Human Inputs, Model Outputs: Dual Sources of Variability in Stance Mining on Gun Rights/Control

Niranjana Sathish Avilery*, **Soumyajit Datta***, **Ashini Anantharaman**, **Ashiqur R. KhudaBukhsh**
Rochester Institute of Technology
na6322@rit.edu, sd3528@rit.edu, aa9162@rit.edu, axkvse@rit.edu

Abstract

Large language models (LLMs) are increasingly used in computational social science, both to analyze human-generated web content and to act as human surrogates capable of answering substantive research questions. As these uses expand, a critical methodological question arises: how stable and reliable are the answers LLMs produce? We examine this question through the lens of stance mining on gun control and gun rights, a mature and widely used NLP task applied to one of the most extensively studied and polarized political issues in the United States. Using a multi-platform dataset spanning diverse online communities, we evaluate the stability of LLM-predicted stance across model families, prompt designs, and conversational contexts. Despite strong expectations of robustness in this setting, we observe substantial variability in predicted stance across LLMs and prompts. To contextualize these findings, we compare model-induced variability with human variability by analyzing annotations from raters with differing political affiliations. We find that political priors and annotator disagreement persist even in this ostensibly straightforward classification task. As a useful resource, we release a stance dataset annotated by raters with diverse political affiliations. Our results show that outcomes in stance mining depend jointly on which LLM is used and which humans provide the inputs, underscoring dual sources of variability that have important implications for using LLMs as research instruments in computational social science.

1 Introduction

Large language models (LLMs) are rapidly transforming the methodological landscape of computational social science (Ziems et al., 2024) (CSS). While early work focused on automating analyses of human-generated web content, recent efforts increasingly treat LLMs themselves as human surrogates, capable of answering sophisticated research questions or simulating diverse sociopolitical perspectives (Argyle et al., 2023; Horton, 2023; Aher et al., 2023). As these capabilities expand, a central question becomes unavoidable: *How stable or reliable are the answers LLMs provide?* The promise of scalable, low-cost inference hinges on the assumption that model outputs are consistent across platforms, prompts, and model families. Yet emerging evidence suggests that LLM predictions can vary substantially (Baumann et al., 2025), raising concerns about their suitability for empirical research, policy analysis, and theory building.

To investigate this issue, we consider a setting where we might reasonably expect stability: stance mining on gun control and gun rights. Stance mining is among the most widely used NLP tasks (Mohammad et al., 2016) in computational social science, supporting research on political polarization, public opinion, civic discourse, misinformation, and agenda-setting (Darwish, 2019; Dutta et al., 2022; Conover et al., 2011; Darwish et al., 2020; Burnham, 2025). Because the task requires models to infer latent positions embedded in natural language—often marked by ambiguity, culturally contingent cues, or subtle framing—it provides a rigorous test of how well machine outputs align with human interpretations. Its methodological maturity and well-established evaluation conventions make stance mining an ideal lens for studying variability in LLM behavior.

Within this task, we focus on gun control and gun rights, one of the most enduring and normatively charged political debates in the United States. The issue is sharply polarized (Weerasooriya et al., 2023a), empirically rich, and extensively studied (Gyawali et al., 2024), providing a strong foundation for evaluating whether LLMs capture stable indicators of political position or whether their

*Niranjana Sathish Avilery and Soumyajit Datta are equal-contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33629–33642
October 24-29, 2026 ©2026 Association for Computational Linguistics

predictions drift with changes in prompt structure, model architecture, or conversational context. Prior computational social science research has used this domain to investigate affective polarization, identity signaling, framing, moral reasoning, and cross-platform opinion dynamics, offering a robust interpretive scaffold for our analysis.

Using a multi-platform dataset that spans diverse online communities, we assess variability in LLM-predicted stance on gun rights and control. Contrary to expectations for a well-understood task and a richly characterized issue domain, we observe considerable variability across models and prompts. To contextualize these findings, we contrast model variability with human variability—an equally important, but often under-examined, component of modern computational pipelines. Here, too, we find systematic differences: political leanings, interpretive priors, and annotator disagreement persist even in this ostensibly straightforward classification task.

Taken together, these patterns underscore a broader challenge for CSS: even in a canonical, highly studied NLP task, the outcome depends both on which LLM is used and on which humans provide the inputs. Our analysis shows that human-machine variability is not merely noise but a structural feature of how modern AI systems interact with human judgment. Understanding this dual source of variability is essential for any rigorous use of LLMs as research instruments or as proxies for human reasoning in sociotechnical studies.

Contributions: We make four contributions. **(1)** We empirically evaluate the stability of LLM-predicted stance in the context of gun control and gun rights, revealing substantial sensitivity to model choice and prompt design even for a mature NLP task. **(2)** We explicitly compare LLM-induced variability with human annotator variability, showing that disagreement and political priors persist alongside model-driven variation. **(3)** Through a multi-platform analysis, we demonstrate that observed opinion dynamics reflect the interaction of human and machine variability, raising methodological considerations for deploying LLMs in CSS. **(4)** Finally, our work advocates political representation and inclusion as we release a publicly available stance dataset comprising 1,560 instances[1]. with balanced political representation, each annotated by three raters—one Republican, one Democrat, and one Independent—to support future research on stance, disagreement, and model–human comparisons.

[1]Dataset can be found here: https://github.com/Social-Insights-Lab/gun_rights_Human-Inputs-Model-Outputs

2 Related Work

Our work contributes to a growing literature on rater subjectivity in NLP annotation (Prelec, 2004; Pavlick and Kwiatkowski, 2019; Dumitrache et al., 2018; Poesio et al., 2019; Liu et al., 2019b; Weerasooriya et al., 2020; Basile, 2020; Weerasooriya et al., 2023b; Sap et al., 2022; Weerasooriya et al., 2023a; Pofcher et al., 2025). Prior studies have examined how political orientation influences annotation outcomes in a range of contexts, including offensive speech detection (Sap et al., 2022; Weerasooriya et al., 2023a), identification of supportive content for the LGBTQ+ community (Pofcher et al., 2025), and stance mining related to policing in the United States (Crowl et al., 2025). While annotator subjectivity (Sap et al., 2022; Weerasooriya et al., 2023a) and uneven model representation of political views (Santurkar et al., 2023) have each been documented before, our contribution is to show that these two sources interact and together shape the substantive conclusion.

Crowl et al., 2025 report that their substantive findings are robust to annotator political affiliation, observing similar conclusions when models are trained exclusively on annotations provided by Republicans, Democrats, or Independents. We instead observe that substantive outcomes can vary as a function of rater political identity, alongside other experimental design choices such as model selection and prompting strategy. Similarly, while prompt sensitivity in CSS is well documented (Atreja et al., 2024), our study shows how this sensitivity affects the substantive conclusion, the stance ratio, in a mature and well-studied task. Together, these differences position our work as complementary to existing studies, while emphasizing the need to account for rater heterogeneity when drawing substantive conclusions from politically sensitive annotation tasks.

