Large Language Models Unpack Complex Political Opinions through Target-Stance Extraction

Özgür Togay[1], **Javier Garcia Bernardo**[1],
Florian Kunneman[2], **Anastasia Giachanou**[1]
[1]Department of Methodology and Statistics, Utrecht University
[2] Department of Languages, Literature and Communication, Utrecht University
{o.togay, j.garciabernardo, f.a.kunneman, a.giachanou}@uu.nl

Abstract

Political polarization emerges from a complex interplay of beliefs about policies, figures, and issues. However, most computational analyses reduce discourse to coarse partisan labels, overlooking how these beliefs interact. This is especially evident in online political conversations, which are often nuanced and cover a wide range of subjects, making it difficult to automatically identify the target of discussion and the opinion expressed toward them. In this study, we investigate whether Large Language Models (LLMs) can address this challenge through Target-Stance Extraction (TSE), a recent natural language processing task that combines target identification and stance detection, enabling more granular analysis of political opinions. For this, we construct a dataset of 1,084 Reddit posts from `r/NeutralPolitics`, covering 138 distinct political targets and evaluate a range of proprietary and open-source LLMs using zero-shot, few-shot, and context-augmented prompting strategies. Our results show that the best models perform comparably to highly trained human annotators and remain robust on challenging posts with low inter-annotator agreement. These findings demonstrate that LLMs can extract complex political opinions with minimal supervision, offering a scalable tool for computational social science and political text analysis.

Keywords: political opinions, large language models, target stance extraction

1. Introduction

Political polarization, in which individuals or groups adopt extreme political beliefs and attitudes, has become a major concern in contemporary democracies (McCoy et al., 2018; Orhan, 2022; Ruggeri et al., 2024; Druckman et al., 2024). Although the mechanisms driving polarization and its spread are still debated, social media is often viewed as a key contributor (Suhay et al., 2018; Bail et al., 2018). Platforms such as Facebook, Twitter/X, and Reddit now serve as major venues for news consumption and information exchange (Newman et al., 2025), and have been found to foster affective polarization, echo chambers, and ideological segregation (Barberá et al., 2015; Bakshy et al., 2015). Most studies on polarization in social media have focused on binary classifications such as left and right or partisan labels such as Democrat and Republican (Yang et al., 2017; Garimella and Weber, 2017; Darwish, 2020; Brum et al., 2022; Peng et al., 2024; Bojić et al., 2025). A growing body of work, however, emphasizes that polarization arises from the complex interplay of opinions and beliefs (DellaPosta, 2020; Turner-Zwinkels et al., 2023; Van Noord et al., 2024), which calls for approaches that move beyond coarse labels toward fine-grained, target-aware measurement.

Accurate identification of individual positions on specific issues, known as stance detection, is important for quantifying polarization (Burnham, 2024). Traditional approaches assume that the target is known in advance (Ghosh et al., 2019) and often rely on supervised machine learning models requiring large datasets with stance labels. These models can be context-dependent and may underperform in nuanced or unseen scenarios (Küçük and Can, 2021).

Target-Stance Extraction (TSE) extends stance detection by jointly identifying both the target mentioned in a text and the stance expressed toward it (Li et al., 2023). This removes the need for pre-defined targets and enables analysis across a broader range of topics. Early TSE architectures used two separate fine-tuned transformer models for target identification and stance classification (Li et al., 2023). While this was an improvement over conventional approaches, the need for fine-tuning limited TSE's ability to operate without pre-defined targets.

Recent advances in instruction-tuned large language models (LLMs) provide an opportunity to apply TSE with little or no task-specific training. LLMs have demonstrated strong performance in various NLP tasks, including generating synthetic data to improve stance detection (Wagner et al., 2024) and matching or exceeding human annotators in sentiment and classification tasks (Törnberg, 2024; Bojić et al., 2025). By unifying target identification and stance classification in a single model, LLMs can leverage large-scale pretraining to perform TSE with minimal supervision. This

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 248–260
16 May 2026, Palma, Mallorca (Spain).

approach enables scalable, fine-grained analysis of political beliefs, improving our understanding of the mechanisms that drive polarization.

In this study, we evaluate the ability of LLMs to perform target-stance extraction (TSE) on complex political discussions. We focus on `r/NeutralPolitics`, a nonpartisan and strictly moderated subreddit known for longer, nuanced and evidence-based political conversations, making it a challenging setting for stance analysis. To support this, we constructed a manually annotated dataset of 1,084 posts, with a codebook that covers 138 targets alongside a semi-open "Other {target}" category. From the 1,084 annotated posts, a gold test set of 200 posts was further validated by an expert and used as the primary benchmark for LLM evaluation.[1] Our evaluation of proprietary and open-source LLMs of varying sizes and prompting strategies shows that they can reliably identify political beliefs, with the best model achieving an F1 score of 0.76 for target identification and 0.87 for stance detection. Even when predictions differ from the gold labels, LLMs still produce reasonable, interpretable outputs. These results highlight LLMs' potential as a scalable, nuanced tool for analyzing political opinions beyond partisan proxies.

2. Data and Methodology

This section describes our data collection, the annotation procedure for creating a gold-standard evaluation set, and the configuration and prompting strategies used for evaluating LLMs on target-stance extraction.

2.1. Data Collection

Our evaluation uses posts drawn from `r/NeutralPolitics`, a nonpartisan political forum on Reddit known for evidence-based, highly moderated discussions[2]. Unlike the emotionally charged interactions typical of Twitter/X, this subreddit enforces strict civility and sourcing rules, making it well-suited for nuanced stance analysis.

