

Investigating Vaccine Buyer’s Remorse: Post-Vaccination Decision Regret in COVID-19 Social Media Using Politically Diverse Human Annotation

Miles Stanley^{*†1}, Soumyajit Datta^{*‡2}, Ashutosh Kumar², Ashiqur R. KhudaBukhsh^{2§}

¹University of Washington, Seattle

²Rochester Institute of Technology

milesds@cs.washington.edu, sd3528@rit.edu, ak1825@rit.edu, axkvse@rit.edu

Abstract

A significant gap exists in datasets regarding post-COVID-19 vaccination experiences, particularly “vaccine buyer’s remorse”. Understanding the prevalence and nature of vaccine regret, whether based on personal or vicarious experiences, is vital for addressing vaccine hesitancy and refining public health communication. In this paper, we curate a novel dataset from a large YouTube news corpus capturing COVID-19 vaccination experiences, and construct a benchmark subset focused on vaccine regret, annotated by a politically diverse panel to account for the subjective and often politicized nature of the topic. We utilize large language models (LLMs) to identify posts expressing vaccine regret, analyze the reasons behind this regret, and quantify its occurrence in both first and second-person accounts. This paper aims to (1) quantify the prevalence of vaccine regret; (2) identify common reasons for this sentiment; (3) analyze differences between first-person and vicarious experiences; and (4) assess potential biases introduced by different LLMs. We find that while vaccine buyer’s remorse appears in only $< 2\%$ of public discourse, it is disproportionately concentrated in vaccine-skeptic influencer communities and is predominantly expressed through first-person narratives citing adverse health events.

1 Introduction

The global COVID-19 vaccination campaign unfolded amidst a deeply polarized social landscape (KhudaBukhsh et al. 2021) and a concurrent *infodemic*, an overabundance of information, both accurate and misleading, that makes it difficult for people to find reliable guidance (do Nascimento et al. 2022). This environment, fueled by the rapid circulation of confusing and often contradictory health content on social media, created fertile ground for post-decisional sentiments like vaccine regret to emerge and spread. While a substantial body of research has explored the drivers of pre-vaccination hesitancy, a significant gap remains in our un-

derstanding of post-vaccination sentiment, particularly the phenomenon of *vaccine buyer’s remorse*. Understanding the prevalence and nature of this vaccine regret is vital for refining public health communication, addressing the long-term erosion of institutional trust, and preparing for future health crises.

The concept of *decision regret*, the distress an individual feels after making a health-related choice, is a well-established area of study (Becerra-Perez et al. 2016; Zeelenberg and Beattie 1997; Brehaut et al. 2003). Recent work has begun to explore this phenomenon in the context of COVID-19, linking the experience of adverse events to increased regret and a subsequent unwillingness to receive booster doses (Luo et al. 2022). Studies have also identified perceived coercion and disillusionment with vaccine efficacy as key drivers of this sentiment (Tayhan, Tayhan, and Büyük 2025). This regret is often shaped by social media, where personal anecdotes about side effects can create negative expectations and amplify perceived negative experiences through the nocebo effect (Clemens et al. 2023), especially since compelling narratives can sway medical decisions even when presented alongside contradictory statistical data (Line et al. 2024).

For decades, public health agencies have conducted post-market safety surveillance through formal channels like the Vaccine Adverse Event Reporting System (VAERS), a passive system that relies on voluntary reports from the public and clinicians (Shimabukuro et al. 2015). In the digital age, social media platforms like YouTube have become vast, informal analogs to VAERS, hosting millions of unsolicited, user-generated accounts of personal and vicarious health experiences. Analyzing this discourse offers an opportunity to understand public sentiment at scale, yet it presents significant methodological challenges. Distinguishing first-person accounts from vicarious narratives, interpreting the nuanced emotion of regret, and mitigating potential model biases requires a sophisticated analytical approach.

To address these challenges, this study introduces a novel dataset and a multi-stage hybrid inference pipeline to analyze a large corpus of YouTube comments related to COVID-19 vaccination.

Contributions: First, we focus on an underexplored as-

*These authors contributed equally.

†Work done at RIT

‡Soumyajit Datta and Ashiqur R. Khudabukhsh co-mentored Miles Stanley on the project.

§Ashiqur R. KhudaBukhsh is the corresponding author. Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

pect of vaccine discourse on the social web – vaccine regret or vaccine buyer’s remorse. Second, we create a benchmark dataset for vaccine regret, annotated by a politically diverse panel of raters to account for the subjective and often politicized nature of the topic (Dolman et al. 2023). Our dataset consists of 2,000 YouTube comments each annotated by three raters (overall, 201 unique raters)¹. We develop and evaluate a computational pipeline to classify comments based on narrative perspective (first-person vs. vicarious) and the presence of regret. Finally, we use this pipeline to conduct a large-scale analysis of our corpus, guided by four primary research questions: (1) quantify the prevalence of vaccine regret; (2) identify the common reasons cited for this sentiment; (3) analyze the differences between first-person and vicarious expressions of regret; and (4) assess potential biases in our models’ classifications.

2 Related Work

Our work is situated at the intersection of public health, psychology, and natural language processing. We draw upon existing literature in three primary areas: the psychological underpinnings and real-world manifestations of vaccine regret, the application of large language models (LLMs) to analyze health-related social media data, and the broader context of public health challenges in the digital age.

