Target Stance Extraction with LLMs in Complex Political Discussions

No Author Given

No Institute Given

Abstract. Target-Stance Extraction (TSE) is a natural language processing task that involves identifying both the entity or issue discussed in a text and the stance expressed toward it. Unlike traditional stance detection, which assumes a small list of of targets, TSE enables more flexible and fine-grained analysis of political discourse. With recent advances in large language models (LLMs), it is now feasible to approach this task without extensive annotation or domain-specific training. In this study, we evaluate the capabilities of proprietary and open-source LLMs across multiple sizes and prompting strategies. Our findings show that LLMs can reliably perform TSE with minimal supervision, offering a scalable alternative to existing methods.

Keywords: target stance extraction · stance detection · political polarization · large language models · natural language processing

1 Introduction

Political polarization has become an urgent concern in contemporary democracies, with online platforms now serving as the primary venue for news consumption and information exchange [9]. While social media enables wide participation in political discourse, their open and fragmented nature has been linked to rising political polarization [17, 1], posing a growing threat to social cohesion and democratic politics. To understand how polarization emerges and spreads in these environments, researchers require tools that can extract political attitudes at scale and with nuance.

Stance detection is one such tool. As defined by Burnham [4], stance detection is the task of assessing whether a text entails a particular stance toward a target, such that a reader, given contextual information, would infer the stance. By identifying opinions users express about political figures, policies, or issues, stance detection allows researchers to map belief systems and measure polarization beyond simple partisan labels. However, most prior approaches rely on narrow sets of predefined targets or partisan proxies, limiting their ability to capture the full range of political attitudes. In addition, supervised models typically require costly and time-consuming annotation efforts and often struggle to generalize across topics or domains.

Recent advances in large language models (LLMs) offer a promising direction for automated political text analysis. LLMs have demonstrated strong performance in related tasks such as sentiment analysis and content classification [**?**,**?**],

and can readily adapt to new contexts with minimal supervision. For stance detection, they hold the potential to expand the scope of analysis while providing more granular insights by identifying both the stance and the specific target it is directed at. By detecting particular beliefs and attitudes, LLMs enable a deeper understanding of how polarization manifests in everyday online discourse.

This study proposes a more flexible and fine-grained framework for detecting political beliefs through Target-Stance Extraction (TSE), which jointly identifies a stance and its target [13]. To evaluate LLM performance on this task, we choose Reddit, a major social media platform where discussions are organized into "subreddits" dedicated to specific themes or interests. We use a manually annotated dataset from `r/NeutralPolitics`, a highly moderated, nonpartisan subreddit that promotes civil and evidence-based political debate. The dataset includes over 1,000 posts with fine-grained stance and target labels. We test a range of proprietary and open-source LLMs of varying sizes and prompting strategies on this dual task. Our results show that LLMs can reliably identify political beliefs in natural conversations without relying on restrictive predefined target lists, enabling scalable annotation and offering new tools for analyzing polarization beyond partisan lines.

2 Literature

Political polarization has been conceptualized in various ways, with no consensus on a unified framework [11]. A useful operationalization by Yarchi et al. [21] identifies three core dimensions: ideological polarization (divergence in policy preferences), interactional or structural polarization (the homogeneity of interactions), and affective polarization (emotional hostility between political groups). These forms of polarization have been linked to democratic backsliding, rising social distrust, and declining institutional accountability[14]. Yet the mechanisms driving polarization and its diffusion remain insufficiently understood.

Social media have become a central arena for political discussion, activism, and ideological conflict. Unlike traditional media, they enable participatory, real-time communication where users act as both producers and consumers of content. This makes it a uniquely valuable environment for studying political discourse. Prior research shows that platforms like Facebook and Twitter foster affective polarization, echo chambers, and ideological segregation [**?**,**?**]. Reddit, by contrast, is organized around topic-specific "subreddits," which allow for more modular and theme-based discussions. Its voting mechanisms and community norms can sometimes support more cross-cutting or deliberative engagement [**?**,**?**]. Nonetheless, growing polarization has also been observed on Reddit, especially during election cycles and in politically charged communities [**?**].

Despite a growing body of research, several limitations persist. Much of the research centers on Twitter (now X), largely due to its historical data accessibility[3]. This platform-specific focus limits the generalizability of findings. Additionally, many studies infer political beliefs using coarse indicators—such as user engagement patterns or partisan labels—that fail to capture the issue-specific,

contextual nature of political attitudes[6]. This is particularly limiting on Reddit, where users often participate in a wide variety of topic-driven communities. Existing work also tends to focus on short timeframes or political flashpoints like elections, leaving long-term and cross-domain dynamics understudied.

Methodological challenges exacerbate these issues. Survey-based approaches are costly, prone to social desirability bias, and fail to capture fringe communities. While supervised machine learning offers greater scalability, it requires domain-specific labeled datasets that are expensive to produce and often fail to generalize across contexts [7]. Many supervised stance detection models rely on overt partisan cues—such as hashtags or slogans—that are rarely present in more organic, issue-based discussions [12]. Network models have been used to capture polarization through structural patterns, but these typically reflect interactional dynamics rather than the expressed beliefs or stances themselves [16].

Recent advances in large language models (LLMs) offer new tools for overcoming these limitations. LLMs have been employed to simulate user behavior on social media [19], generate synthetic data to improve stance detection [20], and match or exceed human annotators in tasks such as sentiment analysis and political classification [2, **?**]. Despite these advances, most existing work on stance detection remains focused on benchmark datasets with a limited set of predefined targets, restricting its applicability in real-world settings(placeholder).