We retrieved historical posts via the Pushshift Archives (Baumgartner et al., 2020) from 2005 until April 2023, when Reddit API changes temporarily halted archiving. The dataset remains publicly available and widely used in academic research (Mok et al., 2023; Veselovsky and Anderson, 2023). Of the 578,041 comments extracted from the subreddit, 130,390 were marked as removed in the archive, meaning their text and author information were unavailable. This likely reflects subreddit moderation policies and is notably higher than the 11% removal rate reported for Reddit overall (Hofmann et al., 2022).

2.2. Data Annotation

We created an annotated evaluation dataset to cover a diverse set of political targets and issues. Three research assistants with knowledge of U.S. politics and social media were trained over three weeks. They were provided with a preliminary codebook of frequent and polarizing targets from relevant literature (Iyengar et al., 2019; American National Election Studies, 2021; Davern et al., 2024). Following the iterative open coding approach proposed by Tanweer et al. (2021), annotators were instructed to label posts openly, allowing new targets to emerge inductively. Weekly discussions refined the codebook, resolved ambiguities, and developed a shared understanding. The final codebook includes 138 distinct targets plus an open-ended "Other target" option, significantly more granular than prior studies (Mohammad et al., 2016; Li et al., 2021). In occasional cases of multi-target posts, the target agreed upon by both annotators was selected. The full list of targets and definitions is provided as supplementary material.

To obtain posts that were both politically relevant and balanced across stance categories, our sampling strategy evolved across multiple rounds. Initial random sampling from roughly 500,000 posts returned mostly conversational, non-political comments. Continuing this approach would have made creating a politically balanced dataset extremely time-consuming. To address this, we applied a GPT-4o-mini pre-filter to identify posts with political content, roughly balanced across positive, neutral, and negative stances. Prefilter labels were not shown to annotators, ensuring unbiased labeling. Annotators received both the submission title and the full comment text and completed four rounds of training to become familiar with the task and codebook. Agreement improved as the codebook stabilized, and this refined strategy was adopted for the main annotation phase.

Three annotators labeled posts in rotating pairs, such that each post was annotated by two annotators. In total, 1,084 posts were annotated. Krippendorff's α for target detection was 0.48, indicating moderate agreement (Landis and Koch, 1977). As α penalizes skewed category distributions and treats all mismatches as equally distant, it likely underestimates agreement in our setting with 138 semantically related target categories (Zhao et al., 2013; Lacy et al., 2015).

Annotators agreed on the target for 568 posts and on both target and stance for 385 posts, with

[1] Both datasets are publicly available at `https://github.com/zgrtgy/llm-tse`.

[2] https://www.reddit.com/r/NeutralPolitics/wiki/index/

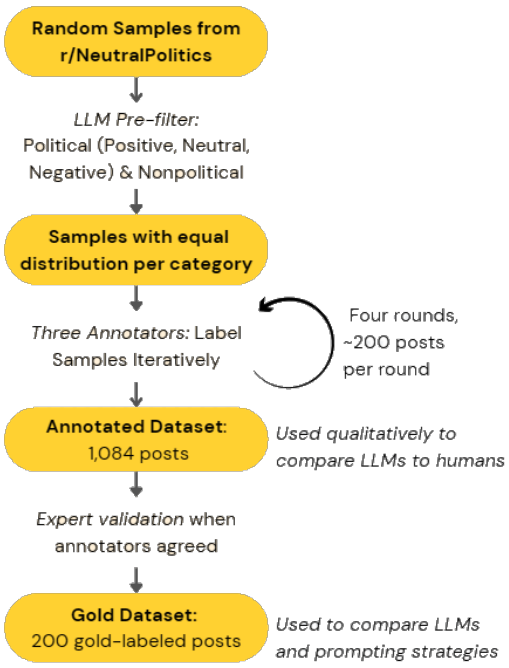

Figure 1: Annotation Process

	Pair	Count	%
Agreements	Positive - Positive	116	20%
	Neutral - Neutral	196	35%
	Negative - Negative	73	13%
	Total	**385**	**68%**
Disagreements	Positive - Neutral	83	15%
	Positive - Negative	34	6%
	Negative - Neutral	66	12%
	Total	**183**	**32%**

Table 1: Stance agreement distribution for samples with matching target labels (N = 568 of 1,084). The remaining 516 samples had mismatched targets.

Category	Count	% of Gold Dataset
Policies & Socioeconomic Issues	50	25%
Key Divisive Issues	29	14.5%
Key Politicians	29	14.5%
Trust & Institutions	20	10%
Political Groups	11	5.5%
Technology & Governance Issues	5	2.5%
International Relations	3	1.5%
Ethnic & Religious Groups	3	1.5%
No Target	50	25%

Table 2: Distribution of target categories for gold-labeled posts. For mapping criteria, see Appendix A.

disagreements most frequent in neutral stances and multi-target posts. Table 1 presents the distribution of agreements and disagreements by stance pairs. From this agreed subset, an expert on political polarization validated a balanced sample of 200 posts (50 with no target, 50 per stance class), forming the gold-labeled dataset used for evaluation. Figure 1 shows the steps followed during the annotation process. This subset covers 58 distinct targets. Although smaller than the full list of 138 targets, balancing all targets across multiple stances would require oversampling rare targets, which could reduce representativeness. For readability, targets were grouped into higher-level categories according to their political role or type when reporting distributions. For example, "Trump" and "Biden" were assigned to *Key Politicians*, "Republicans" and "Democrats" to *Political Groups*, "Christians" and "Hispanics" to *Ethnic & Religious Groups*, and "Trust in the US Military" to *Trust & Institutions*. Table 2 reports category-level distributions; individual target frequencies are provided in Appendix A.

LLMs were evaluated against the full 138-target list, preserving the task's granularity and difficulty. The average post length is 106 words (648 characters), considerably longer and more complex than tweet-based datasets such as PStance (30 words / 180 characters) (Li et al., 2021) and SemEval 2016 (17 words / 108 characters) (Mohammad et al., 2016).