Methodologically, our work leverages LLMs to move beyond traditional sentiment analysis, which is often insufficient for capturing the complex and often ambivalent emotion of regret. Traditional sentiment analysis often fails to capture the complexity of user opinions, such as sarcasm or mixed emotions (Liu et al. 2022). Recent studies have demonstrated that LLMs can perform more sophisticated, multi-layered sentiment and topic analysis, identifying not only positive or negative sentiment but also discrete emotions and their underlying drivers (Yin, Han, and Nie 2024). Similar work has used deep learning models to monitor public opinion and extract reported side effects from Twitter (Portelli et al. 2022), (Jain et al. 2025). LLMs have proven effective in specific information extraction tasks, such as identifying adverse events following vaccination from social media posts with high precision (Li et al. 2025), (Sehgal et al. 2025) and (Zhou et al. 2025). However, the literature also cautions that LLMs are not without limitations; their performance is highly dependent on effective prompt engineering, and they can be prone to factual inaccuracies and inherent biases, necessitating careful validation against human-annotated data (He et al. 2024; Crawl et al. 2025).

Prior work also examines broad vaccine debates, latent arguments, and anti-vaccine themes, but none focus on post-vaccination regret. Pacheco et al. (2022, 2023) analyze vaccine narratives and concept learning; Islam and Goldwasser (2025b,a, 2022) study campaign messaging and LLM-assisted theme discovery; and Wawrzuta et al. (2021) analyze anti-vaccine arguments. Our paper uniquely targets *vaccine buyer’s remorse*, distinguishing first- vs third-

¹Dataset, annotator details, and an expanded version with supplemental information (SI) are publicly available at <https://github.com/Social-Insights-Lab/Vaccine-Buyers-Remorse>.

person narratives and quantifying regret prevalence using a novel benchmark dataset.

Finally, our annotation strategy is grounded in recent work addressing annotator subjectivity and bias. Labeling politically charged content is inherently subjective. Weerasooriya et al. (2023) demonstrated that annotators’ political beliefs systematically influence how they perceive and label potentially offensive content. They found that recruiting a politically diverse panel of annotators is a crucial step in understanding and accounting for these perceptual differences. Our study employs a similar methodology as in Crawl et al. (2025); Pofcher et al. (2025) to create a benchmark dataset that explicitly accounts for political diversity.

3 Dataset

Data Collection: We curate a dataset $\mathcal{D}_{newspool}$ of 80,307,930 comments posted on 65886 YouTube videos ($\mathcal{V}_{newspool}$) from the official channels of three major U.S. cable news networks: Fox News, CNN, and MSNBC via the official YouTube Data API v3. This dataset has found prior use in studying election-related discourse (KhudaBukhsh et al. 2021, 2022), health-related discourse (Yoo and KhudaBukhsh 2023), LGBTQ+ discourse (Pofcher et al. 2025), rater subjectivity (Weerasooriya et al. 2023; Pandita et al. 2024; Dutta et al. 2025b), and political identity projections (Mittal, Chawla, and KhudaBukhsh 2024). This selection was made to capture a spectrum of potential political viewpoints among commenters, a key factor in exploring variations in sentiment. The comments for the videos selected for inclusion were published within the timeframe from 14th December 2020 to 31st October 2024 and feature all types of content such as daily briefings, news updates, or official announcements but not particularly related to COVID-19. Along with these mainstream sources, we also included 981 videos (\mathcal{V}_{influ}) with 847,702 comments (\mathcal{D}_{influ}) from prominent YouTube influencers taking part in the vaccine discourse within the same timeframe. This addition was made with the understanding that many individuals now consume news from a variety of online creators (Zimmermann, Klee, and Kaspar 2023), and the discussions in their comment sections warrant similar scholarly attention.

Two independent reviewers with a consensus manually categorized influencer channels based on a qualitative review of their content. We define channels as “*pro-vaccine*” if their content consistently aligned with public health guidance and encouraged vaccination. Conversely, channels were defined as “*vaccine-skeptic*” if they frequently questioned the safety or efficacy of vaccines, focused on adverse events, or expressed opposition to vaccine mandates.

Identifying Vaccine-Relevant Comments: To build our corpus, we first performed an initial, keyword-based filtering of all collected comments to isolate those relevant to vaccination. This process identified comments containing keywords related to personal experiences, side effects, and potential regret. After the filtration, we got total comments of 1,370,101 (\mathcal{D}_{news}) from 54,666 videos (\mathcal{V}_{news}).

Constructing the Benchmark and In-the-wild Dataset: From the filtered comments dataset \mathcal{D}_{news} , we construct a

Example Comment	Regret Label	Subject Label
<i>I got my first shot, then tried to get a mamogram and they will not do it until 4 weeks after my second shot. apparently the limp nods have a chance of swelling. I wouldn't have gotten the shot if i knew that. they are not telling us everything.</i>	Positive	First-Person
<i>i didn't vote for world shut down. i can do the research myself and have decided this particular vaccine is not safe for many reasons. my own doctor is no longer recommending it and regrets taking it himself.</i>	Positive	Third-Party
<i>coming home after my booster shot, i just crossed paths with the coroner as he was removing a body from my apt. building - covid is everywhere. get vaxed!</i>	Negative	First-Person
<i>isn't dying the activity you do after the vaccine</i>	Negative	Unspecified

Table 1: Examples of manually annotated comments from the dataset.

dataset designed to categorize content based on sentiments and perspectives related to vaccine regret. As the *Positive for Regret* class is considerably less common than irrelevant or neutral comments, a simple random sampling for our benchmark dataset would be inefficient. To address this, we employ a multi-faceted approach to identify comments more likely to be valuable for human annotation. This involved using a combination of simple regular expressions alongside zero-shot and few-shot prompting with an LLM to classify a large, unseen pool of comments. This process enabled us to purposefully sample comments identified as likely *Positive for Regret*, creating a set for our full crowd-sourced annotation study and ensuring a more balanced and functional benchmark dataset \mathcal{D}_{bench} comprising 2,000 comments. For classification task we divide \mathcal{D}_{bench} into a training split of 80% of our benchmark dataset (\mathcal{D}_{train}), with the remaining portion reserved for testing (\mathcal{D}_{test}). After validation, we processed a total of 600,000 comments. This corpus was balanced between mainstream news sources (300,000 comments, with 100,000 from each of Fox News, CNN, and MSNBC) and influencer channels (300,000 comments, with 150,000 from pro-vaccine influencers and 150,000 from vaccine-skeptic influencers). These comments were categorized as in-the-wild dataset (\mathcal{D}_{wild}).