Computational analyses of gun-related discourse on the social web constitute a substantial body of prior work, spanning a range of platforms including Twitter (Gyawali et al., 2024; Benton et al., 2016; Zhang et al., 2019; Demszky et al., 2019) and Reddit (Magu et al., 2024). We draw on this literature to construct a lexicon of commonly used

keywords (e.g., *gun*, *shoot*, *kill*, *victim*, *shooter*, *firearm*, *Second Amendment*, and *NRA*) to identify and filter relevant content (Demszky et al., 2019; Benton et al., 2016; Liu et al., 2019a; Zhang et al., 2019). While much of this prior work focuses on answering substantive sociopolitical questions, for example, examining how a shooter's race shapes online discourse following mass shootings (Demszky et al., 2019), our study shifts attention to a complementary concern: how human annotator characteristics and LLM design choices jointly influence the substantive conclusions drawn from such analyses.

Our work is also situated within the extensive literature on stance mining (Küçük and Can, 2020; Alturayeif et al., 2023; Zhang et al., 2024), which has explored framing stance detection as natural language inference (Dutta et al., 2022), examined how annotation context shapes label quality (Joseph et al., 2017), and introduced domain-specific resources focused on gun control and gun rights discourse (Gyawali et al., 2024). We build on the task formulation of Gyawali et al., 2024, analyzing multiple social web platforms and explicitly incorporating annotators' political affiliations, allowing us to examine how political heterogeneity interacts with stance labeling and downstream model behavior.

As LLMs are increasingly adopted as "in silico" annotators, critical concerns regarding their reliability have emerged. Baumann et al., 2025 quantified the risks of "LLM Hacking," where arbitrary choices in model selection or prompting lead to incorrect scientific conclusions. Further complicating reliability is the tension between a model's parametric knowledge and user instructions; Mohammadi et al., 2025 found that LLMs' integration of explicit definitions is inconsistent, with models defaulting to internal representations particularly on general tasks. Our work adds to this important conversation.

3 Data

3.1 Reddit Data Collection

We collect data from 49 subreddits likely to contain considerable gun control/rights discourse. These included 42 gun-specific communities (e.g., r/guncontrol, r/progun) and 7 general political forums (e.g., r/politics, r/news) filtered for gun-related keywords. In addition to these subreddits, we include event-related data on mass shootings from various subreddits, such as Sandy Hook Massacre, Parkland Shooting, Robb Shooting, Las Vegas Shooting, Vegas Shooting Research, Highland Park Shooting, and Uvalde Texas Shooting. Since the debate surrounding gun control/rights often show sharp differences along party lines, we also include relevant posts from political subreddits such as r/liberal, r/conservative, r/politics, and r/republican. To ensure gun-specific discourse, we consider posts that contain specific keywords related to gun control, gun rights, and mass shootings. Finally, to capture the full context of the discussion, we extracted data hierarchically: obtaining the original submission (Post), top-level comments, and nested replies. The raw Reddit dataset comprised 1,273,515 data points, consisting of 39,831 posts, 718,500 comments, and 515,184 replies.

3.2 Link Extraction and YouTube Retrieval

From the Reddit corpus, we parsed 47,180 hyperlinks to identify external media sources. YouTube was the second most frequent domain shared (4,897 links), following internal Reddit links of 5,130 links. Using the Google API, we resolved these links to retrieve metadata (title, description, tags) for 3,879 unique videos. As not all linked videos remained relevant to the specific topic of gun violence (e.g., gaming videos or general news), we applied a relevance filter. We employed `gpt-4o-mini` to classify video metadata as either "Relevant" or "Irrelevant" to the gun debate. This process yielded a curated set of 1,108 unique videos containing 2,874,955 comments (collected through publicly available YouTube API). For the experiments detailed in this paper, we utilized a stratified random sample of 10,054 YouTube comments from this filtered dataset ensuring nearly equal representation of videos.

3.3 Linked corpora $\mathcal{D}_{Reddit}$ and $\mathcal{D}_{YouTube}$

Our described steps yield two linked corpora: $\mathcal{D}_{Reddit}$ consisting of 1,116 Reddit posts where a YouTube link relevant to gun-related discourse is shared; and $\mathcal{D}_{YouTube}$ a sample of 10,054 comments from these shared YouTube videos. We obtain $\mathcal{D}_{Reddit}$ through backtracking the relevant YouTube videos to their originating points on Reddit, identifying the exact posts, comments, or replies that contained the external YouTube links. This process yielded 602 unique source records. To capture the immediate conversational context surrounding these links, we expanded the dataset: for top-level

posts, we included their respective sub-comments and replies; for comments, we included their nested sub-replies, resulting in a total corpus of 1,116 Reddit items.

4 Design Considerations

Stance classification taxonomy. We adopt the following stance classification taxonomy, adapted from prior work on gun control and gun rights discourse (Gyawali et al., 2024) and stance mining frameworks for other sociopolitical issues (Pofcher et al., 2025).

- **Gun Control (GC):** The text advocates for stricter regulations, legislation, or oversight of firearms, or expresses criticism of gun culture.
- **Gun Rights (GR):** The text advocates for the right to own firearms, often citing self-defense, the Second Amendment, or opposition to government overreach.
- **Neither:** The text is on-topic regarding guns but expresses a neutral stance, asks a factual question, or discusses a specific incident without advocating for policy change.
- **Irrelevant:** The text is off-topic, spam, or noise (e.g., unrelated personal attacks).

Open weight models. We consider three well-known open models across both platforms: `Mistral` (Mistral-7B-v0.3) [2]; `Llama` (Meta-Llama-3-8B) [3]; and `GPT-OSS` (gpt-oss-20b) (OpenAI, 2025). To check whether our findings hold at larger scale, we additionally evaluate two larger members of the Mistral and Llama families on Reddit, `Mistral-Medium` (Mistral-Medium-3.5) [4]; and `Llama-70B` (Llama-3.1-70B) [5]; these results are reported in Appendix E

Zero-shot classification prompts. We evaluate three prompt settings that vary in the amount and specificity of task guidance provided to the model.

1. **Vanilla (Zero-Shot):** The model is asked to classify the comment based on the label names alone, relying on its internal pre-trained representations.
2. **Definition 1 (Standard):** The prompt includes detailed definitions for the categories.
3. **Definition 2 (Refined):** A stricter variant explicitly designed to disambiguate between "Neither" (on-topic neutral) and "Irrelevant" (off-topic).

