Is External Information Useful for Stance Detection with LLMs?

Quang Minh Nguyen
KAIST
Daejeon, South Korea
qm.nguyen@kaist.ac.kr

Taegyoon Kim
KAIST
Daejeon, South Korea
taegyoon@kaist.ac.kr

Abstract

In the stance detection task, a text is classified as either favorable, opposing, or neutral towards a target. Prior work suggests that the use of external information, e.g., excerpts from Wikipedia, improves stance detection performance. However, whether or not such information can benefit large language models (LLMs) remains an unanswered question, despite their wide adoption in many reasoning tasks. In this study, we conduct a systematic evaluation on how Wikipedia and web search external information can affect stance detection across eight LLMs and in three datasets with 12 targets. Surprisingly, we find that such information degrades performance in most cases, with macro F1 scores dropping by up to 27.9%. We explain this through experiments showing LLMs' tendency to align their predictions with the stance and sentiment of the provided information rather than the ground truth stance of the given text. We also find that performance degradation persists with chain-of-thought prompting, while fine-tuning mitigates but does not fully eliminate it. Our findings, in contrast to previous literature on BERT-based systems which suggests that external information enhances performance, highlight the risks of information biases in LLM-based stance classifiers[1].

1 Introduction

Stance detection is a task that determines whether a given content supports, opposes, or remains neutral toward a target. When the content assumes implicit information about the target, stance detection systems can benefit from external information, such as Wikipedia excerpts, regarding the target. Accordingly, recent research has explored incorporating such information to improve stance detection, highlighting its benefits (Wen and Hauptmann, 2023; Li et al., 2023; He et al., 2022; Zhu et al., 2022).

On the other hand, large language models (LLMs) have demonstrated remarkable capabilities across various reasoning tasks, including mathematical reasoning (Imani et al., 2023), coding (Guo et al., 2024), and language understanding (Wei et al., 2022a). Given these advances, recent research has begun exploring the potential of LLMs for stance detection (Weinzierl and Harabagiu, 2024; Lan et al., 2024).

With these parallel trends, an important question arises: *Can external information enhance LLMs in stance detection?* In this paper, we systematically evaluate how external information about targets impacts the performance of a diverse set of LLMs across a wide range of datasets and targets.

Surprisingly, we find that such information, from Wikipedia or web search, tends to *compromise* stance detection performance. To explain this phenomenon, we show that model predictions often adopt external information stance and sentiment. Despite LLMs' known sensitivity to prompt variations, we show consistent results through experiments on different chain-of-thought prompting methods. Finally, we find that fine-tuning mitigates but does not fully eliminate performance degradation. Our research serves as a caution against the use of external information without proper bias consideration for LLMs in stance detection and natural language reasoning at large.

2 Related Work

Stance Detection with External Information. A key line of related work investigates leveraging external information, often from Wikipedia, to enhance stance detection. He et al. (2022) fine-tuned BERT models which take Wikipedia excerpts, in addition to given texts and targets, as inputs and report significantly improved stance detection performance. Subsequent works in the literature either utilized external information in a different formu-

[1]Code is available at https://github.com/ngqm/acl2025-stance-detection

Findings of the Association for Computational Linguistics: ACL 2025, pages 14798–14807

lation of stance detection (Wen and Hauptmann, 2023) or introduced new knowledge organization and filtering schemes for such information (Li et al., 2023; Zhu et al., 2022). While these works have primarily focused on fine-tuning smaller, BERT-like models for stance detection, we extend this research to LLMs, which possess emergent reasoning abilities but require significantly more resources for fine-tuning.

Stance Detection with LLMs. Relatedly, another stream of works examines how LLMs can be applied to stance detection. Weinzierl and Harabagiu (2024) and Lan et al. (2024) proposed prompting schemes where reasoning on stance is organized as ensembles or multi-agent discussions. Meanwhile, Li et al. (2024) introduced a calibration network which serves to mitigate internal biases of LLMs. Orthogonal yet complementary to these efforts, our work provides a foundational analysis of how external information influences their decision-making, uncovering unintended effects and offering insights to guide future research.