The annotation scheme captures two key dimensions:

Vaccine Regret: This dimension categorizes comments into one of two classes: *Positive for Regret* and *Negative for Regret*. The *Positive for Regret* class includes both *Explicit Regret* (direct statements like “I regret taking the vaccine”) and *Implicit Regret* (statements strongly suggesting dissatisfaction, e.g., “I wish I never got it, my health has been terrible since”). The *Negative for Regret* class includes comments that are unrelated to the topic or are neutral statements about an individual’s vaccination status.

Narrative Perspective: For comments indicating regret, this dimension identifies the narrative point of view. Categories include *First-Person* (personal experience), *Third-Party/Vicarious* (reporting another specific individual’s knowledge), and *Unspecified*. We only categorized regret for comments that were from a first or third party, not unspecified (which includes general statements, e.g., “People regret taking this vaccine”).

Annotation Study Design. We conduct a crowd-sourced annotation study designed to mitigate potential political bias in subjective annotations to generate our \mathcal{D}_{bench} labels.

Annotator Recruitment: Following the methodology of Weerasooriya et al. (2023), Crawl et al. (2025), and Pofcher

et al. (2025), we recruited a panel of annotators with diverse, self-identified political affiliations: Republican, Democrat, and Independent.

Annotation Process: Each comment in the \mathcal{D}_{bench} was independently annotated by one annotator from each of the three political groups.

Disagreement Resolution: Prior literature has considered diverse approaches to resolving inter-annotator disagreements (e.g., majority voting (Davidson et al. 2017; Wiegand, Ruppenhofer, and Kleinbauer 2019) or third objective instance (Gao and Huang 2017)). The final label for each comment on each dimension was determined by a majority vote.

Annotation details: We used Prolific² to recruit annotators, while hosting the annotation questionnaire on our custom-built platform. A total of 2,000 comments were divided into 67 batches of 30, with 201 unique annotators participating in the labeling process. The annotator pool was evenly distributed across political affiliations (Democrat, Republican, Independent), with a near-equal gender ratio (106 male, 95 female) and an average age of 42.5 years. All annotators were US citizens. The median annotation time was ~ 19 minutes. Each annotator was compensated \$4 per batch, estimated for a 30-minute task.

4 Methodology and Experiment Design

Zero and Few Shot Classification: We select eight diverse and widely-used LLMs (both open and closed source and ranging from 7B to 70B in size): *mistral-small13.2:24b* (MistralAI 2025), *mistral:7b* (Jiang et al. 2023), *mixtral:8x22b* (Jiang et al. 2024b), *llama3.1:8b* (Grattafiori et al. 2024), *gemma3:12b* (Team et al. 2025), *qwen2.5:7b* (Bai et al. 2023), *llama3.1:70b* (Grattafiori et al. 2024), and *gpt-4o-mini* (Hurst et al. 2024). We perform zero-shot classification on \mathcal{D}_{bench} using default hyperparameters with enforced JSON outputs, following best practices from (Ziems et al. 2024), to assess the models’ out-of-the-box reasoning ability without task-specific tuning. We also evaluate \mathcal{D}_{bench} on a few-shot setting (Brown et al. 2020) to understand whether minimal supervision improves performance and consistency across models.

Supervised Classification: For supervised classification, we finetune three models with varying architectures

²Prolific: <https://www.prolific.com>

Model	Zero-Shot						Few-Shot					
	Subject		Vaccinated		Regret		Subject		Vaccinated		Regret	
	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy
mistral-small13.2:24b	0.734	0.746	0.824	0.828	0.739	0.757	0.695	0.702	0.836	0.837	0.790	0.816
gpt-4o-mini	0.672	0.674	0.842	0.843	0.803	0.855	0.647	0.650	0.839	0.839	0.804	0.853
mixtral:8x22b	0.745	0.769	0.770	0.783	0.806	0.842	0.733	0.743	0.833	0.836	0.808	0.846
llama3.1:70b	0.744	0.769	0.795	0.803	0.775	0.807	0.709	0.717	0.840	0.841	0.800	0.843
qwen2.5:7b	0.617	0.620	0.793	0.794	0.766	0.796	0.587	0.598	0.803	0.804	0.808	0.841
mistral:7b	0.683	0.702	0.763	0.771	0.767	0.795	0.655	0.665	0.813	0.813	0.783	0.816
gemma3:12b	0.736	0.760	0.773	0.783	0.641	0.648	0.743	0.763	0.796	0.803	0.687	0.698
llama3.1:8b	0.681	0.715	0.750	0.763	0.729	0.748	0.649	0.667	0.792	0.794	0.774	0.807

Table 2: Performance comparison across models in zero-shot and few-shot settings on the held-out test set

llama3.1:70b, llama3.1:8b, and mixtral:8x7b using a LoRA-based approach (Hu et al. 2021) on \mathcal{D}_{train} . This method allows us to assess whether the models can better understand and adapt to the classification task through task-specific supervision. We evaluate their performance on \mathcal{D}_{test} using standard metrics: precision, recall, F1-score, and accuracy on a held-out validation set.

