2022. The role of online misinformation and fake news in ideological polarization: barriers, catalysts, and implications. *Information systems frontiers*, 1–24.
- Barari, S.; Lucas, C.; Munger, K.; et al. 2021. Political deepfake videos misinform the public, but no more than other fake media. *OSF Preprints*, 13: 1–16.
- Bartlett, J.; and Miller, C. 2010. *The power of unreason: Conspiracy theories, extremism and counter-terrorism*. Demos London.
- Blom, J. N.; Rønlev, R.; Hansen, K. R.; and Ljungdalh, A. K. 2021. The potentials and pitfalls of interactional speculations by journalists and experts in the media: The case of Covid-19. *Journalism Studies*, 22(9): 1142–1160.
- Botha, J.; and Pieterse, H. 2020. Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited*, 57.
- Chen, E.; Deb, A.; and Ferrara, E. 2022. # Election2020: the first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science*, 1–18.
- Cinelli, M.; Pelicon, A.; Mozetič, I.; Quattrociocchi, W.; Novak, P. K.; and Zollo, F. 2021. Dynamics of online hate and misinformation. *Scientific reports*, 11(1): 22083.
- Douglas, K. M.; Sutton, R. M.; and Cichocka, A. 2017. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6): 538–542.
- Goel, A.; Gueta, A.; Gilon, O.; Liu, C.; Erell, S.; Nguyen, L. H.; Hao, X.; Jaber, B.; Reddy, S.; Kartha, R.; et al. 2023. LLMs accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, 82–100. PMLR.
- Jindal, S.; Sood, R.; Singh, R.; Vatsa, M.; and Chakraborty, T. 2020. Newsbag: A multimodal benchmark dataset for fake news detection. In *CEUR workshop proc*, volume 2560, 138–145.
- Kruspe, A.; Häberle, M.; Hoffmann, E. J.; Rode-Hasinger, S.; Abdulahhad, K.; and Zhu, X. X. 2021. Changes in Twitter geolocations: Insights and suggestions for future usage. *arXiv preprint arXiv:2108.12251*.
- Kulkarni, A. 2017. Internet meme and Political Discourse: A study on the impact of internet meme as a tool in communicating political satire. *Journal of Content, Community & Communication Amity School of Communication*, 6.
- League, A.-D. 2024. The ADL Debunk: False Narratives Around the 2024 Presidential Election. Accessed: 2025-03-24.
- Mirzakhmedova, N.; Gohsen, M.; Chang, C. H.; and Stein, B. 2024. Are Large Language Models Reliable Argument Quality Annotators? In *Conference on Advances in Robust Argumentation Machines*, 129–146. Springer.
- Mohta, J.; Ak, K.; Xu, Y.; and Shen, M. 2023. Are large language models good annotators? In *Proceedings on*, 38–48. PMLR.
- Moravec, P. L.; Minas, R. K.; and Dennis, A. R. 2019. Fake News on Social Media. *MIS quarterly*, 43(4): 1343–A13.
- Mourão, R. R.; and Robertson, C. T. 2019. Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism studies*, 20(14): 2077–2095.
- Mridha, M. F.; Keya, A. J.; Hamid, M. A.; Monowar, M. M.; and Rahman, M. S. 2021. A comprehensive review on fake news detection with deep learning. *IEEE access*, 9: 156151–156170.
- Nakamura, K.; Levy, S.; and Wang, W. Y. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.
- Pinto, G.; Bickham, C.; Salkar, T.; Luceri, L.; and Ferrara, E. 2024. Tracking the 2024 US Presidential Election Chat-ter on Tiktok: A Public Multimodal Dataset. *arXiv preprint arXiv:2407.01471*.
- Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, 518–527. IEEE.
- Rasaq, A.; Udende, P.; Ibrahim, A.; and Oba, L. 2017. Media, politics, and hate speech: A critical discourse analysis. *E-Academia Journal*, 6(1).
- Raza, S.; Khan, T.; Chatrath, V.; Paulen-Patterson, D.; Rahman, M.; and Bamgbose, O. 2024. FakeWatch: a framework for detecting fake news to ensure credible elections. *Social Network Analysis and Mining*, 14(1): 142.
- Rojecki, A.; and Meraz, S. 2016. Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society*, 18(1): 25–43.
- Said, A. F.; Kashyap, V.; Choudhury, N.; and Akhbari, F. 2017. A cost-effective, fast, and robust annotation tool. In *2017 IEEE applied imagery pattern recognition workshop (AIPR)*, 1–6. IEEE.
- Segura-Bedmar, I.; and Alonso-Bartolome, S. 2022. Multi-modal fake news detection. *Information*, 13(6): 284.
- Solovev, K.; and Pröllochs, N. 2022. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM web conference 2022*, 3656–3661.
- Su, H.; Kasai, J.; Wu, C. H.; Shi, W.; Wang, T.; Xin, J.; Zhang, R.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.
- Su, Q.; Wan, M.; Liu, X.; and Huang, C.-R. 2020. Motivations, methods and metrics of misinformation detection: an NLP perspective. *Natural Language Processing Research*, 1(1): 1–13.
- Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; and Liu, H. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Zeng, F.; Li, W.; Gao, W.; and Pang, Y. 2024. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. *arXiv preprint arXiv:2409.19656*.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **NA**
 - (e) Did you describe the limitations of your work? **Yes**
 - (f) Did you discuss any potential negative societal impacts of your work? **NA**
 - (g) Did you discuss any potential misuse of your work? **NA**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? **NA**
 - (b) Did you mention the license of the assets? **NA**
 - (c) Did you include any new assets in the supplemental material or as a URL? **NA**
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **NA**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **NA**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **Yes, early access is available upon request.**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? **NA**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
 - (d) Did you discuss how data is stored, shared, and de-identified? **NA**

Ethical Statement

The use of LLMs for automated annotation presents ethical challenges, particularly related to bias and mislabeling. Ensuring fair, transparent, and responsible labeling is essential, especially in sensitive contexts. Future work should prioritize cross-validating model outputs and integrating human oversight to reduce bias and enhance reliability.