3 Experimental Setup

3.1 Data, Model, and Metric

We utilize the following datasets, which are all in English and widely used in stance detection research.

1. COVID-19-Stance (Glandt et al., 2021): 6,133 Tweets about COVID-19 in the U.S.: Fauci, school closure, stay-at-home orders, and face masking. Labels are either FAVOR, AGAINST, or NONE.

2. P-Stance (Li et al., 2021): 21,574 Tweets with Trump, Biden, and Sanders as targets. Labels are either FAVOR or AGAINST.

3. SemEval 2016 Task 6 (Mohammad et al., 2016): 4,163 Tweets about atheism, climate change, feminist movement, Hillary Clinton, and abortion. Labels are either FAVOR, AGAINST, or NONE.

For our experiments, we consider a total of eight popular LLMs of various sizes, both open- and closed-source (see Table 1). We utilize the instruction-tuned versions of all open-source models (the "-Instruct" postfix omitted in the paper). Additionally, we use WS-BERT (He et al., 2022) as a BERT baseline. Since stance detection is a task requiring determinism over creativity, we set the inference temperature of all models to zero; as a result, each run produced outputs with negligible variation in performance, and our results are therefore based on single runs. We evaluate models through accuracy and macro F1. More details on data, models, prompts, and output validation are in Appendices A, B, and C.

3.2 External Information

Following previous work, we utilize external information from Wikipedia collected by He et al. (2022) for COVID-19-Stance and P-Stance. The external information for SemEval 2016 Task 6 was collected ourselves through the Wikipedia API. Additionally, we consider both *synthetic* and *real-world* sources of bias: the former is implemented through Against and Favor stance injection with GPT-4o mini, while the latter employs information fetched through OpenAI's Web Search API with GPT-4o mini (OpenAI, 2025a), which may reflect non-objective content retrieved from the web. The stance injection and web search prompts are included in Appendix B.

3.3 Research Questions

RQ1. Effects of external information on performance. We first ask *how the stance detection performance of LLMs changes as external information is introduced*. We evaluate LLMs when external information (from either Wikipedia or web search) is given, relative to when no external information is available. All eight LLMs are evaluated without further training, while WS-BERT is trained using the configuration in He et al. (2022).

RQ2. Analysis on information characteristics. To explain observations in RQ1, we study *how external information stance and sentiment are related to a model's predictions* and *whether external information length correlates with performance changes*. Hypothesizing that the *perceived* stance or sentiment (positive/negative/none) by a model is what influences the model's final prediction, we consider such perception instead of information stance or sentiment *in isolation* (e.g., by having a fixed reference sentiment analysis model). We compare the proportion of Tweets for which the model's prediction is aligned with the external information stance or sentiment, with and without such information[2]. This is captured through the net adoption rate (NAR) in Equation 1. We also

[2]For sentiments, this means information with a *positive* sentiment may lead the stance prediction towards *favorable*.