Multi-Stage Hybrid Inference Pipeline: To classify a large volume of user comments with both nuance and efficiency, we designed a two-stage hybrid inference pipeline to balance computational cost and accuracy:

Stage 1: Relevance Filter: The first stage acts as a high-throughput *Relevance Filter*, using a Natural Language Inference (NLI) model from Sileo (2024), ModernBERT-large-nli. This model determines if a comment is relevant to the topic of vaccines by treating the comment as a premise and evaluating its entailment with a specific hypothesis. We tuned this stage by testing 24 combinations of different hypotheses and acceptance thresholds on our validation set. The best-performing configuration, which achieved an F1-score of 0.8680, utilized the hypothesis: “*This comment mentions or discusses anything related to vaccines, vaccination, or immunization*” with an acceptance threshold of 0.01. This initial filtering step efficiently removes a large volume of irrelevant comments, ensuring our more computationally intensive model is reserved for relevant data and reducing the likelihood of the LLM producing off-task or malformed responses.

Stage 2: Expert Reasoner: Comments that pass the relevance filter proceed to the second stage, which uses a finetuned model. This LLM performs a multi-label classification in a single pass, identifying the subject (self, other, or unspecified), vaccinated status, and regret status using a detailed prompt with specific rules and examples.

5 Results and Discussions

5.1 Classification Task

Table 2 shows that larger models generally perform better across tasks, with mixtral:8x22b achieving the highest F1 score for Regret classification in both zero-shot and few-shot settings; gpt-4o-mini leads in Vaccinated prediction (zero-shot), while surprisingly gemma3:12b performs strongly on Subject in few-shot. Few-shot setups yield

marginal improvements for most models, indicating that even minimal supervision can enhance performance on complex social classification tasks.

Table 3 shows the result for our finetuned models and it improves on zero and few shot metrics. While the mixtral:8x7b (Jiang et al. 2024a) model showed the highest performance on the Subject classification task, the llama3.1:70b model demonstrated the strongest and most consistent performance on the two tasks most central to our RQs by achieving the highest F1-score and accuracy for identifying Furthermore, in our practical testing, we observed that the Llama model had a faster inference speed than the Mixtral model. Given its performance on key tasks and its computational efficiency, we selected llama3.1:70b as the most suitable model for our final inference pipeline.

Pipeline Performance: Our two-stage hybrid inference pipeline was evaluated on the held-out test set of 400 comments from our benchmark dataset. The pipeline achieved an overall Exact Match (requiring all fields to be correct) of 62.00%. The model demonstrated strong performance in identifying vaccinated (F1(macro)=0.87) and regret (F1(macro)=0.82) status, though it faced more challenges in the three-way subject classification (81.50% accuracy & F1(macro)=0.8), particularly with the unspecified class (F1(macro)=0.67).

Pipeline Processing for \mathcal{D}_{wild} : The first stage of our pipeline, the NLI relevance filter, identified 243,547 (40.6%) of the comments as relevant to the topic of vaccines. The relevance rate was considerably higher for mainstream news sources (52.9%) than for influencer channels (28.3%).

5.2 In-the-wild results

Prevalence of Vaccine Regret Across Sources: From the pool of 243,547 relevant comments, the pipeline identified 2,727 (1.1%) as expressing regret. As shown in Figure 1, the rate of regret varied significantly across source categories. The overall rate of regret was significantly higher on influencer channels (1.9%) than on mainstream news channels (0.7%) ($\chi^2=780.29$, $p < 0.001$). Within the influencer category, the difference was even more stark: vaccine-skeptic channels exhibited a regret rate of 2.9%, a statistically significant difference from the 1.0% rate on pro-vaccine channels ($\chi^2=421.10$, $p < 0.001$). In contrast, the variations

Model	Fine-Tuned Performance						Inference Speed (sec/400 comments)
	Subject		Vaccinated		Regret		
	F1 (macro)	Accuracy	F1 (macro)	Accuracy	F1 (macro)	Accuracy	
llama3.1:70b	0.77	80.25%	0.87	87.00%	0.83	87.00%	650
llama3.1:8b	0.72	76.00%	0.86	85.75%	0.81	84.50%	164
mixtral:8x7b	0.79	81.50%	0.86	85.75%	0.83	87.25%	2,703

Table 3: Performance and inference speed comparison across fine-tuned models on the held-out test set.

in regret rates among the three mainstream news outlets were also statistically significant, though less pronounced ($\chi^2=16.23$, $p < 0.001$).

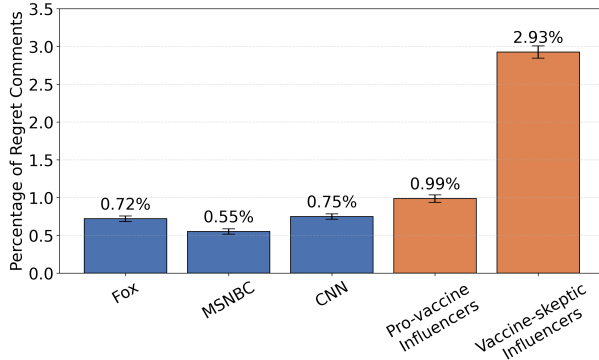


Figure 1: Percentage of regret comments across news sources and influencer categories.

Our analysis reveals that while expressions of vaccine regret are a persistent theme in online discourse, they appear in only 1.1% of relevant comments. This provides a quantitative anchor to a topic often dominated by powerful anecdotes, which are known to impact medical decisions even when presented alongside statistical data (Line et al. 2024). This suggests that while vaccine regret is a salient narrative, its actual prevalence in this discourse is far lower than its potential amplification within the broader *infodemic* might suggest.