Addressing these limitations, Li et al. [13] introduced the Target-Stance Extraction (TSE) task, combining the identification of both the stance and its specific target under a single task. By removing the reliance on predefined target lists, TSE offers greater scalability and flexibility for real-world applications. Their proposed framework uses separate components for target identification (via classification or generation) and stance classification, both relying on fine-tuned models such as BERT and BART. However, recent advances in instruction-tuned, general-purpose LLMs now make it possible to integrate these unify these steps—without the need for task-specific training—as these models inherently encode extensive domain knowledge and exhibit greater robustness across diverse contexts. Building on this insight, we propose a prompting-based approach to TSE using general-purpose LLMs and evaluate their performance on real-world data from Reddit. Our work contributes a novel pipeline for detecting political stances with greater granularity and contextual sensitivity, advancing the use of LLMs for political text analysis in real-world settings.

3 Methodology and Data

Although recent studies have begun exploring the capabilities of large language models, our approach is, to our knowledge, the first to apply instruction-tuned, general-purpose LLMs (such as GPT-4 and Gemini 2.5) to the full TSE task. Unlike prior work using encoder-decoder models like BERT or BART, which require task-specific fine-tuning, our method relies solely on prompting. This enables us to address a broader range of stance expressions—including subtler or more context-dependent forms—and to apply TSE to the domain of political po-

larization, where targets and stances are highly variable and often implicit. This approach raises several important questions, which we may not fully resolve here but aim to highlight and explore through our methodology. First, reproducibility remains a challenge, as many state-of-the-art models are proprietary or costly to use. To address this, we compare a range of models—from proprietary flagships to open-source models of varying sizes, including large models that require high-performance servers; mid-sized models suitable for small research clusters; and smaller models compatible with personal computers. Second, we evaluate a set of prompting and augmentation strategies, including zero-shot prompting, few-shot prompting, coarse-grained labeling (reduced granularity), and contextual augmentation (feeding the comment thread for richer context), to assess whether they improve model performance. Together, these comparisons provide a structured foundation for assessing the feasibility, reliability, and scalability of prompting-based TSE in political domains.

3.1 Data

Our evaluation focuses on samples drawn from the `r/NeutralPolitics` subreddit—a rare nonpartisan political forum on Reddit known for evidence-based, highly moderated discussion [1]. Unlike the emotionally charged interactions typical of Twitter/X, `r/NeutralPolitics` enforces strict rules, including civility guidelines and source requirements, which make it well-suited for studying nuanced political stances. We retrieved historical posts using the Pushshift Archives (link), which span Reddit content from 2005 onward. We used April 2023 as the cutoff date, as Reddit changed its API terms, leading to a temporary halt in archiving and a transition in maintainers. The dataset remains publicly available and widely used in academic research [15, **?**]. From this archive, we extracted all posts from `r/NeutralPolitics`, which includes 578,041 comments under N submissions. Of these, 130.390 were marked as removed—likely by subreddit moderators enforcing community rules.[2]

3.2 Annotation

We created an annotated evaluation dataset from `r/NeutralPolitics` to assess model performance on TSE. The main goal was to obtain a diverse and representative set of annotations reflecting the variety of political targets and expressions in the subreddit.

Three research assistants with knowledge of U.S. politics and social media were trained over three weeks. During training, annotators were provided with a preliminary codebook outlining frequent and polarizing political targets identified from relevant literature [5, **?**]. However, consistent with the open coding

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

[2] This removal rate is higher than the 11% reported in [8], likely due to the subreddit's strict moderation policies.

approach proposed by Tanweer et al. [18], they were instructed to label comments openly, allowing new targets to emerge inductively from the data.

The sampling strategy evolved to improve relevance and balance. Initial random sampling yielded many non-political posts. Subsequent rounds filtered samples using a GPT-4o-mini classifier to select political comments, which were further stratified by stance (positive, neutral, negative) to ensure coverage across categories. Submission bodies were excluded in later rounds due to their variable utility and annotation effort. This final refined sampling strategy was adopted for the main annotation phase, which included four rounds.

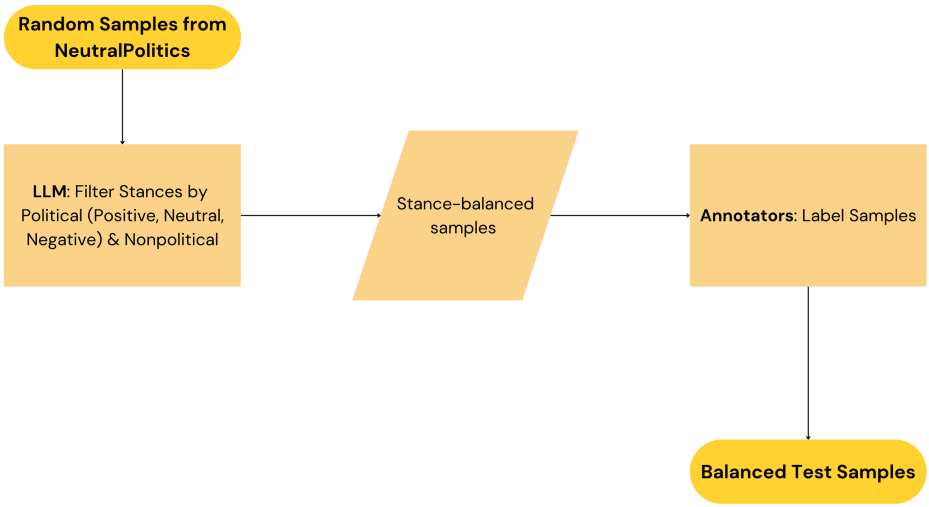

Fig. 1: Annotation Process

Each comment was labeled independently by two assistants to maximize coverage. Labels with agreement between both assistants, and confirmed by an expert, were included in the evaluation dataset for LLM assessment.