Recent literature has investigated the receptivity of LLMs to external label definitions (Mohammadi et al., 2025). We consider two nuanced variants (def_1 and def_2) beyond the vanilla prompt. In simple words, def_1 lets the model read between the lines by paying attention to tone and implied meaning, so it can catch subtle or indirect opinions about guns. Def_2 is more strict and literal, only labeling a stance when it is clearly stated, which makes the results more consistent and easier to reproduce. Appendix B contains prompt details.

Substantive measure. Datasets and resources related to gun control and gun rights support a wide range of substantive research questions. We focus on a simple, interpretable measure that captures the relative presence of these two viewpoints within a given slice of data. For a dataset $\mathcal{D}$, let $\mathcal{N}_{GC}$ and $\mathcal{N}_{GR}$ denote the number of instances expressing gun control (GC) and gun rights (GR) stances, respectively. We quantify relative prevalence using the ratio $\frac{\mathcal{N}_{GC}}{\mathcal{N}_{GR}}$. Values greater than one indicate a higher prevalence of gun-control–oriented content, whereas values less than one indicate a higher prevalence of gun-rights–oriented content. Statistical methods such as design-based supervised learning (Egami et al., 2023) and prediction-powered inference (Angelopoulos et al., 2023) estimate such a quantity from imperfect machine labels by combining them with a smaller set of human labels to correct the bias. In our setting the human labels themselves vary with annotator political identity, so rather than reducing this to one estimate, we examine how the ratio changes across models, prompts, and annotator groups.

5 Noise Audit of Zero-shot Classifiers

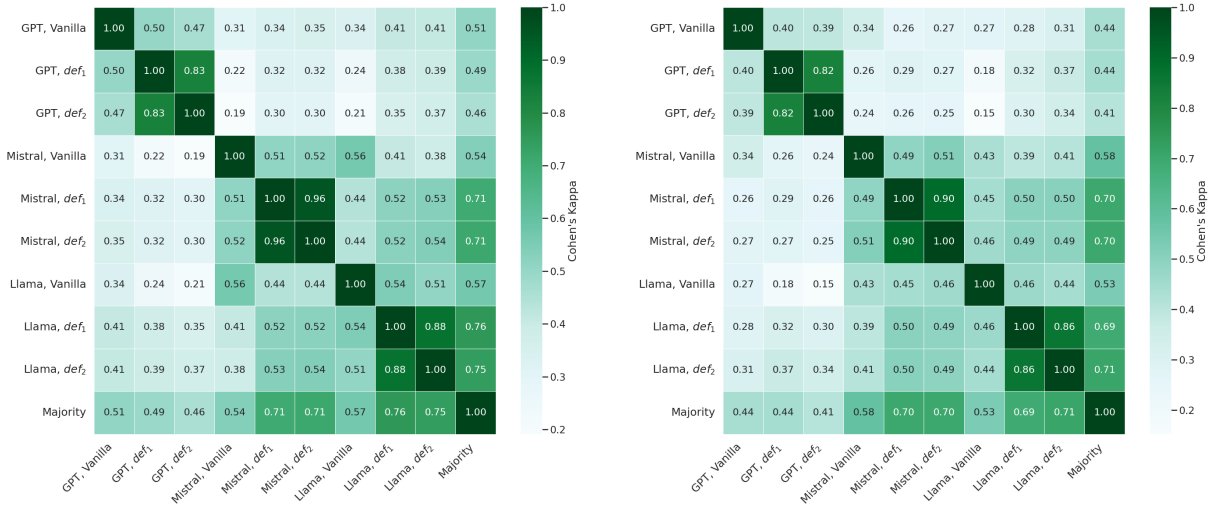
On both Reddit and YouTube we use three models, each with three prompting strategies, giving nine $\langle model, prompt \rangle$ configurations per platform. For each model configuration, we ran inference on $\mathcal{D}_{Youtube}$ and $\mathcal{D}_{Reddit}$. In the absence of established ground truth, behavioral economics literature suggests *noise audit* as a viable path to study outcome variability of multiple competent decision systems (Kahneman et al., 2021). In the NLP literature, the concept of *noise audit* has been extended in the contexts of offensive language (Weerasooriya

[2] https://huggingface.co/mistralai/Mistral-7B-v0.3

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[4] https://huggingface.co/mistralai/Mistral-Medium-3.5-128B

[5] https://huggingface.co/meta-llama/Llama-3.1-70B



(a) Reddit

(b) YouTube Noise Audit

Figure 1: Agreement between $\langle \mathcal{M}, prompt \rangle$ pairs. A cell $\langle i, j \rangle$ presents the Cohen's κ agreement between $\mathcal{M}_i$ and $\mathcal{M}_j$ where $\mathcal{M}_i$ and $\mathcal{M}_j$ are defined by an LLM and a prompt. `Majority` takes the majority vote of the nine individual $\langle \mathcal{M}, prompt \rangle$ choices we considered.

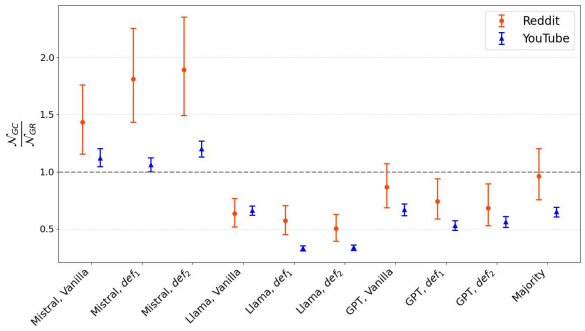


Figure 2: Stance ratio (Gun Control / Gun Rights) produced by different ⟨model, prompt⟩ configurations across Reddit and YouTube. Ratios above the horizontal dashed line (> 1) indicate a greater prevalence of gun control–favoring discourse, while ratios below the line (< 1) indicate a greater prevalence of gun rights–favoring discourse. 95% confidence intervals are estimated via bootstrap sampling.

et al., 2023a) and commonsense reasoning (Kejriwal et al., 2024). We first conduct a noise audit of these configurations to investigate outcome variability. In addition, we also consider a Majority model which assigns the most prevalent label across all configurations on a given platform. For every comment, we calculated a *Majority* based on a majority vote across these configurations:

$$C_{Majority} = \text{argmax}_{c \in C} \sum_{i=1}^{N} \mathbb{I}(M_i(x) = c) \quad (1)$$

where $c \in \{GC, GR, Neither, Irrelevant\}$ and $N = 9$ is the number of configurations per platform.