The prevalence of regret, however, is not uniform across online communities. The rate of regretful comments on influencer channels was more than double that on mainstream news channels, with vaccine-skeptic influencers hosting a rate nearly three times higher than their pro-vaccine counterparts. This aligns with prior work on the *infodemic* (do Nascimento et al. 2022) and the significant role of online creators in shaping public opinion and health discourse (Zimmermann, Klee, and Kaspar 2023). These channels may foster echo chambers where expressions of regret are more common, normalized, and amplified.

Analysis of Narrative Perspectives: Of the 2,727 comments expressing regret, the majority (67.9%) were first-person (self) narratives. This indicates that individuals sharing their own personal stories are the primary source of regretful sentiment in these online spaces.

5.3 Substantive Findings

Analysis of Vicarious Relationships To further understand the social dynamics of vicarious regret narratives, we perform an additional classification on comments identified by the pipeline as having subject *other*.

The core categories for this task are directly informed by empirical research on the social networks that influence vaccination decisions, which identifies a hierarchy of influential relationships (Brunson 2013). During our preliminary review of the data, we also observed that a number of comments referenced the experiences of celebrities, politicians, and other well-known individuals. This type of parasocial, one-to-many influence does not fit into the interpersonal categories, so we added a Public Figure category to capture this distinct form of vicarious narrative. The final categories are: Spouse or Partner, Family Member, Friend, Health Care Provider, Public Figure, Other Acquaintance, and Unspecified.

This classification is performed using a zero-shot prompting approach with the Llama-3.1:70B-Instruct (Llama Team, AI @ Meta 2024) model and is validated by manually verifying the model’s output on a random sample of comments. We ran this on the author’s relationship to the subject in all 875 comments in \mathcal{D}_{wild} identified as “*other*”. This approach was first validated on a manually annotated set, where it achieved 90.71% accuracy.

The analysis showed that *Family Member* was the most frequently cited relationship (29.6%), followed by *Unspecified* (28.5%) (see Figure 2 for full distribution). This underscores the role of intimate social networks in the dissemination of health narratives, as established by Brunson (2013).

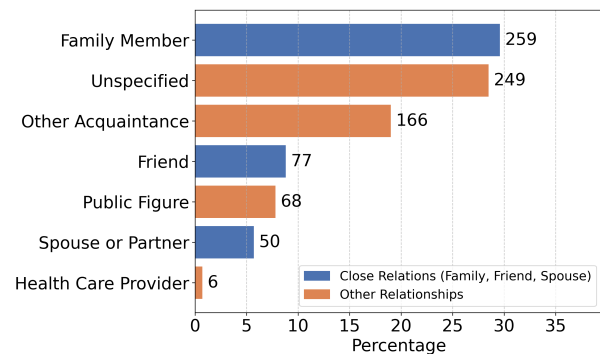


Figure 2: Overall Distribution of Relationships in Vicarious Regret Comments

Analysis of Regret Reasons: To identify the primary themes driving vaccine regret, we developed a set of categories informed by existing literature and our own data. The categories *Adverse Health Event*, *Perceived Coercion*, and *Shift in Beliefs* are directly supported by prior qualitative research (Tayhan, Tayhan, and Büyük 2025; Luo et al. 2022). In addition, we included the *Lack of Efficacy* category after observing a prevalent theme of individuals expressing regret because they believed the vaccine did not work as promised (e.g., they still contracted COVID-19). We then employed a zero-shot prompting approach with the Llama-3.1:70B-Instruct model to perform an information extraction task, categorizing each regretful comment accordingly. This model was first validated on our benchmark test set, where it achieved 92.08% accuracy in extracting the correct reason

Applying this validated method to the 2,727 comments expressing regret on \mathcal{D}_{wild} , our analysis revealed that an *Adverse Health Event* was the most common reason overall, accounting for 55.0% of cases. However, the distribution of reasons differed significantly between news and influencer channels ($\chi^2=248.84$, $p < 0.001$). As shown in Figure 3, influencer channels were dominated by discussions of adverse health events (64.2%). In contrast, news channels featured a more balanced conversation where *Lack of Efficacy* was a prominent secondary reason (26.9%).

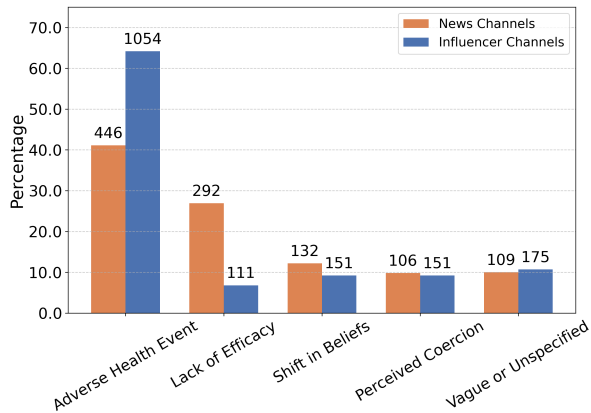


Figure 3: Distribution of Regret Reasons by Source Type.

We also observed a significant difference in the reasons cited between first-person and vicarious accounts ($\chi^2=45.94$, $p < 0.001$). While *Adverse Health Event* was the most common reason for both groups, it was more dominant in vicarious narratives (61.3%) than in first-person ones (52.1%). Conversely, *Perceived Coercion* was more than twice as likely to be cited as a reason for regret in first-person accounts (11.6%) compared to vicarious accounts (4.9%). This supports previous research identifying perceived coercion as a key driver of regret (Tayhan, Tayhan, and Büyük 2025) and highlights that the feeling of diminished autonomy is a powerful and personal component of this sentiment

Perceptual Differences in Annotation: We observed moderate inter-rater agreement among our politically di-

verse annotators across all three annotation tasks, as detailed in Table 4. Overall agreement (Fleiss’ Kappa) was highest for the Subject task ($\kappa = 0.5089$) and slightly lower for Regret ($\kappa = 0.4480$) and Vaccinated ($\kappa = 0.4272$). Our observed agreement aligns with prior literature (Weerasooriya et al. 2023; Crowl et al. 2025).