Our final codebook includes 138 distinct political targets—significantly more granular than prior studies. Interrater reliability, measured by Krippendorff's alpha, was 0.48 overall. This moderate score reflects the inherent difficulty and nuance of the task. Agreement improved across training rounds as the codebook stabilized. Disagreements were most common on nuanced stance distinctions and multi-target comments.

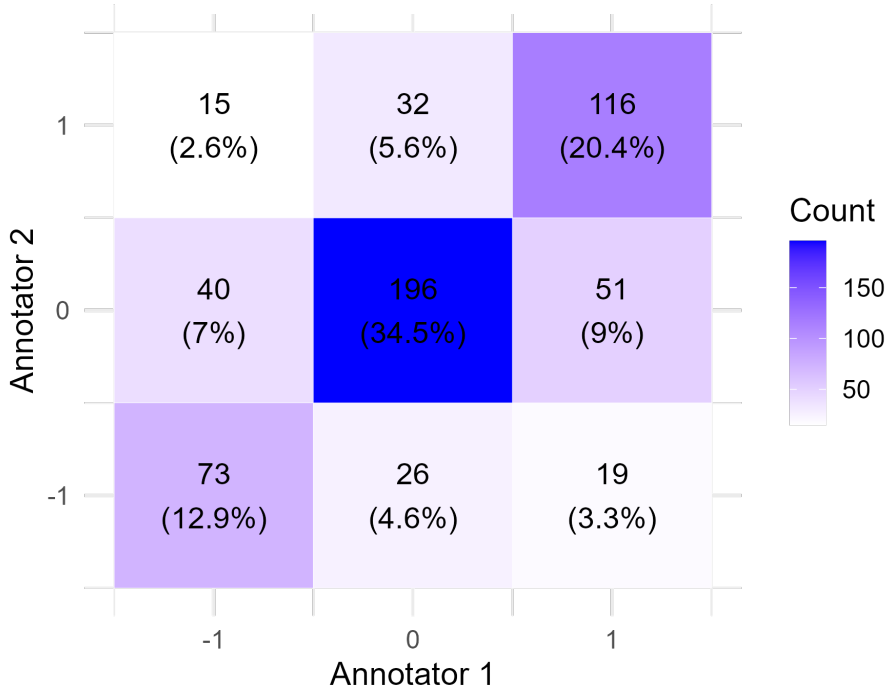

Fig. 2: Inter-rater Agreement Matrix

3.3 Model Configuration and Access

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 comply with GDPR and explicitly state that user inputs sent through their APIs are not used for training purposes [3,4], mitigating concerns over data leakage.

All models were run with a fixed seed and temperature set to 0.0. except for o3, which does not support a temperature of 0.0 because such low values

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

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

[4] Popular open-weight models such as *Deepseek-V3* and *Deepseek-R1* were excluded as Deepseek does not explicitly guarantee that API-submitted data will not be retained or used for training. While we were able to locally deploy Deepseek's distilled *Qwen3-8b* model to avoid this issue, deploying larger Deepseek variants was unfeasible due to cost and infrastructure constraints.

tend to hinder reasoning capabilities. It should be noted that this does not guarantee fully deterministic outputs; however, averages taken over three runs show minimal variation in metrics. The deployed Qwen models had reasoning disabled, as enabling it did not improve performance and significantly increased costs. Additionally, Qwen models used FP8 quantization, which does not affect inference quality [10, **?**].

4 Analysis

To evaluate model performance, we implemented a modular stance detection pipeline. Our experiments explored several configurations:

- zero-shot prompting as the baseline,
- few-shot prompting with examples for each stance,
- including informational context for augmentation,
- including conversational context for augmentation,
- using broader target labels for different levels of detail.

This layered design enables an in-depth investigation of LLM capabilities and limitations in interpreting political attitudes within complex, real-world discourse.

We evaluated a range of LLMs on a balanced subset of 200 samples from our annotated dataset on TSE, ensuring equal representation of each stance class (Negative, Neutral, Positive, None). Targets were not equally distributed, as equally sampling 138 targets over four stance categories is infeasible. The following sections summarize key findings from these experiments across different strategies.

4.1 LLM Strategies for Target-Stance Extraction

Zero-shot Prompting (Baseline) We use zero-shot results as a baseline to compare model performance across different families and sizes. Proprietary models like *GPT-4.1*, *O3*, and *Gemini-2.5* show strong and consistent performance, with minor drops in their smaller variants, achieving stance detection scores above 0.80. although target detection remains notably lower.

Reasoning-enabled models do not improve target identification but consistently outperform non-reasoning counterparts on stance classification. Target detection scores are generally 0.1 to 0.2 points lower than stance scores, reflecting the greater difficulty of accurately identifying political targets. This gap widens as model size decreases.

Performance declines sharply when moving from 8B to 4B parameters and becomes unacceptable for models below 2B. Meanwhile, gains between 8B and 70B models are modest, with Qwen leading among similar sizes, and distillation from Deepseek R1 shows no clear benefit. The significant jump seen with Llama-3.1-405B suggests that richer embedded knowledge substantially aids target identification, also explaining the poor performance of smaller models.

A key qualitative observation from manual review of results is that many disagreements between model outputs and gold labels stem from labeling inconsistencies rather than conceptual errors. For instance, a model might label a target as "Trust in the US Military" while the gold annotation reads "Stance toward the US Military." Both refer to the same underlying concept but differ in phrasing. This suggests that, despite modest F1 scores, high-performing models often correctly capture the semantic focus of a statement.