Model	No Info	Wiki	Web	Wiki (Against)	Wiki (Favor)
COVID-19-Stance (Accuracy / Macro F1 in %)					
Llama-3.2-3B	49.4 / 44.8	-12.4 / -14.9	-12.9 / -14.8	-17.4 / -26.2	-9.5 / -15.8
Llama-3.1-8B	64.1 / 63.5	-0.2 / -1.2	-6.2 / -8.3	-4.1 / -5.8	-1.9 / -3.1
Qwen2.5-3B	59.6 / 54.3	-6.4 / -2.6	-15.6 / -14.3	-10.2 / -9.0	-10.6 / -10.3
Qwen2.5-7B	63.5 / 58.1	-3.4 / -6.1	-4.9 / -5.7	-4.8 / -8.0	-2.4 / -6.0
Qwen2.5-14B	72.5 / 70.8	-3.2 / -3.1	-3.6 / -3.8	-5.9 / -11.1	-1.9 / -6.8
Qwen2.5-32B	72.0 / 71.0	-3.2 / -3.8	-2.6 / -3.1	-6.1 / -7.2	-1.1 / -1.7
GPT-4o mini	64.1 / 62.9	-9.1 / -9.4	-10.9 / -11.8	-12.0 / -14.9	-11.9 / -12.3
Claude 3 Haiku	64.4 / 63.7	-7.5 / -9.0	-11.4 / -18.3	-14.0 / -16.9	-7.9 / -12.4
WS-BERT	83.9 / 82.7	+0.2 / +0.2	+0.1 / +0.1	+0.4 / +0.4	+0.2 / +0.2
P-Stance (Accuracy / Macro F1 in %)					
Llama-3.2-3B	77.4 / 59.0	-12.6 / -15.7	-18.4 / -16.5	-19.7 / -21.5	-20.9 / -22.5
Llama-3.1-8B	81.7 / 63.0	-5.3 / +12.1	-6.0 / +11.5	-13.4 / -0.1	-9.8 / +7.8
Qwen2.5-3B	65.4 / 64.4	-9.2 / -14.0	-6.4 / -14.0	-1.7 / -19.1	-3.8 / -6.7
Qwen2.5-7B	74.3 / 48.4	-1.4 / -1.6	-0.2 / +6.0	-8.0 / -8.0	-0.1 / -0.3
Qwen2.5-14B	81.7 / 54.2	-1.8 / +7.1	-1.3 / +7.3	-2.8 / +6.1	-1.0 / +8.0
Qwen2.5-32B	84.3 / 65.2	-0.8 / +8.2	-0.6 / +8.4	-2.4 / +15.9	-0.3 / +8.8
GPT-4o mini	80.9 / 53.8	-3.8 / +4.5	-3.7 / -3.1	-6.9 / +1.3	-3.0 / -2.6
Claude 3 Haiku	83.6 / 73.5	-7.3 / -9.6	-5.0 / -5.9	-18.0 / -15.4	-2.4 / +6.5
WS-BERT	80.9 / 80.3	+1.0 / +1.1	+0.1 / +0.1	+1.0 / +1.1	+1.0 / +1.1
SemEval 2016 Task 6 (Accuracy / Macro F1 in %)					
Llama-3.2-3B	63.4 / 46.3	-2.4 / -1.1	-4.6 / -8.1	-16.3 / -27.9	-22.9 / -18.0
Llama-3.1-8B	68.9 / 63.3	-5.7 / -5.3	-6.3 / -7.0	-12.6 / -16.8	-9.5 / -8.4
Qwen2.5-3B	45.7 / 43.3	-2.6 / -7.1	+3.8 / -0.3	+4.6 / -0.8	+0.1 / -4.5
Qwen2.5-7B	58.7 / 53.2	-0.3 / -3.5	-1.0 / -5.1	-4.4 / -11.2	-1.6 / -8.4
Qwen2.5-14B	71.7 / 67.1	-2.2 / -2.8	-1.6 / -3.9	-4.2 / -9.1	-3.4 / -2.8
Qwen2.5-32B	70.1 / 67.2	-0.9 / -1.8	-1.0 / -2.1	-2.9 / -6.2	-0.7 / -1.3
GPT-4o mini	74.1 / 67.6	-2.5 / -5.4	-3.4 / -6.4	-4.6 / -9.6	-2.5 / -6.3
Claude 3 Haiku	71.8 / 67.6	-1.7 / -5.1	+2.8 / -4.5	-8.6 / -20.5	+0.4 / -4.7
WS-BERT	70.9 / 57.2	-1.2 / -2.3	-0.1 / -0.7	-1.4 / -2.5	-1.4 / -2.4

Table 1: Use of external information degrades LLM stance detection performance. In each row, *No Info* stands for the performance when there is no external information; other columns contain the performance relative to *No Info*, in the presence of external information: Wikipedia, web search, Wikipedia with synthetic Against bias, and Wikipedia with synthetic Favor bias, respectively. We color positive and negative changes in blue and red, respectively.