2.3. Model Configuration

We accessed *GPT-4.1*, *GPT-4.1-mini*, and *o3* through OpenAI's API, while other models were accessed or deployed via Google Cloud's Vertex AI platform. Smaller open-weight models (such as *Qwen* and *Gemma*) were deployed directly on Vertex AI, while larger models (e.g., *LLaMA*, *Gemini*) were accessed via Vertex AI endpoints. Both providers explicitly state that user inputs sent through their APIs are not used for training purposes [3][4], mitigating concerns over data leakage.

We ran all models with a fixed seed and temperature set to 0.0[5]. It should be noted that these still do not guarantee deterministic outputs; however, averages taken over three runs show minimal variation in metrics.

2.4. Prompting Strategies

Prompting refers to crafting instructions that guide LLMs toward producing desired outputs. Unlike training or fine-tuning, prompting allows researchers to steer model behavior using carefully designed instructions, examples, or contex-

[3] https://platform.openai.com/docs/concepts

[4] https://cloud.google.com/vertex-ai/generative-ai/docs/data-governance

[5] Except o3, which does not support a temperature of 0.0, as it hinders reasoning.

tual cues.

We evaluate the most commonly used prompting strategies to assess their effect on target-stance extraction:

Zero-shot prompting: The prompt includes only the task instruction, serving as a baseline for comparison.

Few-shot prompting: Prompt includes one example per stance category to provide the model with a small reference set.

Conversational context augmentation: Up to four surrounding posts from the same Reddit thread are included, prioritizing parent posts and nearby replies, to help models understand co-references and infer author intent (Niu et al., 2024).

Informational context augmentation: Short descriptions of targets, taken directly from the annotation codebook, are appended to resolve ambiguity and standardize understanding of targets (Shafiei et al., 2025).

Full prompt templates are provided in Appendix C.

3. Results

In this section, we report the results and key findings from evaluating several LLMs under different prompting configurations on our gold-labeled dataset.

3.1. Comparison of LLMs

Table 3 reports the performance of various LLMs on target and stance extraction using the zero-shot strategy, which serves as our baseline. Larger proprietary models (*GPT-4.1, o3, Gemini-2.5*) achieve high stance detection scores (>0.80), while target identification remains around 0.70. This likely reflects the difficulty of detecting 138 possible targets, often implicitly or ambiguously mentioned in comments, compared to only three stance categories. We observe that reasoning models improve stance detection but not target identification, with sharp drops in performance for models under 8B parameters. These results are promising given the general-purpose zero-shot setup and the challenging nature of the task.

To assess robustness to target granularity, we ran a variation in which detailed targets were replaced with broader target labels. For example, "Republican Party," "Republican politicians," "Republican politician (non-listed)," "Republicans," and "Conservatives" were all replaced with a single label ("Republicans/Conservatives"). As shown in Table 4, this slightly improves target identification, especially in larger models, while effects on stance detection are mixed. The full mapping of detailed to broader labels is provided in Appendix B.

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
gpt-4.1	**0.73**	**0.70**	0.81	0.81
o3	0.71	0.68	**0.84**	**0.85**
gemini-2.5-pro	0.71	0.67	0.84	0.84
gemini-2.5-flash	0.70	0.66	0.84	0.84
llama-3.1-405b	0.66	0.66	0.80	0.80
gpt-4.1-mini	0.63	0.61	0.79	0.79
qwen-3-32b	0.61	0.58	0.73	0.73
llama-3.1-70b	0.59	0.60	0.78	0.78
llama-4-maverick	0.59	0.58	0.63	0.63
qwen-3-8b	0.58	0.56	0.72	0.72
gemini-2.5-flash-lite	0.57	0.55	0.74	0.75
qwen-3-14b	0.54	0.55	0.76	0.77
gemma-3-27b	0.55	0.50	0.73	0.73
gemma-3-12b	0.51	0.49	0.72	0.72
llama-4-scout	0.50	0.49	0.69	0.68
ds-r1-qwen3-8b	0.50	0.51	0.70	0.71
llama-3.1-8b	0.36	0.34	0.67	0.68
gemma-3-4b	0.28	0.27	0.53	0.57
qwen-3-1.7b	0.13	0.17	0.46	0.48
qwen-3-0.6b	0.04	0.06	0.35	0.36
gemma-3-1b	0.02	0.02	0.53	0.67

Table 3: Zero-shot performance of all tested models, sorted by Target F1. Bold indicates the best score per metric.

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
gpt-4.1	0.74↑	0.73↑	0.81	0.82↑
gemini2.5-flash	0.74↑	0.70↑	0.86↑	0.86↑
gemini2.5-pro	0.72↑	0.68↑	0.85↑	0.85↑
o3	0.72↑	0.70↑	0.84	0.85
llama3-405b	0.67↑	0.67↑	0.81↑	0.80
gpt-4.1-mini	0.65↑	0.64↑	0.77↓	0.77↓
llama3.1-70b	0.60↑	0.62↑	0.79↑	0.80↑
llama4-maverick	0.60↑	0.60↑	0.68↑	0.67↑
gemini2.5-flash-lite	0.60↑	0.59↑	0.71↓	0.71↓
qwen3-32b	0.57↓	0.56↓	0.76↑	0.77↑
qwen3-14b	0.57↑	0.57↑	0.74↓	0.75↓

Table 4: TSE metrics using broad target labels, ordered by Target F1. Up arrows (↑) indicate increases from zero-shot performance, down arrows (↓) indicate decreases.

3.2. LLM Performance Across Prompting Strategies

We next evaluate how different prompting strategies affect LLM performance, focusing on the best performing models.

Few-shot Prompting Few-shot prompting consistently improves target identification across models, while stance detection remains largely stable or shows only slight gains (Table 5). This strategy provides strong overall performance without substantially increasing computational cost, making it an effective way to guide models with a few representative examples.