While these results fall short of state-of-the-art performance on benchmark datasets, they remain promising given the general-purpose, zero-shot setup and the open-ended nature of both target and stance identification in our task.

To address these limitations and explore potential gains, we next evaluate a set of strategies designed to improve performance across model sizes.

Table 1: Zero-shot TSE metrics, ordered by Target Accuracy. Best values for model categories are in bold.

Model	Target Acc.	Target F1	Stance Acc.	Stance F1
gpt-4.1	**0.70**	**0.73**	0.81	0.81
o3	0.68	0.71	**0.85**	**0.84**
gemini-2.5-pro	0.67	0.71	**0.84**	**0.84**
gemini-2.5-flash	0.66	0.70	**0.84**	**0.84**
llama-3.1-405b	**0.66**	**0.66**	**0.80**	**0.80**
gpt-4.1-mini	0.61	0.63	0.79	0.79
llama-3.1-70b	0.60	0.59	0.78	0.78
llama-4-maverick	0.58	0.59	0.63	0.63
qwen-3-32b-fp8	**0.58**	**0.61**	**0.73**	0.73
qwen-3-8b-fp8	0.56	0.58	0.72	0.72
qwen-3-14b-fp8	0.55	0.54	0.77	0.76
gemini-2.5-flash-lite	0.55	0.57	0.75	0.74
deepseek-r1-0528-qwen3-8b	0.51	0.50	0.71	0.70
gemma-3-27b	0.50	0.55	0.73	0.73
gemma-3-12b	0.49	0.51	0.72	0.72
llama-4-scout	0.49	0.50	0.68	0.69
llama-3.1-8b	0.34	0.36	0.68	0.67
gemma-3-4b	0.27	0.28	0.57	0.53
qwen-3-1.7b-base	0.17	0.13	0.48	0.46
qwen-3-0.6b-base	0.06	0.04	0.36	0.35
gemma-3-1b	0.02	0.02	0.67	0.53

Context Augmentation Providing contextual information is generally expected to improve LLM performance by offering additional cues for interpretation. In the case of social media conversations, we identify two relevant types of context: informational context, which refers to background knowledge about the topic being discussed, and conversational context, which includes the surround-

ing thread or dialogue in which a post appears. This is particularly important on platforms like Reddit, where posts are embedded within extended threads.

Informational Context. We tested the effects of informational context by augmenting the task input with short textual descriptions of political targets, taken directly from the annotation codebook. These descriptions aim to resolve ambiguity and standardize understanding of targets, especially for models with smaller context windows or weaker world knowledge.

As shown in Table 2, this strategy consistently improves target identification across all models. Both accuracy and F1 scores for target classification increase—often substantially. Stance classification, however, shows mixed results. While some models improve on stance metrics as well, others exhibit slight degradations in stance metrics. This may be due to a shift in the model's focus toward interpreting the augmented target explanation at the expense of capturing the author's stance in the original post.

Table 2: TSE metrics using **informational context**, excluding models with overall low performance. Increases from zero-shot performance are in **bold**, decreases are marked in red.

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

Conversational Context. To evaluate conversational context, we appended up to four surrounding posts from the same Reddit thread to each example. Priority was given to the parent post (i.e., the one being replied to), followed by replies to the focal post and sibling comments. This approach aims to help models resolve coreferences and better infer the author's intent. Full prompt formatting details are provided in Appendix X.

However, results are mixed (Table 3). While conversational context can be intuitively useful, several models—particularly those in the GPT-4.1 family—experience a notable decline across all four metrics when thread history is included. These results suggest that even high-performing models may be sensitive to increased prompt length or to off-topic distractors introduced by the additional posts.

At the same time, some models benefit from the added context. The Gemini 2.5 family shows consistent gains, especially in target identification, as do Llama 4 Maverick and o3. Overall, while thread context seem to enhance performance for models with stronger reasoning capabilities, it is not universally helpful and can incur substantial computational cost due to increased token counts.

Table 3: TSE metrics using **conversational context**.

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

Broader Target Labels Our codebook defines highly specific political targets, which can pose challenges for generalization. To address this, we test whether grouping fine-grained targets into broader categories improves performance. Table 4 shows representative examples.[5] This simplification may reduce ambiguity and improve model accuracy.

We find that broader target labels often lead to improvements in target classification metrics, particularly among larger models, as shown in Table 5. Stance performance shows more mixed results: while some models maintain or slightly improve their scores, others—especially smaller models—experience a decline. This suggests that abstraction aids target identification but may introduce ambiguity that hampers stance detection in less capable models.

[5] We did not consolidate key issues such as *Abortion* and *Gun Control*, even though they are strongly associated with particular ideologies, as they represent major points of political conflict.

Table 4: Examples of fine-to-broad label mappings used in the broad-labels evaluation. Each broad label aggregates multiple related targets.

Default Label (Target)	**Broad Label**
Republican Party	Republicans/Conservatives
Republican politicians	Republicans/Conservatives
Conservatives	Republicans/Conservatives
...	
Democratic Party	Democrats/Progressives
Democratic politicians	Democrats/Progressives
Progressives	Democrats/Progressives
...	
Social spending	Progressive socioeconomic policies
Higher corporate tax	Progressive socioeconomic policies
...	
Military spending	Conservative socioeconomic policies
Tax cuts for the wealthy	Conservative socioeconomic policies
...	

Table 5: TSE metrics using **broad target labels**.