A Chi-square test revealed that political affiliation significantly influenced the subjective task of classifying a comment’s *Subject* ($\chi^2=22.85$, $p < 0.001$). Conversely, we found no systematic impact on the more factual *Vaccinated* or sentiment-based *Regret* judgments, as detailed in Table 5.

Task	Fleiss’ κ	Dem vs Rep κ	Dem vs Ind κ	Rep vs Ind κ
Subject	0.5089	0.5093	0.5140	0.5046
Vaccinated	0.4272	0.4380	0.4286	0.4330
Regret	0.4480	0.4246	0.4967	0.3677

Table 4: Inter-Annotator Agreement by Task

Task	Chi-square (χ^2)	p-value
Subject	22.8515	0.0001
Vaccinated	1.6875	0.4301
Regret	3.1468	0.2073

Table 5: Systematic Differences in Annotation by Political Affiliation

Model Alignment with Annotator Politics: On the subset of 398 comments where annotators disagreed on the Regret label (always in a 2-vs-1 split), we analyzed our pipeline’s predictions to check for political alignment. The model’s final output sided with the majority opinion 70.1% of the time. The model’s alignment with Democratic (55.8%), Republican (53.3%), and Independent (61.1%) annotators was not statistically different ($\chi^2 = 2.22$, $p = 0.3298$). This suggests that for this task, the model does not systematically favor the perspective of one political group over the others.

Temporal Study of Regret: Figure 4 shows the temporal distribution of vaccine-related regret comments expressed by pro-vaccine and vaccine-skeptic influencers from 2020 through mid-2024, overlaid with key U.S. COVID vaccine policy phases. The graph is divided into four zones: Zone A (EUA & Rollout) marks the earliest phase of public awareness, characterized by high uncertainty and the emergence of early regret commentary; Zone B (Mandate Peak) corresponds to the introduction of federal employee and contractor mandates, a period associated with heightened polarization and visible spikes in regret, particularly among vaccine-skeptic influencers; Zone C (Rollback) captures the easing of legal enforcement and restrictions, potentially correlating with a decline in overt regret discourse; and Zone D (Post-Mandate Era) reflects a stabilized regulatory environment where vaccine sentiment is increasingly shaped by individual and localized narratives. Throughout these phases, vaccine-skeptic influencers exhibit a rising trajectory of regret commentary, peaking dramatically by the end, while

pro-vaccine influencers maintain comparatively lower and more sporadic levels of regret expression. While on the other end we do not see any noticeable differences in trend for the news channels.

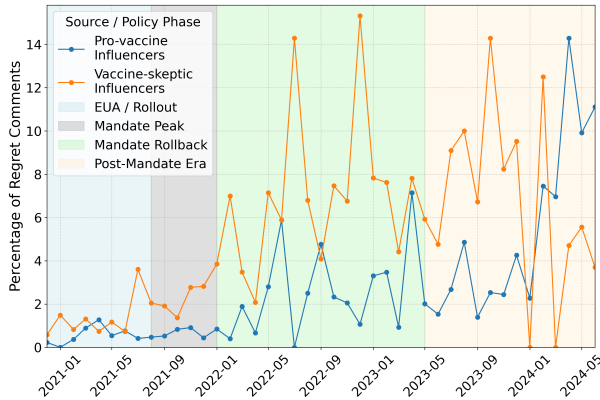


Figure 4: Temporal Distribution of Regret Comments over time across zones

Qualitative Error Analysis: To better understand the model’s performance, we conducted a qualitative analysis of misclassified examples in \mathcal{D}_{test} . This revealed that the lower F1-score for subject classification is an expected artifact of our pipeline’s design, which prioritizes efficiency. The initial NLI relevance filter effectively discards large volumes of irrelevant comments, which substantially improves processing speed and the stability of the downstream LLM. This efficiency creates an occasional mismatch with human annotations on irrelevant comments, artificially lowering the performance metric for the *unspecified* subject class. This is an accepted trade-off, however, as it does not impact the accuracy of the primary regret classification task.

5.4 Limitations

While our study offers significant insights to public’s vaccine stance, it has some limitations that must be considered when interpreting the results.

First, our dataset of YouTube comments, like the official Vaccine Adverse Event Reporting System (VAERS) (Shimabukuro et al. 2015) rely on spontaneous, voluntary reports subject to significant self-selection bias; therefore, our findings cannot be used to infer causality or calculate the true incidence rate of vaccine regret or adverse events in the general population. Second, while our pipeline performed well, the classification of a nuanced human emotion like regret is inherently challenging. Sarcasm and complex expressions can still be misinterpreted by advanced LLMs (Crowl et al. 2025), and the F1-score of 0.83 for the regret class indicates some degree of model error is unavoidable. Finally, our analysis is limited to English-language comments on a single platform (YouTube). The dynamics of post-vaccination sentiment may differ significantly in other languages, cultural contexts, and on other social media platforms.

5.5 Implications and Future Work

Our research has several implications. For public health officials, our work provides a scalable method for monitoring public sentiment and identifying the primary drivers of vaccine regret. Understanding that narratives of coercion, adverse events, and lack of efficacy are central to this sentiment can help refine public health communication and improve Proactively addressing public concerns about adverse events and communicating the rationale behind public health mandates are also essential to maintain institutional trust (Souvatzi et al. 2024). The stark differences between mainstream news and influencer communities underscore the need for tailored outreach strategies that address the specific concerns circulating in different online ecosystems.