Figure 1 summarizes our noise audit findings. Each cell $\langle i, j \rangle$ represents Cohen's κ agreement between model configurations $\mathcal{M}_i$ and $\mathcal{M}_j$ on the dataset $\mathcal{D}$. We highlight three key observations. First, holding the model fixed, variability across prompting strategies is comparatively lower. However, despite gun control and gun rights being well-established concepts, vanilla prompting exhibits substantially lower agreement with prompts that include nuanced definitions. Second, for a fixed model, prompts that incorporate explicit and refined definitions yield relatively consistent inference behavior. Taken together, as the technical barriers to deploying LLMs continue to fall, our findings suggest that the absence of precise prompt definitions—even for ostensibly familiar concepts—can lead to unreliable and unstable model inferences. Finally, we also notice that the aforementioned findings are qualitatively consistent across Reddit and YouTube, the platforms we investigate. We also investigated how different factors (e.g., context length) affect model agreement. Our secondary findings indicate that disagreements are often a function of text length—both models and humans tend to disagree less on longer comments where more context is available. Figures 7 and 8 in the Appendix illustrates this cumulative agreement trend.

The noise audit reveals considerable outcome variability depending on model configurations when we consider inference at the instance level.

However, it could be the case that substantive findings may remain robust and stable across model configurations. We now turn our focus to the stance ratio measure $\frac{\mathcal{N}_{GC}}{\mathcal{N}_{GR}}$. At a macro level, Figure 2 presents the ratio of gun control posts over gun rights posts (a number greater than one indicates gun control posts outnumbers gun rights posts). We notice that certain model and prompt combination will support a substantive finding that gun control posts were outnumbered by gun rights posts and certain confirmations will support exactly the opposite. Our results corroborate the key finding of Baumann et al., 2025 that we must adopt caution when using LLMs for annotation. Furthermore, zero-shot prompting can also create scenarios in which the LLM–prompt configuration yields observations that mirror the researcher's anticipated substantive narrative rather than the underlying data.

6 Human Subjectivity

Our noise audit demonstrates that substantive empirical findings can vary meaningfully as a function of both model choice and prompting strategy. While fine-tuning models to better reflect aggregate human preferences may improve alignment with human judgments, such variability underscores the importance of interrogating not only model behavior but also the human judgments to which models are compared. We therefore next examine how human judgments themselves vary on stance mining tasks involving gun rights and gun control.

In the NLP annotation literature spanning several decades, early work largely assumed the existence of gold-standard labels, treating inter-rater disagreement primarily as a marker of annotation noise or poor data quality rather than as an expression of inherent subjectivity in the task. Under the assumption that annotation was largely substitutable across raters, little attention was paid to annotator characteristics, including demographic or ideological attributes. One of the earliest NLP studies to explicitly incorporate political ideology into the analysis of political discourse is Johnson and Goldwasser, 2018, followed by a growing body of work (Sap et al., 2022; Weerasooriya et al., 2023a; Crowl et al., 2025; Stanley et al., 2026) examining how rater political ideology shapes annotations of politically sensitive or offensive content. Our work builds on this emerging line of inquiry by explicitly incorporating political heterogeneity in the annotation process and extending it to include Independents alongside partisans (Weerasooriya et al., 2023a). In what follows, we describe the design considerations made in our annotation study.

6.1 Seed Set for Annotation

We first consider the inference results on $\mathcal{D}_{Reddit}$ and $\mathcal{D}_{YouTube}$ from nine model configurations. For a given instance and a specific class (e.g., Neither or GC), we count how many models classified the instance as that particular class and assign votes. Based on this voting, we stratified the dataset into three agreement bins: High (Yes) (8–9 votes), Medium (Maybe) (4–5 votes), and High (No) (0–1 vote). For each class, we sample equally from these bins. We obtain 900 unique YouTube comments and 240 Reddit comments to cover the spectrum of model agreement. These samples are also balanced across the four categories (Gun Rights, Gun Control, Neither, Irrelevant).

6.2 Annotation Specifics

Each of these 1,140 instances (900 YouTube comments and 240 Reddit posts) is annotated by three raters (one Republican, one Democrat, and one Independent). Following prior literature (Weerasooriya et al., 2023a; Pofcher et al., 2025), for each annotator, we set a batch size of 30 instances. Annotator compensation is also grounded in prior literature (3 USD per batch) (Weerasooriya et al., 2023a; Pofcher et al., 2025); Appendix F contains further details. We use Prolific to conduct this study using the instructions in the appendix section A. When we train model involving all political voices, we consider majority voting and discard any data point where all three raters disagreed (3.42% of all data). While training individual political models, we consider the full dataset.

6.3 Findings on Human Seed Set Annotation

On the 1,140 instances we obtain an overall Fleiss' κ agreement of 0.43 suggesting *moderate agreement*. This agreement is in line with similar politically charged tasks as annotating supportive content (Pofcher et al., 2025) or annotating offensive content (Weerasooriya et al., 2023a). Table 1 further breaks down this analysis to study pairwise agreement between parties between political groups. Contrary to prior findings that characterize Independents as occupying a "political middle" between partisan extremes (Weerasooriya et al., 2023a; Pofcher et al., 2025; Crowl et al., 2025), in

our study we notice that Republican raters agree slightly more with Democrats ($\kappa = 0.38$) than with Independents ($\kappa = 0.34$). A bootstrap (1,000 resamples) over the human annotations gives a 95% confidence interval of $[0.005, 0.081]$ for this difference. We thus present this as suggestive rather than definitive evidence of the risks of treating political groups as monolithic or universally ordered along a single ideological spectrum; patterns of agreement may appear to be task- and context-dependent, cautioning against broad generalizations that may inadvertently reinforce political stereotypes.

Recall that, when constructing our seed set for a given class c, we explicitly stratify instances into three agreement regimes: (1) near-consensus model agreement that the instance belongs to class c (8–9 votes); (2) near-consensus model agreement that the instance does not belong to class c (0–1 votes); and (3) moderate model agreement that the instance belongs to class c (4–5 votes). Across these regimes, we observe Fleiss' κ values of 0.53, 0.44, and 0.18, respectively. Hence, we observe that in regions where models exhibit substantial disagreement, human annotators likewise struggle to reach consensus. Conversely, instances on which models converge tend to elicit higher levels of human agreement. We further note an asymmetry between near-consensus agreement that an instance belongs to a class and near-consensus agreement that it does not. Agreement that a post does not belong to a particular class does not necessarily imply agreement on an alternative label, introducing an additional source of uncertainty. This distinction is reflected in the comparatively lower agreement observed in the second category.

Content Analysis of Disagreement. Manual inspection of the most contested comments, those where raters or our model configurations split most evenly, revealed three recurring patterns: semantic ambiguity, indirect or tangential signaling, and sarcastic mimicry. In all three cases, the same features that make these comments hard for human raters also mislead models. Appendix D contains a detailed exposition with examples.