Conversational Context Augmentation Including up to four surrounding thread posts yields mixed results (Table 6). Some models (e.g., GPT-4.1 variants) perform worse, suggesting sensitivity to longer prompts. Strong reasoning models (Gemini-2.5, Llama 4 Maverick, o3) show more consistent benefits, especially for target identifica-

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
gemini-2.5-pro	0.76↑	0.73↑	0.82↓	0.82↓
o3	0.75↑	0.73↑	0.84	0.84↓
gemini-2.5-flash	0.74↑	0.71↑	0.81↓	0.82↓
gpt-4.1	0.74↑	0.72↑	0.82↑	0.82↑
gpt-4.1-mini	0.70↑	0.68↑	0.78↓	0.78↓
gemini-2.5-flash-lite	0.61↑	0.59↑	0.78↑	0.78↑
qwen-3-32b-fp8	0.63↑	0.61↑	0.72↓	0.72↓
llama-4-maverick	0.62↑	0.60↑	0.70↑	0.69↑
llama-3.1-405b	0.65↓	0.65↓	0.84↑	0.84↑
llama-3.1-70b	0.55↓	0.54↓	0.78↑	0.79↑
qwen-3-14b-fp8	0.57↑	0.57↑	0.75↓	0.75↓

Table 5: TSE metrics using few-shot prompting, ordered by Target F1. Arrows indicate changes from zero-shot performance.

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
gemini-2.5-pro	0.74↑	0.70↑	0.84	0.84
o3	0.73↑	0.72↑	0.84	0.85
gemini-2.5-flash	0.72↑	0.68↑	0.84	0.84
llama-4-maverick	0.62↑	0.59↑	0.73↑	0.73↑
gemini-2.5-flash-lite	0.60↑	0.58↑	0.77↑	0.77↑
llama-3.1-405b	0.62↓	0.61↓	0.83↑	0.83↑
llama-3.1-70b	0.58↓	0.57↓	0.84↑	0.84↑
qwen-3-32b-fp8	0.58↓	0.57↓	0.76↑	0.77↑
gpt-4.1	0.66↓	0.64↓	0.78↓	0.79↓
qwen-3-14b-fp8	0.50↓	0.52↓	0.83↑	0.84↑
gpt-4.1-mini	0.56↓	0.53↓	0.75↓	0.76↓

Table 6: TSE metrics using conversational context, ordered by Target F1. Arrows indicate changes from zero-shot performance.

tion. These gains, however, come with substantially higher token costs.

Informational Context Augmentation Providing short explanations of or background information about the targets consistently helps models identify the correct targets, as shown in Table 7. The effect on stance detection is less consistent.

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
gemini-2.5-flash	0.75↑	0.71↑	0.82↑	0.83↑
gemini-2.5-pro	0.75↑	0.73↑	0.84↑	0.84↑
gpt-4.1	0.74↑	0.73↑	0.81↓	0.81↓
o3	0.73↑	0.72↑	0.84	0.85
llama-3.1-405b	0.70↑	0.68↑	0.81↑	0.81↑
gpt-4.1-mini	0.67↑	0.64↑	0.75↓	0.75↓
gemini-2.5-flash-lite	0.64↑	0.63↑	0.76↑	0.77↑
llama-3.1-70b	0.62↑	0.62↑	0.74↓	0.74↓
qwen-3-32b-fp8	0.64↑	0.61↑	0.73	0.73
llama-4-maverick	0.59	0.58	0.81↑	0.81↑
qwen-3-14b-fp8	0.57↑	0.56↑	0.71↓	0.72↓

Table 7: TSE metrics using informational context, ordered by Target F1. Arrows indicate changes from zero-shot performance.

Few-shot with Informational Context To further enhance performance, we tested combining few-shot prompting with additional informational context (i.e., the codebook descriptions of targets) in the prompt (Table 8). While this does not consistently improve all models, pairing it with the o3 model yielded the best results in our tests, achieving the highest combined Target and Stance F1 scores (Figure 2).

Model	Target F1	Target Acc.	Stance F1	Stance Acc.
o3	0.76↑	0.75↑	0.87↑	0.87↑
gemini-2.5-pro	0.73↓	0.72↓	0.84↑	0.85↑
gemini-2.5-flash	0.73↓	0.70↓	0.83↑	0.84↑
gpt-4.1	0.74↑	0.72↑	0.82↑	0.82↑
gpt-4.1-mini	0.70	0.68	0.77↑	0.78
gemini-2.5-flash-lite	0.63↑	0.62↑	0.75↓	0.75↓
llama-4-maverick	0.63↑	0.61↑	0.68↓	0.67↓
qwen-3-14b-fp8	0.58↑	0.57	0.75	0.75
llama-3.1-70b	0.61↑	0.60↑	0.75↓	0.76↓
qwen-3-32b-fp8	0.63	0.61	0.71↓	0.71↓
llama-3.1-405b	0.65↓	0.65↓	0.84↓	0.84↓

Table 8: TSE metrics using few-shot prompting with informational context, ordered by Target F1. Arrows indicate changes from few-shot performance.

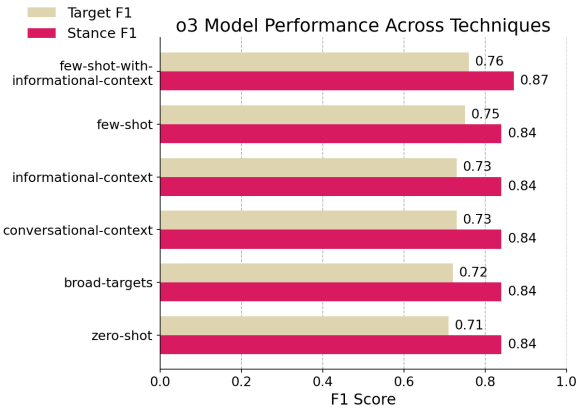

Figure 2: Target and Stance F1 scores for o3 model across different prompting strategies.