Model	Target Acc.	Target F1	Stance Acc.	Stance F1
gemini-2.5-pro	**0.68**	**0.72**	**0.85**	**0.85**
gemini-2.5-flash	**0.70**	**0.74**	**0.86**	**0.86**
gemini-2.5-flash-lite	**0.59**	**0.60**	0.71	0.71
o3	**0.70**	**0.72**	0.85	0.84
gpt-4.1	**0.73**	**0.74**	**0.82**	0.81
gpt-4.1-mini	**0.64**	**0.65**	0.77	0.77
llama-4-maverick	**0.60**	**0.60**	**0.67**	**0.68**
llama-3.1-405b	**0.67**	**0.67**	0.80	**0.81**
llama-3.1-70b	**0.62**	**0.60**	**0.80**	**0.79**
qwen-3-32b-fp8	0.56	0.57	**0.77**	**0.76**
qwen-3-14b-fp8	**0.57**	**0.57**	0.75	0.74

Few-shot Prompting Comparing zero-shot and few-shot prompting results reveals a consistent pattern across most models. Target classification metrics—both accuracy and F1—show a clear and marked improvement under few-shot prompting, with increases typically ranging from about 0.03 to 0.08 points. This suggests that providing a few examples helps models better identify political targets in the text. In contrast, stance classification metrics (accuracy and F1) exhibit slight decreases or remain largely stable under few-shot prompting, generally dropping by only about 0.01 to 0.02 points or improving marginally. These small fluctuations indicate that stance detection is less affected by few-shot examples than

target classification, or possibly that the few-shot prompts emphasize target identification more directly. Overall, few-shot prompting improves target detection noticeably while maintaining comparable stance performance, making it a valuable and low-cost method for improving model output quality.

Table 6: TSE metrics using **few-shot prompting**.

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

To further test whether few-shot prompting can be strengthened without significantly increasing cost, we introduced additional informational context alongside the examples.

Table 7: TSE metrics using **few shot prompting with informational context**. Increases from **few-shot** performance are in **bold**, decreases are marked in red.

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

The results show that the effects of informational context on target classification are mixed: while some models benefit modestly, others — particularly those in the Gemini family — exhibit slight regressions in target-related metrics. Its impact on stance classification is similarly model-dependent. Overall, this strategy yields clear stance improvements for stronger models, but its benefits for both target and stance classification vary by model.

5 Contributions

This study makes three primary contributions to computational social science and natural language processing. First, it introduces and systematically evaluates the use of large language models (LLMs) for the Target Stance Extraction (TSE) task, which involves jointly identifying both the stance expressed in a text and its specific target. To our knowledge, this is the first comprehensive application of instruction-tuned, general-purpose LLMs on TSE, offering a scalable and model-agnostic alternative to earlier supervised or task-specific approaches. Second, we develop a modular and reproducible evaluation pipeline that supports multiple prompting strategies—including zero-shot, few-shot, and context-augmented variants. The pipeline works with both proprietary and open-source models, enabling comparisons across systems and configurations. Third, we release a new dataset of 1,084 Reddit comments from `r/NeutralPolitics`, 200 of which have gold-standard labels. The dataset includes detailed annotations of stance and target, covering 138 distinct political issues relevant to US politics, and offers a valuable benchmark for future research on online political communication and LLM evaluation in high-context settings.

6 Limitations

This study has several limitations. First, while we evaluate a diverse set of LLMs, our comparisons are shaped by constraints in compute, API access, and the fast-changing nature of large language models. Second, the dataset is drawn exclusively from `r/NeutralPolitics`, a highly moderated subreddit focused on evidence-based, civil discussion, which may limit the generalizability of our findings to other platforms or contexts. Finally, although we test several prompting techniques, we do not fine-tune any models or perform detailed hyperparameter tuning, both of which could further improve performance and are left for future work.

References

1. Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences

115(37), 9216–9221 (Sep 2018). https://doi.org/10.1073/pnas.1804840115, https://www.pnas.org/doi/full/10.1073/pnas.1804840115, publisher: Proceedings of the National Academy of Sciences
2. Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., Jovančević, A.: Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm **15**(1), 11477. https://doi.org/10.1038/s41598-025-96508-3, https://www.nature.com/articles/s41598-025-96508-3
3. Bruns, A.: Big social data approaches in internet studies: The case of twitter. In: Hunsinger, J., Allen, M.M., Klastrup, L. (eds.) Second International Handbook of Internet Research, pp. 65–81. Springer Netherlands. https://doi.org/10.1007/978-94-024-1555-1_3, https://doi.org/10.1007/978-94-024-1555-1$_3$
4. Burnham, M.: Stance Detection: A Practical Guide to Classifying Political Beliefs in Text (May 2024), http://arxiv.org/abs/2305.01723, arXiv:2305.01723 [cs]
5. Davern, M., Bautista, R., Freese, J., Herd, P., Morgan, S.L.: General social survey 1972-2024 [machine-readable data file] (2024), 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.
6. DellaPosta, D.: Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization. American Sociological Review **85**(3), 507–536 (Jun 2020). https://doi.org/10.1177/0003122420922989, http://journals.sagepub.com/doi/10.1177/0003122420922989
7. Ghosh, S., Singhania, P., Singh, S., Rudra, K., Ghosh, S.: Stance Detection in Web and Social Media: A Comparative Study. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 75–87. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_4
8. Hofmann, V., Schütze, H., Pierrehumbert, J.B.: The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse **16**, 1259–1267. https://doi.org/10.1609/icwsm.v16i1.19377, https://ojs.aaai.org/index.php/ICWSM/article/view/19377
9. Institute, R.: Digital News Report 2021 | Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (2021)
10. Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., Xiong, D.: A comprehensive evaluation of quantization strategies for large language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 12186–12215. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.726, https://aclanthology.org/2024.findings-acl.726/
11. Kubin, E., Von Sikorski, C.: The role of (social) media in political polarization: A systematic review **45**(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070, https://www.tandfonline.com/doi/full/10.1080/23808985.2021.1976070
12. Küçük, D., Can, F.: Stance Detection: A Survey. ACM Computing Surveys **53**(1), 1–37 (Jan 2021). https://doi.org/10.1145/3369026
13. Li, Y., Garg, K., Caragea, C.: A New Direction in Stance Detection: Target-Stance Extraction in the Wild. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10071–10085. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.560