	Dem	*Rep*	*Ind*
Dem	-	0.38	0.38
Rep	0.38	-	0.34
Ind	0.38	0.34	-

Table 1: Inter-annotator agreement between Republican (*Rep*), Democratic (*Dem*), and Independent (*Ind*) raters, measured using Cohen's κ, on the seed annotation dataset comprising 1,140 instances (900 from YouTube and 240 from Reddit).

6.4 Data Augmentation Via Active Learning

The initial distribution of our annotated dataset (1,140 comments) revealed moderate class imbalance with GC comments constituting 15% of the seed dataset. To mitigate class imbalance and to obtain a meaningful training set with diverse examples of varying difficulty, we employ two well-known active learning sampling strategies: minority-class certainty sampling, and uncertainty sampling. Brief introduction of active learning concepts and the two sampling strategies follow next.

Active learning is a well-established supervised machine learning paradigm in which model training proceeds through iterative interaction between a learner and a human supervisor (Settles, 2009). In the commonly used pool-based setting, the learner is initialized on a small labeled seed set and granted access to a large pool of unlabeled instances. At each iteration, the learner selects instances to query based on a sampling strategy, receives labels from human annotators, and retrains on the augmented dataset until a stopping criterion (e.g., budget exhaustion or performance convergence) is reached. Some of the well-known active learning sampling strategies are uncertainty sampling (the model selects the instances it is most unsure about) (Scheffer et al., 2001; Settles, 2009); certainty sampling (the model selects the instances it is most confident about) (Sindhwani et al., 2009); or density-based sampling (Settles, 2009). In order to rectify high-confidence misclassifications, and perform robustly at decision boundaries, we perform certainty sampling (Sindhwani et al., 2009) and margin sampling (Scheffer et al., 2001), a variant of uncertainty sampling tailored for multi-class classification.

The active learning literature is extensive, and adjacent work combining human and LLM annotation exists. For example, (Li et al., 2023) use LLM uncertainty to determine which instances an LLM should label instead of a human, while (Gligorić et al., 2025) use LLM confidence to route selected instances to human annotators while preserving valid statistical estimation. In both cases, human labels serve as the gold standard, and uncertainty is used to improve annotation efficiency. Our setting is fundamentally different: we do not propose a new sampling strategy or estimate a target quantity. Instead, we use standard certainty and margin sampling to construct a class-balanced training set

LLM	Macro F_1
Mistral	73.99 ± 4.13
Llama	$\mathbf{78.30 \pm 3.63}$
GPT	72.54 ± 3.91

Table 2: Performance comparison of baselines on our test set over five independent training runs with an 80/10/10 train/test/validation split. For all models and a given run, we use the identical test/train/validation split. The best-performing model result is highlighted in bold.

containing both confident and ambiguous examples. Hence, prior work treats human–LLM disagreement as noise to minimize; in our study, that disagreement is itself the phenomenon of interest.

We initially fine-tune three models (`Mistral`, `GPT`, and `Llama`) on a seed set of 1,140 instances using an 80/10/10 train–validation–test split. These models are then used to perform certainty and uncertainty sampling, expanding the dataset to a total of 1,560 instances. Training on the final dataset, again using an 80/10/10 split and averaged over five independent runs, we observe that `Llama` achieves the strongest performance (Table 2).

To address our substantive research question, we additionally train `Llama` models using annotations from individual political groups: $\mathcal{M}_{red}$, $\mathcal{M}_{dem}$, $\mathcal{M}_{ind}$. As an aside, we report Fleiss' κ values of 0.23 for instances selected via uncertainty sampling and 0.57 for those selected via certainty sampling. This pattern is intuitive: human annotators exhibit substantially higher agreement on instances for which models are more confident, while disagreement increases on instances where model predictions are uncertain.

6.5 "Subjective" Substantive Findings

Figure 3 examines our substantive questions through the lens of models trained on annotations from politically distinct groups. Models trained on Republican and Independent annotations indicate that pro–gun rights discourse outweighs pro–gun control discourse across both platforms. In contrast, models trained on Democratic annotations, as well as the aggregate model, yield the opposite conclusion. Figure 2 shows that the baseline `LLaMA` model under zero-shot prompting consistently suggests that pro–gun control discourse is less prevalent than pro–gun rights discourse on both platforms. Our findings thus illustrate substantive conclusions can diverge depending on whose judgments are used for supervision and how the data are curated, underscoring the sensitivity of social web analyses

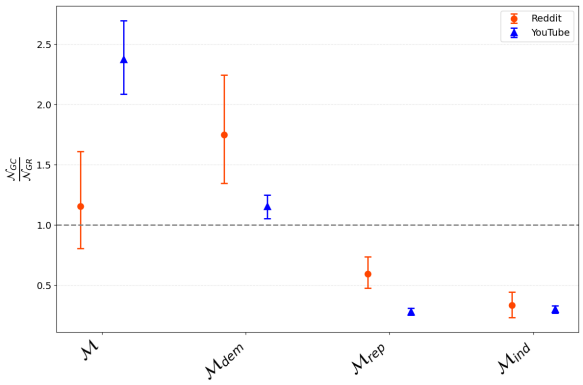


Figure 3: Stance ratio (Gun Control / Gun Rights) produced by Llama models trained on labels provided by different rater groups. $\mathcal{M}_{dem}$, $\mathcal{M}_{rep}$, and $\mathcal{M}_{ind}$ represent the models trained on labels provided by Democrats, Republicans, and Independents, respectively. $\mathcal{M}$ denotes the model trained on aggregate labels obtained through majority voting. Ratios above the horizontal dashed line (> 1) indicate a greater prevalence of gun control–favoring discourse, while ratios below the line (< 1) indicate a greater prevalence of gun rights–favoring discourse. 95% confidence intervals are estimated via bootstrap sampling.

to both human and modeling choices.

7 Conclusions and Future Work

This work examines how both machine-generated and human judgments introduce subjectivity into the analysis of socially consequential discourse, and how this subjectivity can shape substantive empirical findings. Focusing on the well-studied task of stance detection in gun-related discussions, we show that observed conclusions are sensitive to multiple design choices, including the selection of models, prompting strategies, sampling policies, and the political orientations of human raters. As the technical barriers to deploying LLMs continue to decline and researchers from adjacent fields increasingly rely on these tools to study social and public policy questions, it becomes important to recognize that the "overall story" emerging from such analyses may depend critically on these interacting factors.