Future research directions include expanding this analysis to other social media platforms, languages, and cultural contexts to create a more holistic picture of global post-vaccination sentiment. A longitudinal study tracking how these regret narratives evolve over time, particularly in response to new public health developments, would be highly valuable. While our human annotation pipeline accounts for political diversity, extending the annotation pipeline to vicarious interactions (Weerasooriya et al. 2023; Dutta et al. 2025a) can help us understand how well raters can represent out-group values and opinions on politically sensitive topics. Finally, refining the classification of regret reasons, for instance by distinguishing between mild and severe adverse events, could provide even more granular insights for public health intervention.

6 Conclusion

In this study, we introduced a novel computational framework to quantify and analyze the phenomenon of *vaccine buyer’s remorse* within a large corpus of YouTube comments. Our findings reveal that while expressions of regret are a persistent feature of online discourse, they represent a small fraction of the overall conversation. This sentiment is most prevalent on influencer channels and is primarily driven by narratives of adverse health events, perceived lack of efficacy, and feelings of coercion. By differentiating between first-person and vicarious reports and identifying primary drivers of regret, such as adverse health events and perceived lack of efficacy, this study provides key insights into post-vaccination attitudes. In the end, this approach provides a scalable method for public health monitoring and highlights the need for targeted communication efforts that address specific public worries to build trust in a growingly divided and polarized information environment.

Acknowledgments

Miles Stanley and Ashiqur R. KhudaBukhsh were partially supported by NSF Award #2447631.

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

- Becerra-Perez, M.-M.; Menear, M.; Turcotte, S.; Labrecque, M.; and Légaré, F. 2016. More primary care patients regret health decisions if they experienced decisional conflict in the consultation: a secondary analysis of a multicenter descriptive study. *BMC Family Practice*, 17(1): 156.
- Brehaut, J. C.; O'Connor, A. M.; Wood, T. J.; Hack, T. F.; Siminoff, L.; Gordon, E.; and Feldman-Stewart, D. 2003. Validation of a decision regret scale. *Medical decision making*, 23(4): 281–292.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.
- Brunson, E. K. 2013. The Impact of Social Networks on Parents' Vaccination Decisions. *Pediatrics*, 131(5): e1397–e1404.
- Clemens, K. S.; Faasse, K.; Tan, W.; Colagiuri, B.; Colloca, L.; Webster, R.; Vase, L.; Jason, E.; and Geers, A. L. 2023. Social communication pathways to COVID-19 vaccine side-effect expectations and experience. *Journal of Psychosomatic Research*, 164: 111081.
- Crowl, L.; Dutta, S.; KhudaBukhsh, A. R.; Severnini, E.; and Nagin, D. S. 2025. Measuring criticism of the police in the local news media using large language models. *Proceedings of the National Academy of Sciences*, 122(9): e2418821122.
- Davidson, T.; Warmesley, D.; Macy, M.; and Weber, I. 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, 512–515.
- do Nascimento, I. J. B.; Pizarro, A. B.; Almeida, J. M.; Azzopardi-Muscat, N.; Gonçalves, M. A.; Björklund, M.; and Novillo-Ortiz, D. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(8): 544–561.
- Dolman, A. J.; Fraser, T.; Panagopoulos, C.; Aldrich, D. P.; and Kim, D. 2023. Opposing views: associations of political polarization, political party affiliation, and social trust with COVID-19 vaccination intent and receipt. *Journal of Public Health*, 45(1): 36–39.
- Dutta, A.; Ali, S. M. S.; Naseem, U.; and KhudaBukhsh, A. R. 2025a. Towards a Bipartisan Understanding of Peace and Vicarious Interactions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025*, 9628–9637. ijcai.org.
- Dutta, S.; Pandita, D.; Weerasooriya, T. C.; Zampieri, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2025b. ARTICLE: Annotator Reliability Through In-Context Learning. In *AAAI-25, Association for the Advancement of Artificial Intelligence*, 14230–14237. AAAI Press.
- Gao, L.; and Huang, R. 2017. Detecting Online Hate Speech Using Context Aware Models. In Mitkov, R.; and Angelova, G., eds., *RANLP 2017*, 260–266. INCOMA Ltd.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- He, L.; Omranian, S.; McRoy, S.; and Zheng, K. 2024. Using Large Language Models for sentiment analysis of health-related social media data: empirical evaluation and practical tips. *Journal of the American Medical Informatics Association*. Working paper/Preprint.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv 2021. arXiv preprint arXiv:2106.09685*, 10.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Islam, T.; and Goldwasser, D. 2022. Understanding covid-19 vaccine campaign on Facebook using minimal supervision. In *2022 IEEE-Big Data*, 585–595. IEEE.
- Islam, T.; and Goldwasser, D. 2025a. Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating LLMs. In *ICWSM*, volume 19, 859–884.
- Islam, T.; and Goldwasser, D. 2025b. Uncovering latent arguments in social media messaging by employing llms-in-the-loop strategy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 7397–7429.
- Jain, D.; Rai, S.; Mittal, J.; Andy, A.; Buttenheim, A. M.; and Guntuku, S. C. 2025. Twitter reveals spatio-temporal variation in vaccine concerns in Sub-Saharan Africa. *medRxiv*, 2025–08.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024b. Mixtral of Experts. *ArXiv*, abs/2401.04088.
- KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14893–14901.
- KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2022. Fringe News Networks: Dynamics of US News Viewership following the 2020 Presidential Election. In *WebSci '22: 14th ACM Web Science Conference 2022*, 269–278. ACM.
- Li, Y.; Viswaroopan, D.; He, W.; Li, J.; Zuo, X.; Xu, H.; and Tao, C. 2025. Enhancing Relation Extraction for COVID-19