14. McCoy, J., Rahman, T., Somer, M.: Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. American Behavioral Scientist **62**(1), 16–42 (Jan 2018). https://doi.org/10.1177/0002764218759576, https://doi.org/10.1177/0002764218759576, publisher: SAGE Publications Inc
15. Mok, L., Inzlicht, M., Anderson, A.: Echo Tunnels: Polarized News Sharing Online Runs Narrow but Deep **17**, 662–673. https://doi.org/10.1609/icwsm.v17i1.22177, https://ojs.aaai.org/index.php/ICWSM/article/view/22177
16. Peng, X., Zhou, Z., Zhang, C., Xu, K.: Online Social Behavior Enhanced Detection of Political Stances in Tweets **18**, 1207–1219. https://doi.org/10.1609/icwsm.v18i1.31383, https://ojs.aaai.org/index.php/ICWSM/article/view/31383
17. Suhay, E., Bello-Pardo, E., Maurer, B.: The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. The International Journal of Press/Politics **23**(1), 95–115 (Jan 2018). https://doi.org/10.1177/1940161217740697, https://doi.org/10.1177/1940161217740697, publisher: SAGE Publications Inc
18. Tanweer, A., Gade, E.K., Krafft, P.M., Dreier, S.: Why the Data Revolution Needs Qualitative Thinking. Harvard Data Science Review **3**(3) (Jul 2021). https://doi.org/10.1162/99608f92.eee0b0da
19. Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. https://doi.org/10.48550/arXiv.2310.05984, http://arxiv.org/abs/2310.05984
20. Wagner, S.S., Behrendt, M., Ziegele, M., Harmeling, S.: The Power of LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions (Jun 2024)
21. Yarchi, M., Baden, C., Kligler-Vilenchik, N.: Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. Political Communication **38**(1-2), 98–139 (Mar 2021). https://doi.org/10.1080/10584609.2020.1785067, https://doi.org/10.1080/10584609.2020.1785067, publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1785067

Acknowledgments. This study received funding from research grants awarded to the ******* Generative AI models have been employed for formatting and grammar checks.

Disclosure of Interests. The authors have no competing interests.

A Appendix A: Codebook with Targets, Explanations, and Broad Targets

Table 8: Codebook for Target Labels

Target	Explanation	Broad Target
Republican Party	Republican Party as an organization	Republicans/Conservatives
Republican politicians	Republican politicians as a group, including representatives, senators and governors	Republicans/Conservatives
Republican politician (generic)	a Republican politician not on the list.	Republicans/Conservatives
Republicans	Republican party supporters. A comment about both the Republican politicians and Republican party supporters would fall here. If it is only about Republican politicians, use the Republican politicians label.	Republicans/Conservatives
Conservatives	Conservative people and/or ideology	Republicans/Conservatives
Right-wingers	right-wing people and/or ideology	Right-wingers
Alt-right	alt-right people and/or ideology	Right-wingers
Independents	independent people (often used for voters without a party affiliation)	Independents
Democratic Party	Democratic Party as an organization	Democrats/Liberals
Democratic politicians	Democratic politicians as a group, including representatives, senators and governors	Democrats/Liberals
Democratic politician (generic)	a Democratic politician not on the list	Democrats/Liberals
Democrats	Democratic party supporters. A comment about both the Democratic politicians and Democratic party supporters would fall here. If it is only about Democratic politicians, use the Democratic politicians label.	Democrats/Liberals
Centrists	Centrist people and/or ideology	Democrats/Liberals
Liberals	Liberal people and/or ideology	Democrats/Liberals
Leftists	Leftist people and/or ideology	Leftists/Socialists
Greens	Green people and/or ideology	Greens

Continued on next page

Table – Continued from previous page

Target	**Explanation**	**Broad Target**
Socialists	Socialist people and/or ideology	Leftists/Socialists
Antifa	Antifa political group and its supporters	Antifa
Abortion	Planned Parenthood also almost always fall under here, unless they specifically focus on Planned Parenthood as a public healthcare organization.	Abortion
Gun Control		Gun Control
LGBTQ rights		LGBTQ rights
Vaccination		Vaccination
Belief in Institutional Racism		Racial equality
Belief in Climate Change		Belief in Climate Change
Immigration		Immigration
Affirmative Action for Racial Equality		Racial equality
Affirmative Action for Gender Equality		Gender equality
Muslims		Muslims
Jews		Jews
Catholics		Catholics
Protestants		Protestants
Christians		Christians
Atheists		Atheists
Whites		Whites
Blacks		Blacks
Hispanics		Hispanics
Asians		Asians
Indians		Indians
Government spending	Fiscal policies that favor an increased government budget Use the specific label if it refers to a higher budget for education, healthcare, military, security, or any other item on the list.	Progressive socioeconomic policies