Our study also motivates rethinking the sociotechnical aspects of traditional machine learning literature developed over decades such as active learning. In socially and politically sensitive domains, annotation is not merely a mechanical labeling task but a subjective judgment shaped by annotators' ideological positions, lived experiences, and interpretive frames. As a result, the "oracle"

in active learning is not a neutral source of ground truth but a heterogeneous human population whose judgments may systematically vary. This reframing highlights active learning as a socio-technical process in which sampling policies interact with political heterogeneity in the annotator pool, with direct implications for whose interpretations are learned, reinforced, and ultimately deployed at scale.

8 Limitations

While our study is intended to be systematic and comparative, it necessarily reflects a set of scoped design choices that suggest several directions for future work. Our analysis is situated within the sociolinguistic context of U.S. politics and examines English-language content drawn from prominent U.S.-based platforms (Reddit and YouTube). We focus on a stance-mining task that has received sustained attention in the literature and evaluate a set of widely used open-weight models. Future research could broaden this scope by extending the analysis to non-English contexts, alternative platforms—including both other U.S.-based spaces (e.g., Gab, Truth Social, X) and platforms prominent in other cultural settings (e.g., Sina Weibo)—as well as additional stance or opinion-mining tasks and proprietary models such as GPT, Claude, or Gemini. Exploring these dimensions would help assess the generality of our findings across linguistic, cultural, and technical settings.

Our annotation design also leaves room for refinement. Although we explicitly include Independent annotators alongside Republican and Democrat raters, political identity remains coarsely defined. Future work could consider more fine-grained ideological positions, such as combinations of social and fiscal preferences, to better capture the multidimensional nature of political belief. In addition, our analysis focuses on first-person judgments—how annotators themselves interpret the stance of a given piece of content. Building on recent work on vicarious interactions, future studies could examine second-order judgments, such as how annotators reason about how members of other political groups might label the same content. Finally, an interesting direction is to connect human–machine alignment more directly with emerging in silico surrogate approaches, examining when and how persona-conditioned language models approximate or diverge from the judgments of politically situated human annotators.

9 Ethical Statement

This study uses publicly available data collected through official platform APIs. Although the majority of the content in our annotation task is not expected to be overtly offensive, exposure to potentially harmful or disturbing language is an inherent risk when working with politically charged social web discourse. To safeguard annotators' mental well-being, we limited annotation batches to 30 instances, reducing the likelihood of prolonged exposure to toxic or distressing content. Annotators were adequately compensated and provided informed consent through a standard crowdsourcing platform prior to participation. The study adopts data collection and annotation protocols closely aligned with prior work by the authors that received an IRB exempt determination and was developed under comparable ethical guidelines. The study design and annotation protocol were overseen by an experienced social scientist with prior service on ethics committees for major academic conferences. Finally, we used ChatGPT as a writing assistant for flow and grammar. All scientific decisions, interpretations, and core arguments were made by the authors.

Acknowledgments

Datta was supported by an RIT GCI Faculty Growth Award. We thank Cornelia Caragea and Doina Caragea for their input.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science*, 382(6671):669–674.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv preprint arXiv:2406.11980*.

Fan Bai, Alan Ritter, and Wei Xu. 2021. Pre-train or annotate? domain adaptation with a constrained budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*.

Adrian Benton, Braden Hancock, Glen Coppersmith, John W Ayers, and Mark Dredze. 2016. After sandy hook elementary: A year in the gun control debate on twitter. *arXiv preprint arXiv:1610.02060*.

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story@ECIR*.

Michael Burnham. 2025. Stance detection: a practical guide to classifying political beliefs in text. *Political Science Research and Methods*, 13(3):611–628.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.

Logan Crowl, Sujan Dutta, Ashiqur R KhudaBukhsh, Edson Severnini, and Daniel S Nagin. 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.

Kareem Darwish. 2019. Quantifying polarization on twitter: the kavanaugh nomination. In *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings 11*, pages 188–201. Springer.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 141–152.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*.

Sujan Dutta, Beibei Li, Daniel S Nagin, and Ashiqur R KhudaBukhsh. 2022. A murder and protests, the capitol riot, and the chauvin trial: Estimating disparate news media stance. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5059–5065.

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

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. 2025. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533.

Nikesh Gyawali, Iustin Sirbu, Tiberiu Sosea, Sarthak Khanal, Doina Caragea, Traian Rebedea, and Cornelia Caragea. 2024. Gunstance: Stance detection for gun control and gun regulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12027–12044. Association for Computational Linguistics.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 720–730.

Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. 2017. Constance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1115–1124.

Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A flaw in human judgment*. Little, Brown Spark.

Mayank Kejriwal, Henrique Santos, Ke Shen, Alice M Mulvehill, and Deborah L McGuinness. 2024. A noise audit of human-labeled benchmarks for machine commonsense reasoning. *Scientific Reports*, 14(1):8609.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 504–514.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019b. Learning to predict population-level label distributions. In *Companion of The 2019 World Wide Web Conference, WWW 2019*, pages 1111–1120. ACM.

Rijul Magu, Nivedhitha Mathan Kumar, Yihe Liu, Xander Koo, Diyi Yang, and Amy Bruckman. 2024. Understanding online discussion across difference: Insights from gun discourse on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–28.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Seyedali Mohammadi, Bhaskara Hanuma Vedula, Hemank Lamba, Edward Raff, Ponnurangam Kumaraguru, Francis Ferraro, and Manas Gaur. 2025. Do llms adhere to label definitions? examining their receptivity to external label definitions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32380–32393, Suzhou, China. Association for Computational Linguistics.

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

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Pofcher, Christopher M Homan, Randall Sell, and Ashiqur R KhudaBukhsh. 2025. Hope vs. hate: Understanding user interactions with lgbtq+ news content in mainstream us news media through the lens of hope speech. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19862–19888.

Drazen Prelec. 2004. A bayesian truth serum for subjective data. *science*, 306(5695):462–466.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 29971–30004.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden Markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Vikas Sindhwani, Prem Melville, and Richard D. Lawrence. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, pages 953–960.

Miles Stanley, Soumyajit Datta, Ashutosh Kumar, and Ashiqur R KhudaBukhsh. 2026. Investigating vaccine buyer's remorse: Post-vaccination decision regret in covid-19 social media using politically diverse human annotation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39294–39302.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023a. 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*, pages 11648–11668.

Tharindu Cyril Weerasooriya, Tong Liu, and Christopher M. Homan. 2020. Neighborhood-based pooling for population-level label distribution learning. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 490–497. IOS Press.

Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher M. Homan. 2023b. Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–966, Toronto, Canada. Association for Computational Linguistics.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, Senzhang Wang, Xiaochun Cao, and Hu Huang. 2024. A survey of stance detection on social media: New directions and perspectives. *arXiv preprint arXiv:2409.15690*.