compare the number of changes leading to incorrect (harmful) and correct predictions through the harmful adoption rate (HAR). Given a model M for a target T and external information t, we have

$$\begin{aligned}\mathrm{NAR}(t,T) &= P(M(x|t,T)=s) \\ &\quad - P(M(x|\varnothing,T)=s) \\ \mathrm{HAR}(t,T) &= P(M(x|t,T)=s, g(x)\neq s, \\ &\qquad M(x|\varnothing,T)\neq s) \\ &\quad - P(M(x|t,T)=s, g(x)=s, \\ &\qquad M(x|\varnothing,T)\neq s)\,, \end{aligned} \tag{1}$$

where $M(x|t,T)$ is the stance detected by M for a text x given information t and target T; $s = M(t|\varnothing,T)$ is the stance or sentiment (depending on the context) M detects for t; $P(c)$ is the proportion of Tweets associated with T satisfying a condition c; $g(x)$ is the ground truth stance of x. When $\mathrm{NAR} > 0$, the model *adopts* the stance or sentiment of the information. When $\mathrm{HAR} > 0$, such adoptions are more *harmful* than helpful. As for length, the number of tokens in each piece of information is counted using tiktoken (OpenAI, 2025b). We analyze one model from each model family (Llama-3.1-8B, Qwen2.5-7B, GPT-4o mini, and Claude 3 Haiku) for this experiment.

RQ3. Variations in prompting strategies. As LLMs are known to be sensitive to prompting variations (Sclar et al., 2024), we identify *how perfor-*

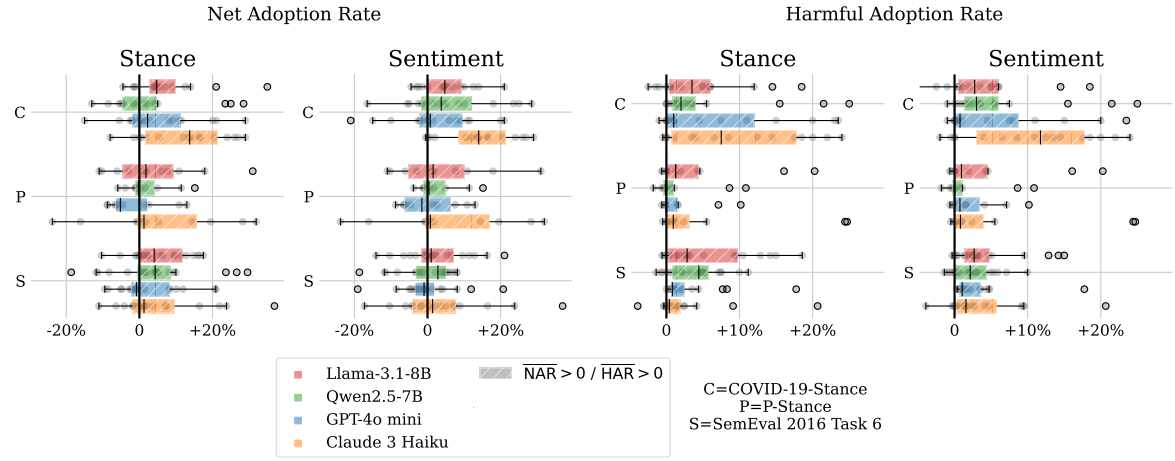

Figure 1: *Left*: Predictions are more likely to move towards information stance and sentiment than they are to move away. *Right*: When a model adopts information stance or sentiment, predictions are more likely to be wrong (harmful) than to be correct (helpful).

mance changes with zero-shot (Kojima et al., 2022) and few-shot (Wei et al., 2022b) chain-of-thought (CoT) prompting. Our prompts explicitly instruct the model to incorporate the provided external information without adopting its stance. Experiments are conducted using Llama-3.1-8B, Qwen2.5-7B, GPT-4o mini, and Claude 3 Haiku.

RQ4. Fine-tuning. Finally, we examine *how fine-tuning affects our findings*, as performance changes may vary when models are fine-tuned alongside with external information. For each target and information type, we train two low-rank adapters (LoRA) (Hu et al., 2022) of ranks 16 and 32. We fine-tune both Llama-3.1–8B and Qwen2.5–7B for 3 epochs with a batch size of 8 and a learning rate of 1×10^{-4}. This results in 96 LoRAs in total. To ensure fair comparisons with identical training settings, closed-source models are not included.