Overall, our results indicate that larger, proprietary models perform best under all conditions on both target and stance detection, and that some prompting strategies, especially few-shot combined with contextual information, can noticeably improve target identification without sacrificing stance detection.

4. Qualitative Analysis

This section presents a qualitative analysis of the top-performing model–technique pair, o3 with few-shot prompting and informational context. We examine errors on the gold-labeled dataset and the model's behavior on posts where annotators disagreed, highlighting common misclassifications and its informative role in challenging cases.

4.1. Error Analysis

We manually reviewed cases where o3 predictions disagreed with gold labels. This review is interpretive and not intended as a statistical evaluation.

Two main causes of target misclassification emerged. Table 9 shows three rephrased exam-

ple posts where the o3 model made a prediction error. First, some posts mention multiple plausible targets, with annotators and the model sometimes emphasizing different ones. Second, o3 occasionally used titles and linked text to infer targets, while annotators focused on the post body (Ex. 2). For stance, o3 often interpreted weak cues as indicative of a stance, whereas annotators relied on stronger signals (Ex. 3).

#	Post	Labels	Note
1	I've never had an issue with funding social programs, especially compared to the money Republicans pouring into the war machine.	**o3:** Republican Party, Negative **Gold:** Social spending, Positive	Focus on alternate target
2	[Ongoing.]www.cnn.com/politics-flynn-russia-calls-investigation	**o3:** Belief in Trump-Russia Collusion, Neutral **Gold:** None, N/A	Target inferred by URL text
3	Why the Democratic Party doesn't treat campaign finance reform as a major issue?	**o3:** Democratic Party, Negative **Gold:** Democratic Party, Neutral	Using weak cues

Table 9: Example posts showing o3 prediction errors compared with gold labels. Posts are rephrased to protect anonymity.

Grouping targets into categories shows that performance is highest for more tangible or specific categories, such as *Key Politicians*, and lowest for abstract or diffuse categories, such as *Trust & Institutions* (Table 10). It should be noted that the sample size is too small for statistical interpretation.

Category	Count	F1	Accuracy
International Relations	3	1.00	1.00
Key Politicians	29	0.93	0.90
Key Divisive Issues	29	0.90	0.86
Political Groups	11	0.88	0.91
Policies & Socioeconomic Issues	50	0.86	0.82
No Target	50	0.75	0.60
Technology & Governance Issues	5	0.67	0.60
Trust & Institutions	20	0.60	0.50
Ethnic & Religious Groups	3	0.33	0.33

Table 10: Target identification performance across target categories.

Finally, we analyzed o3's performance over stance categories when it selected the same target as annotators. We find that o3 has the most difficulty with neutral labels (Table 11), mirroring annotators' experience (see Table 1).

4.2. Disagreement Cases

We also evaluated o3 on posts where annotators disagreed, representing particularly challenging cases. In a random sample of 20 such posts, o3 matched the expert label in 12, produced reasonable labels in 3, and was incorrect in 5. These findings suggest that using o3 as a third annotator could often yield labels consistent with expert judgment (see Table 12). When errors occurred, o3 tended to overinterpret cues, inferring stronger stances (Ex. 1).

True Stance	Count	F1	Accuracy
Positive	43	0.90	0.93
Negative	40	0.84	0.88
Neutral	36	0.76	0.69

Table 11: Performance of o3 by stance (for posts with matching target).

#	Post	Labels
1	It's not about raising the minimum wage. There are awful gaps where you lose your benefits if you earn more, so making less ends up better.	**o3:** Social spending, Negative **A1:** Minimum wage, Negative **A2:** Social spending, Neutral **EX:** Social spending, Neutral
2	Biden sees the Green New Deal as an important foundation for tackling the climate crisis we're facing.	**o3:** Joe Biden, Neutral **A1:** Joe Biden, Neutral **A2:** Green Energy, Neutral **EX:** Joe Biden, Neutral
3	Good thing about Libertarians is they don't chase popularity. Dems and Reps need government support for influence and funding.	**o3:** Democrats, Negative **A1:** Liberals, Positive **A2:** Democrats, Negative **EX:** Democrats, Negative

Table 12: Example posts from disagreement cases. Labels show o3 predictions, human annotator labels (A1, A2), and the expert label (EX).

5. Conclusion

This study evaluates LLM performance on Target-Stance Extraction, a key task for political opinion mining, with a focus on nuanced political discussions. It makes three primary contributions to computational social science and natural language processing. First, we show that LLMs can unpack complex political opinions through TSE, providing an efficient way to study how different beliefs interact in polarization dynamics. Effective application, however, still requires a clear codebook, prompts and understanding of the relevant targets and literature. Second, we present a modular and reproducible analysis framework that supports multiple prompting strategies—including zero-shot, few-shot, and context-augmented variants. This works with both proprietary and open-source models, enabling systematic comparisons across systems and configurations. Third, we release a dataset of 1,084 Reddit comments from `r/NeutralPolitics`, 200 with gold-standard labels. It includes detailed stance and target annotations across 138 distinct political issues relevant to US politics, providing a valuable benchmark for future research on online political com-

munication and LLM evaluation in high-context settings. Together, these contributions demonstrate that LLMs can serve as effective tools for extracting nuanced political opinions at scale. This may help researchers to gain deeper insight into the structure and evolution of political beliefs, offering new opportunities to study polarization and discourse dynamics in online communities.

Limitations

This study has a few limitations. First, our dataset is drawn exclusively from `r/NeutralPolitics`, a forum with complex, nuanced discussions. While this focus may limit generalizability, it also highlights the robustness of our methods in high-context, challenging texts and suggests that applying them to simpler political content could be even more straightforward.