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
Tax cuts (general)	Tax cuts that apply to everyone	Conservative socioeconomic policies
Tax cuts (for rich)	Tax cuts that apply to the rich	Conservative socioeconomic policies
Tax cuts (for low-middle classes)	Tax cuts that apply to the low and/or middle class	Progressive socioeconomic policies
Higher corporate tax	A comment calling for lower corporate tax rate would be labelled as Anti (so - 1) Higher Corporate Tax,	Progressive socioeconomic policies
Higher taxes for the rich		Progressive socioeconomic policies
Minimum wage/Higher minimum wage	establishing a minimum wage, or increasing it	Progressive socioeconomic policies
Social spending	social spending, except healthcare, education, Coronavirus stimulus checks and security Examples: unemployment or childcare benefits, public funding of parks etc.	Progressive socioeconomic policies
Universal basic income	Universal Basic Income (UBI) in which people regularly receive a minimum income without any checks on conditions.	Universal basic income
Public healthcare	public healthcare, including Affordable Care Act (Obamacare), Medicare, Medicaid, healthcare reforms that would require more government funding	Progressive socioeconomic policies
Private healthcare	private healthcare. Examples include calls for more privatization, removal of existing social security programs (ACA/Medicaid etc.)	Conservative socioeconomic policies
Public education	public education. Examples include calls for more government funding for schools	Progressive socioeconomic policies
Private education	private education. Examples include calls for more privatization	Conservative socioeconomic policies

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
Student loan forgiveness	whether student loans and debt should be forgiven.	Progressive socioeconomic policies
Government bailouts	government bailout of private corporations, a common example would be bank bailouts during the 2008 financial crisis.	Conservative socioeconomic policies
Military spending	the US military spending/budget	Conservative socioeconomic policies
Police spending	the US police/law enforcement spending/budget. This is often referred within the state level, but covers both state and federal budgets.	Conservative socioeconomic policies
International Trade	International Trade. Examples include Trans-Pacific Partnership, US trade with China, US trade with Canada and Mexico (NAFTA), World Trade Organization etc.	International Trade
Protectionism	protectionist international trade policies, such as Tariffs and trade barriers.	International Trade
Nuclear Energy		Nuclear Energy
Renewable/Green Energy		Renewable/Green Energy
Fossil fuels		Fossil fuels
Trust in the US government (general)	the Trust the US government as an organization that do not mention a party/president	Trust in the US government (general)
Trust in the Biden administration		Joe Biden
Trust in the Trump administration		Donald Trump
Trust in the Obama administration		Barack Obama
Trust in the G.W. Bush administration		George W. Bush

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
Trust in the US Military		Trust in the US Military
Trust in the US Law Enforcement		Trust in the US Law Enforcement
Trust in the US electoral system	the US electoral system, including electoral college, two-party system, first past the post voting, electoral districts, campaign donations and funding (political action comittes, SuperPACs)	Trust in the US electoral system
Restrictions on Voting Access	the regulations that make voting less accessible. Examples include registration or ID checks. Counter examples, i.e. cases that make voting easier, would include mail-in ballots, ways to reduce voting queues, or helping those waiting in the voting queue.	Restrictions on Voting Access
Trust in the US judicial system/courts	the US judicial system and courts in general. Examples include fair sentencing, costly court cases where corporations can outspend individuals	Trust in the US judicial system/courts
Trust in the US Supreme Court	Specifically refering to the US Supreme Court	Trust in the US judicial system/courts
Coronavirus stimulus checks		Coronavirus stimulus checks
Discrimination based on Political Beliefs or Activities		Discrimination based on Political Beliefs or Activities
Capital Punishment		Capital Punishment
Diversity, equity, and inclusion initiatives	DEI initiatives that seek to promote the fair treatment and full participation of all people, particularly groups who have historically been underrepresented or subject to discrimination. Criticized by Conservatives on the basis that it creates preferential treatment of minorities, forgoes merit as the basis of quality, and restricts freedom of speech.	Diversity, equity, and inclusion initiatives

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
Critical Race Theory	CRT, which focuses on the relationships between social conceptions of race and ethnicity, social and political laws, and mass media. CRT also considers racism to be systemic in various laws and rules, not based only on individuals' prejudices. It is criticisized by (often) Conservatives for being anti-American and anti-White, for undermining American history and traditions. Mostly referred in the context of right-wing backlash and censure of CRT in schools.	Critical Race Theory
Multiculturalism	multiple cultures living together or interacting	Multiculturalism
Cannabis legalization		Cannabis legalization
Environmental Protections		Environmental Protections
Trust in the scientific community	the scientific community. Examples include the academia, advisory organizations such as World Health Organization, Centers for Disease Control and Prevention (CDC)	Trust in the scientific community
Pornography		Pornography
Euthanasia		Euthanasia
AI-driven surveillance	whether the government should use AI technologies in surveillance.	AI-driven surveillance
Social Media Regulations	whether social media platforms and the associated companies should be regulated more	Social Media Regulations
Cryptocurrencies	cryptocurrencies, decentralized finance, and financial use of blockchain technology. A positive stance would favor those, a negative stance would be against or would call for more regulations.	Cryptocurrencies
AI Regulations	the general attitudes toward AI regulation or ethics	AI Regulations
Use of AI in the workplace	the use of AI , often referred in the context of AI replacing human workers.	Use of AI in the workplace
Religion (general)	religion that speak on general terms without referring to a specific belief system	Religion (general)