- Vaccine Shot-Adverse Event Associations with Large Language Models. *Research Square*. Preprint.
- Line, E. N.; Jaramillo, S.; Goldwater, M.; and Horne, Z. 2024. Anecdotes impact medical decisions even when presented with statistical information or decision aids. *Cognitive Research: Principles and Implications*, 9(1): 51.
- Liu, Y.; Wang, Y.; Sun, A.; Meng, X.; Li, J.; and Guo, J. 2022. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1670–1680.
- Llama Team, AI @ Meta. 2024. The Llama 3 Herd of Models. Technical report, Meta. Technical Report.
- Luo, C.; Jiang, W.; Chen, H.-X.; and Tung, T.-H. 2022. Post-vaccination adverse reactions, decision regret, and willingness to pay for the booster dose of COVID-19 vaccine among healthcare workers: A mediation analysis. *Human Vaccines & Immunotherapeutics*, 18(6): e2146964.
- MistralAI. 2025. Mistral Small 3.2 24B Instruct (2506). <https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506>.
- Mittal, S.; Chawla, T.; and KhudaBukhsh, A. R. 2024. You Must Be a Trump Supporter: Political Identity Projections on the Social Web. In *Social Networks Analysis and Mining, ASONAM 2024*, 391–404. Springer.
- Pacheco, M. L.; Islam, T.; Mahajan, M.; Shor, A.; Yin, M.; Ungar, L.; and Goldwasser, D. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. *arXiv preprint arXiv:2205.01817*.
- Pacheco, M. L.; Islam, T.; Ungar, L.; Yin, M.; and Goldwasser, D. 2023. Interactive concept learning for uncovering latent themes in large text collections. *arXiv preprint arXiv:2305.05094*.
- Pandita, D.; Weerasooriya, T. C.; Dutta, S.; Luger, S.; Ranasinghe, T.; KhudaBukhsh, A. R.; Zampieri, M.; and Homan, C. 2024. Rater Cohesion and Quality from a Vicarious Perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5149–5162.
- Pofcher, J.; Homan, C. M.; Sell, R.; and KhudaBukhsh, A. R. 2025. Hope vs. Hate: Understanding User Interactions with LGBTQ+ News Content in Mainstream US News Media through the Lens of Hope Speech. In *EMNLP 2025*, 19873–19899.
- Portelli, B.; Scaboro, S.; Tonino, R.; Chersoni, E.; Santus, E.; and Serra, G. 2022. Monitoring User Opinions and Side Effects on COVID-19 Vaccines in the Twittersphere: Infodemiology Study of Tweets. *Journal of Medical Internet Research*, 24(5): e35115.
- Sehgal, N. K.; Rai, S.; Tonneau, M.; Agarwal, A. K.; Cappella, J.; Kornides, M.; Ungar, L.; Bittenheim, A.; and Guntuku, S. C. 2025. Conversations with AI Chatbots Increase Short-Term Vaccine Intentions But Do Not Outperform Standard Public Health Messaging. *arXiv preprint arXiv:2504.20519*.
- Shimabukuro, T. T.; Nguyen, M.; Martin, D.; and DeStefano, F. 2015. Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). *Vaccine*, 33(36): 4398–4405.
- Sileo, D. 2024. tasksource: A Large Collection of NLP tasks with a Structured Dataset Preprocessing Framework. In *LREC-COLING 2024*, 15655–15684.
- Souvatzi, E.; Katsikidou, M.; Arvaniti, A.; Plakias, S.; Tsiakiri, A.; and Samakouri, M. 2024. Trust in Healthcare, Medical Mistrust, and Health Outcomes in Times of Health Crisis: A Narrative Review. *Societies*, 14(12): 269.
- Tayhan, A.; Tayhan, E. B.; and Büyük, D. Ş. 2025. Nursing and Midwifery Students' COVID-19 Vaccine Regrets and Future Vaccination Intentions: A Mixed Methods Study. *Nursing & Health Sciences*, 27: e70039.
- Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Wawrzuta, D.; Jaworski, M.; Gotlib, J.; and Panczyk, M. 2021. What arguments against COVID-19 vaccines run on Facebook in Poland: content analysis of comments. *Vaccines*, 9(5): 481.
- Weerasooriya, T. C.; Dutta, S.; Ranasinghe, T.; Zampieri, M.; Homan, C.; and KhudaBukhsh, A. R. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What Is Offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, 11648–11668.
- Wiegand, M.; Ruppenhofer, J.; and Kleinbauer, T. 2019. Detection of abusive language: the problem of biased datasets. In *NAACL-HLT*, 602–608.
- Yin, L.; Han, M.; and Nie, X. 2024. Unlocking Blended Emotions and Underlying Drivers: A Deep Dive into COVID-19 Vaccination Insights on Twitter Across Digital and Physical Realms in New York, Using ChatGPT. *Urban Science*, 8(4): 222.
- Yoo, C. H.; and KhudaBukhsh, A. R. 2023. Auditing and robustifying COVID-19 misinformation datasets via anticontent sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15260–15268.
- Zeelenberg, M.; and Beattie, J. 1997. Consequences of Regret Aversion 2: Additional Evidence for Effects of Feedback on Decision Making. *Organizational Behavior and Human Decision Processes*, 72(1): 63–78.
- Zhou, J.; Zhang, L.; Li, M.; Horne, B. D.; and De Choudhury, M. 2025. AI as We Describe It: How Large Language Models and Their Applications in Health are Represented Across Channels of Public Discourse. *arXiv preprint arXiv:2511.03174*.
- Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Comput. Linguistics*, 50(1): 237–291.
- Zimmermann, D.; Klee, A.; and Kaspar, K. 2023. Political news on Instagram: influencer versus traditional magazine and the role of their expertise in consumers' credibility perceptions and news engagement. *Frontiers in Psychology*, 14: 1257994.