4 Results and Discussion

4.1 Effects of External Information on Performance

Table 1 shows the performance of all models with different types of external information, relative to their performance without it.[3] While there are variations depending on the model, information type, and dataset, we see an overall trend of performance degradation. Using information from Wikipedia, the most extreme drop in macro F1 is observed for Llama-3.2-3B for P-Stance, by 15.7%, which becomes more severe when synthetic biases are introduced (by up to 27.9% for Llama-3.2-3B, SemEval 2016 Task 6, and Against bias); with web-search information, the steepest drop in macro F1 reaches 18.3% for Claude 3 Haiku on COVID-19-Stance. This behavior of LLMs contrasts with the fine-tuned WS-BERT, for which the performance generally increases, consistent with He et al. (2022).[4] Together, these results suggest that *external information often decreases LLMs' stance detection performance*.

4.2 Analysis on Information Characteristics

To gain insights into why external information often reduces performance, we consider the effects of information stance, sentiment, and length. Figure 1 visualizes NAR and HAR as defined in Equation 1. We observe that *in the presence of external information, a model is likely to adopt the stance of such information in its predictions*: the mean net adoption rate is positive in 11 out of 12 model–dataset combinations for stance, and in 10 out of 12 combinations for sentiment. Figure 1 (*Right*) shows that *when such adoption happens, more incorrect predictions are made than correct ones*: the mean harmful adoption rates are positive for all models and datasets.

In Figure 2, we visualize the correlation between information length and rates of predicted classes as well as relative performance. We observe a *correlation between length and the rate of FAVOR predictions but no significant correlation in other cases*. These results align with the work of Sclar

[3]Note that macro F1 is an unweighted average across tasks in each dataset while accuracy is weighted.

[4]Note that WS-BERT's performance decreases, though by a small margin, on the SemEval 2016 Task 6 dataset.

et al. (2024), where the correlation between prompt length and task performance is minimal across many multiple-choice and classification tasks.

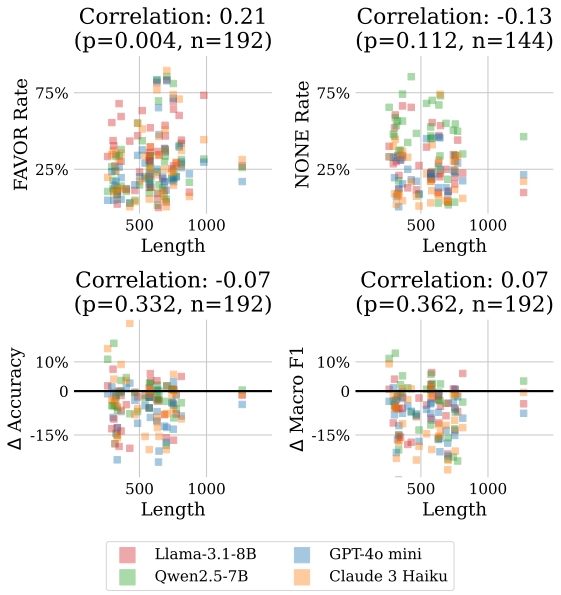

Figure 2: *Top*: There is some correlation between external information length and FAVOR predictions but not NONE predictions. *Bottom*: There is no correlation between external information length and performance change due to such information. Details on correlation calculations are in Appendix D.

4.3 Variations in Prompting Strategies

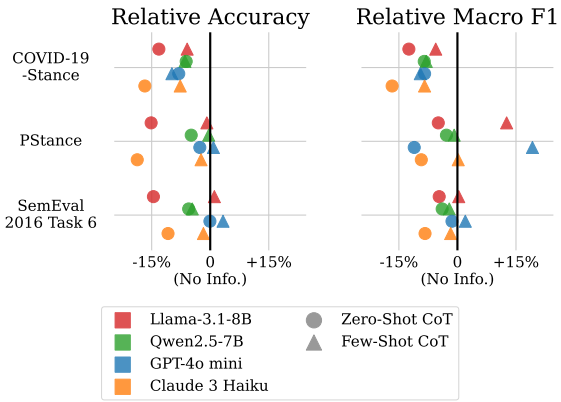

Figure 3: Use of external information degrades stance detection performance when zero-shot and few-shot CoTs are employed. We plot the average results for all types of information.