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
Net Neutrality	Net Neutrality, the principle that Internet service providers (ISPs) must treat all Internet communications equally.	Net Neutrality
Zoning Laws	(the US, and often state or city level) Zoning Regulations, that divide land in a municipality into zones in which certain land uses are permitted or prohibited. Often referred in the context of certain municipalities blocking new constructions to keep house/land values higher, which exacerbate housing crisis.	Zoning Laws
Whistleblowing/Leaks	whistleblowing or journalistic practice of leaking. Examples include Wikileaks, Panama Papers, Boeing whistleblowers, and often discussed through people such as Julian Assange, Edward Snowden, Chelsea Manning	Whistleblowing/Leaks
Censorship	censorship by an administrative or regulative body (the government, the state, and also municipal education boards). Often referred in the context of some US schools banning certain books.	Censorship
War on Drugs	the War on Drugs, the policy of a global campaign led by the United States federal government, of drug prohibition, foreign assistance, and military intervention, with the aim of reducing the illegal drug trade in the US. This is often referred in discussion of whether repression or legalization is better to reduce social harms caused by drugs.	War on Drugs
War on Terror	the War on Terror, a global campaign led by the US against militant Islamist movements, such as Al-Qaeda, Taliban, ISIS and offshoots. Does not include the 2003 invasion of Iraq	War on Terror
Invasion of Iraq (2003)	the invasion and occupation of Saddam's Iraq.	Foreign military interventions by the US

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
NATO/US Allies	the NATO and other US Allies. Mainly referred within the context of the US' bases abroad, and whether the Allies are doing enough (financially or militarily)	NATO/US Allies
United Nations		United Nations
European Union		European Union
Foreign military interventions by the US	the US military involvement in foreign conflicts. Examples include US involvement in the Syrian Civil War against Assad. Does not include Invasion or Iraq, War on Terror or War on Drugs.	Foreign military interventions by the US
US Support for Israel		Israel
Israel		Israel
Palestine	Palestinian people. A pro-Palestinian stance would stress the rights of Palestinians to self-governance and highlight Israeli oppresion, an anti-Palestinian stance would focus more on Hamas' leadership of Palestine, or poor governance by the Palestinian Authority over the West Bank.	Palestine
Black Lives Matter movement		Black Lives Matter movement
Blue Lives Matter movement	Blue Lives Matter movement was a counter-movement against Black Lives Matter and stressed the dangers the law enforcement faces	Blue Lives Matter movement
MeToo movement	the MeToo movement, and the practices employed or popularized by it, such as public exposures, social media campaigns against sexual harrasment and discrimination, and also "cancelling" of accused figures.	MeToo movement
Donald Trump		Donald Trump
Joe Biden		Joe Biden
Barack Obama		Barack Obama

Continued on next page

Table – Continued from previous page

Target	Explanation	Broad Target
G.W. Bush		George W. Bush
Bill Clinton		Bill Clinton
Ronald Reagan		Ronald Reagan
Richard Nixon		Richard Nixon
Franklin D. Roosevelt		Franklin D. Roosevelt
Theodore Roosevelt		Theodore Roosevelt
Abraham Lincoln		Abraham Lincoln
Hillary Clinton		Hillary Clinton
Bernie Sanders		Bernie Sanders
Alexandria Ocasio-Cortez		Alexandria Ocasio-Cortez
Elizabeth Warren		Elizabeth Warren
George Soros		George Soros
Elon Musk		Elon Musk
Ted Cruz		Ted Cruz
Marco Rubio		Marco Rubio
Mitt Romney		Mitt Romney
Ron Paul		Ron Paul
Newt Gingrich		Rand Paul
John McCain		Newt Gingrich
Trump supporters	Specifically referring to Trump supporters, rather than Republicans, Conservatives, or Right-wingers	John McCain
Biden supporters	Specifically referring to Biden supporters, rather than Democrats, Centrists, Liberals etc.	Donald Trump
Obama supporters	Specifically referring to Obama supporters, rather than Democrats, Centrists, Liberals etc.	Joe Biden
Bush supporters	Specifically referring to G.W.Bush supporters, rather than Republican, Conservatives, or Right-wingers	Barack Obama

Continued on next page

Table – Continued from previous page

Target	**Explanation**	**Broad Target**
Hillary Clinton supporters	Specifically referring to H.Clinton supporters, rather than Democrats, Centrists, Liberals etc.	George W. Bush
Sanders supporters	Specifically referring to Sanders supporters, rather than Democrats, Leftists, Socialists etc.	Hillary Clinton
N/A	When target cannot be identified	Bernie Sanders
Other	When there's an explicit target but is not listed	N/A
Belief in Trump-Russia Collusion	Regarding alleged collusion between Donald Trump and Russian authorities for political gain, including influencing elections	Other
Belief in Russian Interference in Elections	Regarding attempts by Russia to influence US elections, without asserting whether this was intended to support Trump	Belief in Trump-Russia Collusion
Increased Surveillance	the increase in state or law enforcement surveillance activities. If the main focus is on the use of AI in surveillance, categorize under 'AI-Surveillance	Belief in Russian Interference in Elections

B Appendix B: Best performing models and strategies

Model	Technique	Target Acc.	Target F1	Stance Acc.	Stance F1
o3	few-shot-with-info	**0.75**	**0.76**	**0.87**	**0.87**
o3	few-shot	0.73	0.75	0.84	0.84
gemini-2.5-pro	few-shot	**0.73**	**0.76**	0.82	0.82
gpt-4.1	broad-labels	**0.73**	0.74	0.82	0.81
gpt-4.1	info.context	**0.73**	0.75	0.81	0.81
gemini-2.5-pro	info.context	**0.73**	0.75	0.84	0.84
gpt-4.1	few-shot	0.72	0.74	0.82	0.82
o3	info.context	0.72	0.73	0.85	0.84
o3	conv.-context	0.72	0.73	0.85	0.84
gemini-2.5-pro	few-shot-with-info	0.72	0.73	0.85	0.84