Second, our approach relies on a curated list of targets. Future work could explore Open-Target Stance Detection (OTSD) (Akash et al., 2025), which identifies targets in a fully open-ended manner rather than guiding the model with a predefined list. While OTSD is valuable for exploratory studies, some degree of target standardization remains necessary for practical analysis and comparability. In use cases like ours, where researchers can identify relevant targets beforehand through the literature, this is not necessary.

Data Availability

The datasets, the codebook, the code, and other supplementary materials associated with this study are publicly available at the following link https://github.com/zgrtgy/llm-tse.

Ethical Statement

This project received ethical approval from the Ethical Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University. All example posts were rephrased to protect anonymity of users.

Acknowledgements

This study received funding from a research grant awarded by the Applied Data Science focus area at Utrecht University. Generative AI tools were used only for formatting issues and grammatical checks.

References

Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. 2025. Can Large Language Models Address Open-Target Stance Detection? (arXiv:2409.00222).

American National Election Studies. 2021. Anes 2020 time series study full release [dataset and documentation]. Available at https://www.electionstudies.org.

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10):1531–1542.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 15(1):11477.

Pedro Brum, Matheus Cândido Teixeira, Renato Vimieiro, Eric Araújo, Wagner Meira Jr, and Gisele Lobo Pappa. 2022. Political polarization on Twitter during the COVID-19 pandemic: A case study in Brazil. *Social Network Analysis and Mining*, 12(1):140.

Michael Burnham. 2024. Stance Detection: A Practical Guide to Classifying Political Beliefs in Text. (arXiv:2305.01723).

Kareem Darwish. 2020. Quantifying Polarization on Twitter: The Kavanaugh Nomination. (arXiv:2001.02125).

Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. 2024. General social survey 1972-2024 [machine-readable data file]. Principal Investigator: Michael Davern; Co-Principal Investigators: Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. Sponsored by the National Science Foundation. Data accessed from the GSS Data Explorer website at https://gssdataexplorer.norc.org.

Daniel DellaPosta. 2020. Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization. *American Sociological Review*, 85(3):507–536.

James N. Druckman, Donald P. Green, and Shanto Iyengar. 2024. Does Affective Polarization Contribute to Democratic Backsliding in America? *The ANNALS of the American Academy of Political and Social Science*.

Venkata Rama Kiran Garimella and Ingmar Weber. 2017. A Long-Term Analysis of Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):528–531.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance Detection in Web and Social Media: A Comparative Study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87. Springer International Publishing.

Valentin Hofmann, Hinrich Schütze, and Janet B. Pierrehumbert. 2022. The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1259–1267.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(Volume 22, 2019):129–146.

Dilek Küçük and Fazli Can. 2021. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1):1–37.

Stephen Lacy, Brendan R. Watson, Daniel Riffe, and Jennette Lovejoy. 2015. Issues and Best Practices in Content Analysis. *Journalism & Mass Communication Quarterly*, 92(4):791–811.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A New Direction in Stance Detection: Target-Stance Extraction in the Wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365. Association for Computational Linguistics.

Jennifer McCoy, Tahmina Rahman, and Murat Somer. 2018. Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist*, 62(1):16–42.

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. Association for Computational Linguistics.

Lillio Mok, Michael Inzlicht, and Ashton Anderson. 2023. Echo Tunnels: Polarized News Sharing Online Runs Narrow but Deep. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:662–673.

N Newman, A Ross Arguedas, CT Robertson, RK Nielsen, and R Fletcher. 2025. Digital news report 2025. Technical report, Reuters Institute for the Study of Journalism.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A Challenge Dataset and Effective Models for Conversational Stance Detection. (arXiv:2403.11145).

Yunus Emre Orhan. 2022. The relationship between affective polarization and democratic backsliding: Comparative evidence. *Democratization*, 29(4):714–735.

Xingyu Peng, Zhenkun Zhou, Chong Zhang, and Ke Xu. 2024. Online Social Behavior Enhanced Detection of Political Stances in Tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1207–1219.

Kai Ruggeri, Friederike Stock, S. Alexander Haslam, Valerio Capraro, Paulo Boggio, Naomi

Ellemers, Aleksandra Cichocka, Karen M. Douglas, David G. Rand, Sander van der Linden, Mina Cikara, Eli J. Finkel, James N. Druckman, Michael J. A. Wohl, Richard E. Petty, Joshua A. Tucker, Azim Shariff, Michele Gelfand, Dominic Packer, Jolanda Jetten, Paul A. M. Van Lange, Gordon Pennycook, Ellen Peters, Katherine Baicker, Alia Crum, Kim A. Weeden, Lucy Napper, Nassim Tabri, Jamil Zaki, Linda Skitka, Shinobu Kitayama, Dean Mobbs, Cass R. Sunstein, Sarah Ashcroft-Jones, Anna Louise Todsen, Ali Hajian, Sanne Verra, Vanessa Buehler, Maja Friedemann, Marlene Hecht, Rayyan S. Mobarak, Ralitsa Karakasheva, Markus R. Tünte, Siu Kit Yeung, R. Shayna Rosenbaum, Žan Lep, Yuki Yamada, Sa-kiera Tiarra Jolynn Hudson, Lucía Macchia, Irina Soboleva, Eugen Dimant, Sandra J. Geiger, Hannes Jarke, Tobias Wingen, Jana B. Berkessel, Silvana Mareva, Lucy McGill, Francesca Papa, Bojana Većkalov, Zeina Afif, Eike K. Buabang, Marna Landman, Felice Tavera, Jack L. Andrews, Aslı Bursalıoğlu, Zorana Zupan, Lisa Wagner, Joaquín Navajas, Marek Vranka, David Kasdan, Patricia Chen, Kathleen R. Hudson, Lindsay M. Novak, Paul Teas, Nikolay R. Rachev, Matteo M. Galizzi, Katherine L. Milkman, Marija Petrović, Jay J. Van Bavel, and Robb Willer. 2024. A synthesis of evidence for policy from behavioural science during COVID-19. *Nature*, 625(7993):134–147.