Figure 3 illustrates the performance changes of Llama-3.1-8B, Qwen2.5-7B, GPT-4o-mini, and Claude 3 Haiku with zero-shot and few-shot CoT reasoning. Surprisingly, in most cases, *external information also decreases the performance of the considered LLMs with CoTs*. The steepest average F1 drops are by 16.7% (with Claude 3 Haiku and COVID-19-Stance) for zero-shot CoT and 9.5% (with GPT-4o-mini and COVID-19-Stance) for few-shot CoT.

4.4 Fine-Tuning

To examine how fine-tuning shapes the role of external information, we visualize the relative performance of models through 3 epochs of fine-tuning in Figure 4. Overall, we observe a monotonic increase with variances contracting with more epochs. Nevertheless, even at the third epoch, we still observe 25 out of 48 rank-32 instances falling below zero for both relative accuracy and relative macro F1 (see Appendix E for more details). This means that fine-tuned LLMs still *do not benefit from external information* in many cases.

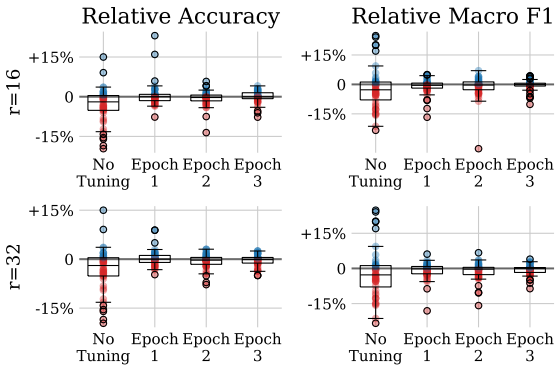

Figure 4: Models become more robust but do not completely eliminate the negative effects of external information through fine-tuning. Performance visualized is relative to the no-information setting. *Top*: Results for LoRA rank 16. *Bottom*: Results for LoRA rank 32.

5 Conclusion

We investigated the question of whether external information can benefit LLMs for stance detection. Contradicting previous literature on BERT-based stance detection with external information, our experiments indicated that such information can actually harm the performance of LLMs. We also verified that this phenomenon is partly caused by LLMs being biased by the stance or sentiment they perceive in external information. Furthermore, CoT prompting is of little benefit, while fine-tuning lessens but does not completely alleviate this problem. Given such observations, we call for more consideration of bias factors in LLM stance detection and natural language reasoning at large.

6 Limitations

Our work provided a systematic evaluation of how external information can affect the performance of LLM stance detection systems. This research can serve as a foundation for a number of crucial future research directions.

Our analysis involving information characteristics represents one of many possible perspectives and levels of depth for interpreting the main results. For instance, another perspective could involve probing the attention heads of models (Marks and Tegmark; Kim et al., 2025). We hope future work will explore such complementary approaches in depth.

Additionally, even though we have conducted experiments on prompting variations and fine-tuning, there could still be advanced methods which may more efficiently improve the stance detection performance of LLMs under external information. Since our focus is on the analysis of effects external information can have on LLMs, investigations on more optimal mitigation measures are left open for future research.

7 Ethical Considerations

Given the tendency of LLMs to be biased by the stance of external information, as investigated in our paper, it is possible for malicious actors to manipulate open information sources such as Wikipedia to alter the outputs of LLM stance detection systems. We caution against the use of external information without proper curation of the information source and also encourage future research on mitigation measures.

Furthermore, even though Tweets in the datasets we utilized have been anonymized by their respective authors (Glandt et al., 2021; Li et al., 2021; Mohammad et al., 2016), their content might contain offensive language against targets. Our work reports aggregated statistics and analysis from such data but does not present any offensive information individually.

Acknowledgements

This work was supported by the National Research Foundation of Korea (RS-2022-NR068758) and Underscore.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, volume 1.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, and 1 others. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Kam-Fai Wong, and Ruifeng Xu. 2024. Mitigating biases of large language models in stance detection with calibration. *arXiv preprint arXiv:2402.14296*.

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.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

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.

OpenAI. 2025a. Openai api web search.