Yini Zhang, Dhavan Shah, Jordan Foley, Aman Abhishek, Josephine Lukito, Jiyoun Suk, Sang Jung Kim, Zhongkai Sun, Jon Pevehouse, and Christine Garlough. 2019. Whose lives matter? mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication*, 24(4):182–202.

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

A Annotation Instruction

In this study, you will read short social-media comments and classify each one into one of four categories related to gun policy: Gun Control, Gun Rights, Neither, or Irrelevant. For every comment shown to you, you will select the category that best describes the stance expressed.

Your task involves carefully interpreting each comment, deciding which category it fits into based on the definitions provided. There are no right or wrong answers—your judgment and interpretation are what matter.

B Prompts

Figures 4, 5, and 6 depict the three prompt settings vanilla, definition 1 (standard), and definition 2 (refined), respectively.

Vanilla Prompt

This is posted by an anonymous user on Social Media.
FIRST, Classify the comment as a whole using the following instructions as either 'Gun Rights', 'Gun Control', 'Neither' or 'Irrelevant'.
SECOND, Provide a brief reason for your classification in one sentence.
Instructions:

1. You should use the context and the meaning of the sentence to classify.
2. Do not use just the keywords to classify.
3. Always return your response as a JSON object with the keys 'classification' and 'reason'.

Comment : "{comment}"

Figure 4: Vanilla LLM prompt

1st Definition Prompt

You will be provided with a social media comment. Classify it into one of the following four categories. Your classification should consider not only what is explicitly stated but also the comment's primary focus, tone, and implied sentiment to capture nuances that are not clearly expressed. Provide a brief, one-sentence reason for your classification.
Categories and Definitions:
1. Gun Control

- **Definition:** Advocates for, or implies a desire for, greater regulation and oversight of firearms. This includes explicit calls for legislation restricting civilian gun ownership as well as comments where the primary focus is a critique of how firearms are used, including by law enforcement.
- **Common ways this stance is expressed include:**
  - Calling for Legislative Action (e.g., universal background checks, banning certain weapons).
  - Critiquing Law Enforcement's Use of Firearms (e.g., calling them "trigger happy," questioning the use of lethal force).
  - Expressing Emotional Outrage or Fear about gun violence.
  - Advocating for Non-Lethal Alternatives (e.g., "They should have used a taser").
  - Criticizing "Gun Culture" or the prevalence of firearms in society.

2. Gun Rights

- **Definition:** Advocates for, or implies support for, the right of individuals to own firearms. This includes explicit opposition to restrictions and appeals to the Second Amendment, as well as comments where the primary focus is on the necessity of guns for self-defense, deterring crime, or as a check on government power.
- **Common ways this stance is expressed include:**
  - Advocating for Self-Defense as a primary need.
  - Opposing Restrictions and Appealing to the Constitution (e.g., "shall not be infringed").
  - Expressing Distrust in Authority, necessitating personal arms for protection.
  - Justifying Firearm Use in Specific Scenarios based on a perceived threat.
  - Arguing that an armed populace acts as a deterrent to crime.

3. Neither

- **Definition:** The comment is on-topic (related to guns, a specific incident, or law enforcement) but does not express a discernible stance for Gun Control or Gun Rights. This includes comments that are purely descriptive, express ambiguous or vague sentiments, ask a neutral question, or focus on a legal proceeding without critiquing the underlying law.

4. Irrelevant

- **Definition:** The comment is off-topic and has no connection to the discussion of firearms, gun policy, or the incident being discussed. This includes spam, personal conversations, or comments on completely unrelated subjects.

Always return your response as a JSON object with the keys 'classification' and 'reason'.
Comment : {comment}

Figure 5: Definition 1 (Standard) Prompt

2nd Definition Prompt

You will be provided with a social media comment. Classify it into one of the following four categories. Provide a brief, one-sentence reason for your classification.
Categories and Definitions:
1. Gun Control

- **Definition:** Advocates for, or implies a desire for, greater regulation and oversight of firearms. This includes explicit calls for legislation restricting civilian gun ownership as well as comments where the primary focus is a critique of how firearms are used, including by law enforcement.
- **Common ways this stance is expressed include:**
  - Calling for Legislative Action (e.g., universal background checks, banning certain weapons).
  - Critiquing Law Enforcement's Use of Firearms (e.g., calling them "trigger happy," questioning the use of lethal force).
  - Expressing Emotional Outrage or Fear about gun violence.
  - Advocating for Non-Lethal Alternatives (e.g., "They should have used a taser").
  - Criticizing "Gun Culture" or the prevalence of firearms in society.

2. Gun Rights

- **Definition:** Advocates for, or implies support for, the right of individuals to own firearms. This includes explicit opposition to restrictions and appeals to the Second Amendment, as well as comments where the primary focus is on the necessity of guns for self-defense, deterring crime, or as a check on government power.
- **Common ways this stance is expressed include:**
  - Advocating for Self-Defense as a primary need.
  - Opposing Restrictions and Appealing to the Constitution (e.g., "shall not be infringed").
  - Expressing Distrust in Authority, necessitating personal arms for protection.
  - Justifying Firearm Use in Specific Scenarios based on a perceived threat.
  - Arguing that an armed populace acts as a deterrent to crime.

3. Neither

- **Definition:** The comment is on-topic (related to guns, a specific incident, or law enforcement) but does not express a discernible stance for Gun Control or Gun Rights. This includes comments that are purely descriptive, express ambiguous or vague sentiments, ask a neutral question, or focus on a legal proceeding without critiquing the underlying law.

4. Irrelevant

- **Definition:** The comment is off-topic and has no connection to the discussion of firearms, gun policy, or the incident being discussed. This includes spam, personal conversations, or comments on completely unrelated subjects.

Always return your response as a JSON object with the keys 'classification' and 'reason'.
Comment : {comment}

Figure 6: Definition 2 (Refined) Prompt

C Context Length Versus Agreement

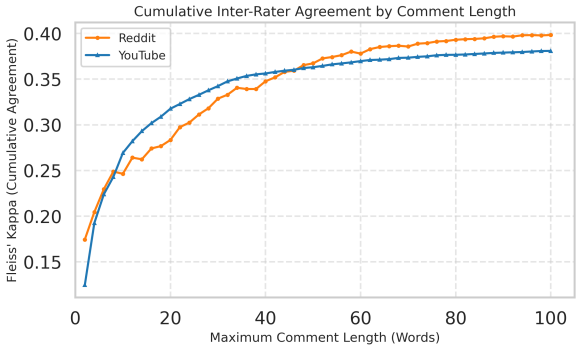


Figure 7: LLM Inter-Model Agreement (Reddit & YouTube): Cumulative Fleiss' Kappa by comment word length. The plot shows how inter-rater agreement among the 9 LLM configurations evolves as progressively longer comments are included in the analysis across both platforms.