Maryam Shafiei, Hossein Rahmani, Amirhossein Derakhshan, and Milad Allahgholi. 2025. Caskow: Context-Aware Stance Detection Using External Knowledge-Augmented LLM. In *2025 11th International Conference on Web Research (ICWR)*, pages 123–129.

Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. 2018. The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. *The International Journal of Press/Politics*, 23(1):95–115.

Anissa Tanweer, Emily Kalah Gade, P. M. Krafft, and Sarah Dreier. 2021. Why the Data Revolution Needs Qualitative Thinking. *Harvard Data Science Review*, 3(3).

Petter Törnberg. 2024. Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*.

Felicity M. Turner-Zwinkels, Jochem Van Noord, Rebekka Kesberg, Efrain García-Sánchez, Mark J. Brandt, Toon Kuppens, Matthew J. Easterbrook, Lien Smets, Paulina Gorska, Marta Marchlewska, and Tomas Turner-Zwinkels. 2023. Affective Polarization and Political Belief Systems: The Role of Political Identity and the Content and Structure of Political Beliefs. *Personality and Social Psychology Bulletin*.

Jochem Van Noord, Felicity M Turner-Zwinkels, Rebekka Kesberg, Mark J Brandt, Matthew J Easterbrook, Toon Kuppens, and Bram Spruyt. 2024. The nature and structure of European belief systems: exploring the varieties of belief systems across 23 European countries. *European Sociological Review*.

Veniamin Veselovsky and Ashton Anderson. 2023. Reddit in the Time of COVID. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:878–889.

Stefan Sylvius Wagner, Maike Behrendt, Marc Ziegele, and Stefan Harmeling. 2024. The Power of LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions. (arXiv:2406.12480).

Muheng Yang, Xidao Wen, Yu-Ru Lin, and Lingjia Deng. 2017. Quantifying Content Polarization on Twitter. In *2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, pages 299–308. IEEE.

Xinshu Zhao, Jun S. Liu, and Ke Deng. 2013. Assumptions behind Intercoder Reliability Indices. *Communication Yearbook*, 36(1):419–480.

Appendix A. Gold Dataset Target Counts and Category Mappings

Category	Target	Count
International Relations	Foreign military interventions by the US	1
	United Nations	1
	Israel	1
Key Divisive Issues	Gun Control	14
	Immigration	5
	LGBTQ rights	3
	Religion (general)	3
	Belief in Climate Change	2
	Abortion	2
Ethnic & Religious Groups	Muslims	1
	Hispanics	1
	Christians	1
Key Politicians	Donald Trump	12
	Hillary Clinton	5
	Bernie Sanders	3
	Barack Obama	2
	G.W. Bush	2
	Ron Paul	1
	Ted Cruz	1
	Joe Biden	1
	Elizabeth Warren	1
	Ronald Reagan	1
None	None	50
Policies & Socioeconomic Issues	Public healthcare	12
	Restrictions on Voting Access	8
	Social spending	6
	Minimum wage/Higher minimum wage	4
	Universal basic income	4
	International Trade	3
	Tax cuts (general)	3
	Student loan forgiveness	2
	Military spending	2
	Private healthcare	2
	Capital Punishment	1
	Protectionism	1
	Police spending	1
	Government spending	1
Political Groups	Republican Party	2
	Democratic Party	2
	Black Lives Matter movement	1
	Democratic politicians	1
	Republicans	1
	Republican politicians	1
	Democratic politician (generic)	1
	Antifa	1
	Republican politician (generic)	1
Technology & Governance Issues	Net Neutrality	2
	Use of AI in the workplace	1
	Social Media Regulations	1
	Increased Surveillance	1
Trust & Institutions	Trust in the US electoral system	7
	Trust in the US government (general)	4
	Trust in the US Military	2
	Trust in the Obama administration	2
	Trust in the US judicial system/courts	2
	Trust in the US Supreme Court	1
	Trust in the Trump administration	1
	Trust in the Biden administration	1

Table 13: Target counts and category mappings for the gold dataset (N = 200), ordered by count descending within each category. Covers 58 of 138 codebook targets. The full target list with category mappings is provided as supplementary material.

Appendix B. Detailed to Broader Labels Mappings

Broad Label	Detailed Label	Broad Label	Detailed Label
Republicans/Conservatives	Republican Party	Joe Biden	Trust in the Biden administration
	Republican politicians		Joe Biden
	Republican politician (generic)	Donald Trump	Trust in the Trump administration
	Republicans		Donald Trump
Democrats/Liberals	Democratic Party	Barack Obama	Trust in the Obama administration
	Democratic politicians		Barack Obama
	Democratic politician (generic)	Trust in the US Military	Trust in the US Military
Antifa	Antifa	Trust in the US electoral system	Trust in the US electoral system
Abortion	Abortion	Restrictions on Voting Access	Restrictions on Voting Access
Gun Control	Gun Control	Trust in the US judicial system/courts	Trust in the US judicial system/courts
LGBTQ rights	LGBTQ rights		Trust in the US Supreme Court
Belief in Climate Change	Belief in Climate Change	Capital Punishment	Capital Punishment
Immigration	Immigration	Social Media Regulations	Social Media Regulations
Muslims	Muslims	Use of AI in the workplace	Use of AI in the workplace
Christians	Christians	Religion (general)	Religion (general)
Hispanics	Hispanics	Net Neutrality	Net Neutrality
Progressive socioeconomic policies	Government spending	United Nations	United Nations
	Minimum wage/Higher minimum wage	Foreign military interventions by the US	Foreign military interventions by the US
	Social spending	Israel	Israel
	Public healthcare	Black Lives Matter movement	Black Lives Matter movement
	Student loan forgiveness	George W. Bush	George W. Bush
Conservative socioeconomic policies	Tax cuts (general)	Ronald Reagan	Ronald Reagan
	Private healthcare	Hillary Clinton	Hillary Clinton
	Military spending	Bernie Sanders	Bernie Sanders
	Police spending	Elizabeth Warren	Elizabeth Warren
Universal basic income	Universal basic income	Ted Cruz	Ted Cruz
International Trade	International Trade	Ron Paul	Ron Paul
	Protectionism	Increased Surveillance	Increased Surveillance
Trust in the US government (general)	Trust in the US government (general)	None	None

Table 14: Mapping of detailed targets to broad labels used in the broad labels prompting variant.