OpenAI. 2025b. tiktoken.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880.

Haoyang Wen and Alexander G Hauptmann. 2023. Zero-shot and few-shot stance detection on varied topics via conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1491–1499.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2070–2075.

A Details of Data and Models

All of the datasets we use are made publicly available by their respective authors (Glandt et al., 2021; Li et al., 2021; Mohammad et al., 2016). Readers can refer to the orignal papers for instructions on how to access the data. Their statistics are included in Table A1. Note that since the SemEval data only has training and testing sets, we perform stratified sampling with the `scikit-learn` library to use 20% of the training sets as validation sets.

The LLMs we utilize include Claude 3 Haiku (snapshot `20240307`) (Anthropic, 2024), GPT-4o mini (version `2024-07-18`) (Hurst et al., 2024), Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct (Dubey et al., 2024), and Qwen2.5-{3B, 7B, 14B, 32B}-Instruct (Yang et al., 2024). We used 4-bit quantizations for all Llama and Qwen models provided by Unsloth (Daniel Han and team, 2023). We also made use of Unsloth's fine-tuning library. Inference is done through the vLLM library (Kwon et al., 2023).

All of our inference temperatures are set to zero, yielding negligible performance stochasticity, because of which results are reported with single runs. Our hardware was 4× NVIDIA RTX A5000, with which a single run over all of our training and evaluation (for both LLMs and BERT models) takes approximately 25 GPU hours. Our evaluation through OpenAI API (for GPT-4o mini) and Anthropic API (for Claude 3 Haiku) cost approximately 75 USD.

Target	Train			Val			Test		
	Favor	Against	None	Favor	Against	None	Favor	Against	None
COVID-19-Stance									
Face Masks	531	512	264	81	78	41	81	78	41
Fauci	388	480	596	52	65	83	52	65	83
School Closures	409	166	215	103	42	55	103	42	55
Stay at Home Orders	136	284	552	27	58	115	27	58	115
P-Stance									
Bernie Sanders	2858	2198	0	350	284	0	343	292	0
Joe Biden	2552	3254	0	328	417	0	337	408	0
Donald Trump	2937	3425	0	374	440	0	352	425	0
SemEval 2016 Task 6									
Atheism	74	243	93	18	61	24	32	160	28
Climate Change	170	12	134	42	3	34	123	11	35
Feminist Movement	168	262	101	42	66	25	58	183	44
Hillary Clinton	89	289	133	23	72	33	45	172	78
Abortion	84	267	131	21	67	33	46	189	45

Table A1: The number of samples in each target and split of the datasets. “Climate Change" is short for “Climate Change is a Real Concern". “Abortion" is short for “Legalization of Abortion".

B Prompts

Non-CoT Stance Detection Prompt. For non-CoT stance detection, all models are prompted with the following instruction:

```
USER: You are given the following text: {text}. What is the stance of the text towards the target ‘{target}’? The following information can be helpful: {wiki}. Options: {options}. Do not explain. Just provide the stance in a single word.
ASSISTANT:
```

Here `{wiki}` stands for the external information excerpt that can also be empty, in which case the sentence in red is omitted. Meanwhile, `{options}` is `[FAVOR, AGAINST]` for P-Stance and `[FAVOR, AGAINST, NONE]` for COVID-19-Stance and SemEval 2016 Task 6. USER and ASSISTANT are replaced with tokens corresponding to the chat prompt template of each model.

Zero-Shot CoT Stance Detection Prompt. Following Kojima et al. (2022), we prompt models as follows:

```
USER: You are given the following text: {text}. What is the stance of the text towards the target ‘{target}’? Integrate the following external information and do NOT automatically adopt the stance of it: {wiki}. Options: {options}
ASSISTANT: Let’s think step by step.
```

Again, the sentence in red is omitted when there is no external information.