D Content Analysis of Disagreement

We elaborate on the three patterns introduced in Section 6.3.

Semantic ambiguity. Many contested comments express vague frustration without naming a specific policy. The comment "So what's stopping the government from stepping in to stop all this" is read inconsistently across our model configurations, since it can mean either curbing gun violence or resisting government overreach. Comments that mix opposing signals also divide raters: one user

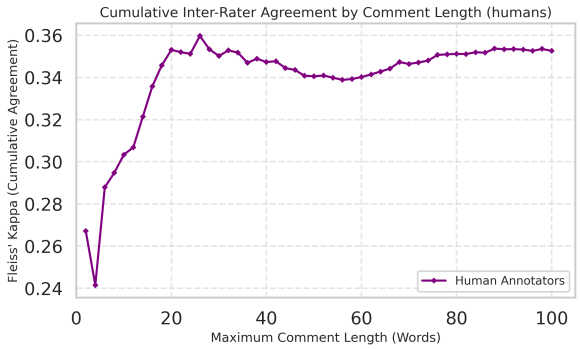


Figure 8: Human Agreement (Reddit & YouTube): Cumulative Fleiss' Kappa by comment word length.

criticizes firearms while defending the right to own them, writing "this just proves that guns are just for show.. When it counts people wont excerise there right." Complex policy proposals are also divisive, as with a comment urging "FORT KNOX EVERY SCHOOL... armed guards, bulletproof glass, metal detectors."

Indirect and tangential signaling. Agreement is also lower when stance is expressed through analogy or comparison rather than stated directly. Users rely on analogies such as "If you subtract suicides out of child deaths it is car accidents... So how about we ban kids from cars?" or international comparisons such as "BS in Australia the skegs use knives and fists violence is dreadful here." Weak or indirect signals lower agreement further, as when a comment focuses on a side issue such as police inaction instead of gun policy.

Sarcastic mimicry. Sarcasm is a further source of disagreement. When a user mocks the other side, writing "Lemme guess ..all them automatic pistols by legal owners!" followed by an eye-roll emoji, raters divide on whether they catch the irony. It also misleads models, since the repetition of an opponent's keywords can pull them toward the wrong label, as when a pro-gun-rights user writes "In the span of 3 minutes he says, 'common sense' seven times... This is an Everytown talking point," which can be misread as a Gun Control stance.

E Inferencing with Larger Models

To check whether the noise-audit patterns reported in Section 5 hold at larger scale, we repeat the Reddit analysis with two larger models from the Mistral and Llama families: `Mistral-Medium` (Mistral-Medium-3.5) and `Llama-70B` (Llama-3.1-70B). Together with the three models from the main text (`Mistral`, `Llama`, `GPT-OSS`) and three prompting strategies, this yields fifteen $\langle model, prompt \rangle$ con-

figurations on Reddit. Figure 9 reports pairwise Cohen's κ across these fifteen configurations, and Figure 10 reports the corresponding stance ratios together with the nine YouTube configurations for comparison.

Within both the Mistral and Llama families, the larger model gives stance ratios closer to one. Mistral-7B's ratios range from 1.47 to 1.90 across prompts, while Mistral-Medium-3.5's range from 0.68 to 1.15. For Llama the same pattern holds: 0.51–0.65 for Llama-3-8B and 0.74–1.05 for Llama-3.1-70B. Prompt sensitivity persists at this scale. Mistral-Medium-3.5's ratio nearly doubles between vanilla and definition 1 (0.68 to 1.15), and Llama-3.1-70B shifts from 0.74 to 1.05 across the same prompts. The main-text observation that prompt choice alters the substantive conclusion thus holds in the larger setup. Overall inter-configuration agreement on Reddit is also nearly unchanged, with Fleiss' $\kappa = 0.43$ in both the nine- and fifteen-configuration setups.

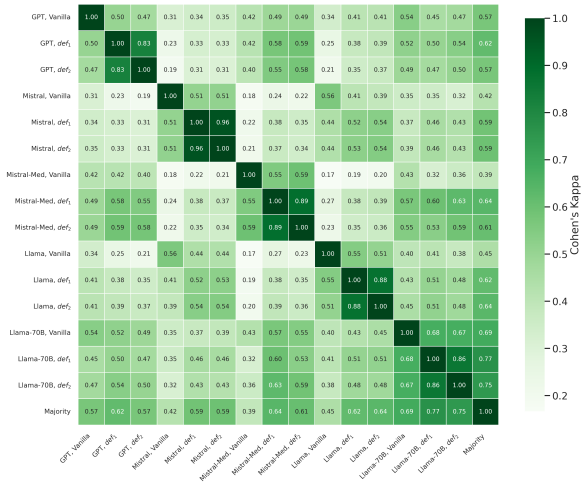


Figure 9: Agreement between $\langle \mathcal{M}, prompt \rangle$ pairs on Reddit with five models (the three main-paper models plus `Mistral-Medium` and `Llama-70B`). Cell $\langle i, j \rangle$ presents Cohen's κ between $\mathcal{M}_i$ and $\mathcal{M}_j$. `Majority` takes the majority vote of the fifteen individual configurations.

F Annotator Compensation

Annotator compensation is grounded in prior literature (3 USD per batch) (Weerasooriya et al., 2023a; Pofcher et al., 2025). A small pilot (median completion time: 9 minutes 59 seconds; mean: 12 minutes 11 seconds) confirms this exceeds the minimum hourly wage, consistent with rates reported in prior annotation studies (Bugert et al., 2020; Bai et al., 2021).

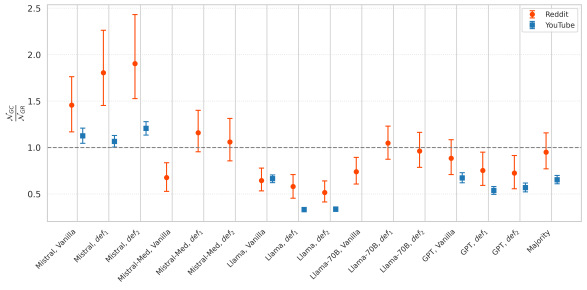


Figure 10: Stance ratio (Gun Control / Gun Rights) across all $\langle model, prompt \rangle$ configurations: fifteen on Reddit (five models $\times$ three prompts) and nine on YouTube (three models $\times$ three prompts). Ratios above the dashed line (> 1) indicate a greater prevalence of gun-control–favoring discourse; ratios below (< 1) indicate a greater prevalence of gun-rights–favoring discourse. 95% confidence intervals are estimated via bootstrap sampling.