Appendix C. Prompts and Few-Shot Examples

All prompting strategies share the same base structure. We present the zero-shot prompt in full below, followed by a description of the modifications applied in each variant. Table 15 summarizes the differences.

Variant	Change from zero-shot baseline
Few-shot	Adds `Examples` section (one per stance) before the target list
Broad labels	Replaces fine-grained target list with merged categories (see Appendix B)
Conversational context	Restructures input format to include surrounding posts; adds `Context Weighting` section
Few-shot with informational context	Combines few-shot examples with per-target descriptions from the codebook

Table 15: Prompt variants and their differences from the zero-shot baseline.

C.1 Zero-Shot Prompt (Baseline)

```
You will be provided with a Reddit
comment and the title of the submission
under which it was posted.

Your task is to identify the **target**
and the **stance** expressed toward
it in the Comment, using the predefined
list of political targets at the end.

Only classify stances related to a
target from the list. If the Comment
refers to a similar target with
different wording, select the exact
matching
entry from the list.
**Do not rename, paraphrase, or invent
new target names.**
- If the post refers to a target not on
the list, label it **"Other"**.
- If no political target is mentioned,
label it **"None"**.

Use the Submission **only** to resolve
ambiguity (e.g., resolving pronouns
or vague references). **Do not infer
stance from the Submission.**

---

## Classification Steps

1. **Identify the Political Target**
   - Choose a target from the predefined
    list at the end.
   - If the post refers to a target not
   on the list, label it **"Other"**.
   - If no target is mentioned, label it
    **"None"** and skip stance
     classification.

2. **Determine the Stance** (only if a
target is identified)
   - **Positive**: clear support or
   praise
   - **Neutral**: mention without a
   clear stance
   - **Negative**: criticism or
   disapproval
   - **None**: no stance can be
   identified

3. **Return the classification**

---

## Output Format

Your response must be a JSON object with
 the following fields:
- "target" (string): The identified
political target, from the list of
  Predefined Targets.
- "stance" (string): The determined
stance. Must be one of: "Positive",
  "Neutral", "Negative", or "None".
- "confidence" (float): A confidence
score between 0.0 and 1.0 (inclusive)
  representing how certain the model is
  about its classification.

Example JSON output if a target and
stance are identified:
{
  "target": "Donald Trump",
  "stance": "Positive",
  "confidence": 0.95
}

---

## Predefined Targets

{predefined_targets}

(Use the exact wording. Do not rename,
paraphrase, or invent targets.)
```

C.2 Few-Shot Variant

The few-shot prompt is identical to the zero-shot baseline, with one addition: an `Examples` section inserted between the output format specification and the predefined targets list. It contains three labeled examples, one per stance category, drawn from `r/NeutralPolitics` but

held out from the annotation data. The examples are shown below.

```
## Examples

# Example 1
Input:
    Submission Title: New York Primary
    Results Megathread
    Comment: Didn't Sanders vote for
    that crime bill too? Has he evolved
    on that position or does he still
    support it?
Output:
{
  "target": "Bernie Sanders",
  "stance": "Neutral",
  "confidence": 0.8
}

# Example 2
Input:
    Submission Title: Flat-tax in the U.
    S. - a good idea?
    Comment: I'd say it's a good thing,
    but the reason for existing tax
    breaks is to encourage people to
    live in a way that is good for
    society; get educated, own property,
     have kids, etc.
Output:
{
  "target": "Tax cuts (general)",
  "stance": "Positive",
  "confidence": 0.90
}

# Example 3
Input:
    Submission Title: Is drug
    legalization/decriminalization sound
     policy?
    Comment: Treating drug abuse as a
    medical problem instead of a
    criminal
    problem has been successful in
    European countries and has been a
    conclusion by various studies.
Output:
{
  "target": "War on Drugs",
  "stance": "Negative",
  "confidence": 0.90
}
```

C.3 Broad Labels Variant

The broad labels prompt is identical to the zero-shot baseline. The only change is in the content of the `{predefined_targets}` placeholder: fine-grained targets are replaced with broader labels. The full mapping from detailed to broad labels is provided in Appendix B.

C.4 Conversational Context Variant

The conversational context prompt differs from the zero-shot baseline in two ways. First, the input description is expanded to accommodate multiple post types from the same thread: the comment and submission title are supplemented with up to four surrounding posts, each tagged by type (`Submission`, `Parent`, `Focus`, `Children`, `Ancestor`, `Earlier_Sibling`, `Later_Sibling`). Second, a `Context Weighting` section is added after the input description, specifying the order in which context types should be prioritized when resolving ambiguity:

```
## Context Weighting

When using context, prioritize in the
following order:
1. Children
2. Parent
3. Ancestors
4. Earlier and Later Siblings
```

The classification instructions, output format, and target list are otherwise unchanged. The stance is always inferred from the Focus post alone; context is used only to resolve co-references and ambiguous phrasings.

C.5 Few-Shot with Informational Context

This variant combines the additions from Section C.2 and appends descriptions of each target from the annotation codebook directly to the target list. The descriptions follow the exact wording used during annotation training, and are intended to resolve ambiguity between semantically similar targets. The few-shot examples are identical to those in Section C.2.