Few-Shot CoT Stance Detection Prompt. Our prompt for the few-shot CoT setting (Wei et al., 2022b) has the following format:

```
USER: You are given the following text: {FAVOR text}. What is the stance of the text towards the target ‘{target}’? Integrate the following external information and do NOT automatically adopt the stance of it: {wiki}. Options: {options}
ASSISTANT: {FAVOR CoT}

USER: You are given the following text: {AGAINST text}. What is the stance of the text towards the target ‘{target}’? Integrate the following external information and do NOT automatically adopt the stance of it: {wiki}. Options: {options}
ASSISTANT: {AGAINST CoT}
```

```
USER: You are given the following text: {NONE text}. What is the stance of the text towards the target '{target}'? Integrate the following external information and do NOT automatically adopt the stance of it: {wiki}. Options: {options}
ASSISTANT: {NONE CoT}

USER: You are given the following text: {text}. What is the stance of the text towards the target '{target}'? Integrate the following external information and do NOT automatically adopt the stance of it: {wiki}. Options: {options}
ASSISTANT:
```

The sentences in red are omitted when there is no external information. Here {FAVOR text}, {AGAINST text}, and {NONE text} are Tweets with FAVOR, AGAINST, and NONE ground truth labels, respectively. Meanwhile, {FAVOR CoT}, {AGAINST CoT}, and {NONE CoT} are our CoT examples for each class. The last in-context example, for the NONE class, is not included for the P-Stance dataset, which has no such class.

The authors wrote the CoT examples themselves for each target and information type (Wikipedia, Wikipedia with Against bias, Wikipedia with Favor bias, web search, or no information).

Web Search Prompt. For the COVID-19-Stance dataset, we search using the prompt

```
Provide a summary of {target} in the context of COVID-19.
```

For the P-Stance dataset, the prompt is

```
Provide a summary of {target} in the context of the 2020 U.S. presidential election.
```

For the SemEval 2016 Task 6 dataset, we search using the prompt

```
Provide a summary of {target}.
```

The prompts for COVID-19-Stance and P-Stance are written while taking into account the context of each dataset (COVID-19 for the former and the 2020 U.S. presidential election for the latter.)

Synthetic Stance Injection Prompt. We inject synthetic Favor and Against stances into Wikipedia information excerpts using the following prompt:

```
USER: You are given the following Wikipedia entry: {wiki}. Rewrite the Wikipedia entry to have the stance '{stance}' towards the target '{target}'. Be discreet and do not change the factual content.
ASSISTANT:
```

Perceived Stance and Sentiment Prompt. We collect LLMs' perceived sentiment of external information using the prompt

```
USER: You are given the following text: {information}. What is the sentiment of the text? Options: {options}. Do not explain. Just provide the sentiment in a single word.
ASSISTANT:
```

{options} is [POSITIVE, NEGATIVE, NONE] for COVID-19-Stance and SemEval 2016 Task 6 and [POSITIVE, NEGATIVE] for P-Stance.

Similarly, the perceived stance of information towards a target is collected using the prompt

```
USER: You are given the following text: {information}. What is the stance of the text towards the target '{target}'? Options: {options}. Do not explain. Just provide the stance in a single word.
ASSISTANT:
```

{options} is [FAVOR, AGAINST, NONE] for COVID-19-Stance and SemEval 2016 Task 6 and [FAVOR, AGAINST] for P-Stance.

C LLM Output Validation

For the non-CoT setting, if the output is within {favor, favour, favorable, favourable} (non-case-sensitive), we register the final answer as 'FAVOR'. If the output is against, we register 'AGAINST'. For outputs that are within {none, neutral}, we register 'NONE'. Other answers are considered invalid. In the case of stance detection with 2-classes, applying to P-Stance among our datasets, none and neutral outputs are invalid.

Metric	No Tuning	Ep. 1	Ep. 2	Ep. 3
LLMs, LoRA rank 16 (out of 48 instances)				
Accuracy	32	21	25	20
Macro F1	31	27	27	25
LLMs, LoRA rank 32 (out of 48 instances)				
Accuracy	32	17	27	25
Macro F1	31	26	28	25
WS-BERT (out of 24 instances)				
Accuracy	3	12	4	8
Macro F1	5	17	4	10

Table A2: The number of combinations of target-model-type of external information in each fine-tuning epoch for which the performance is lower than that of the same target and model without external information.

Metric	No Tuning	Ep. 1	Ep. 2	Ep. 3
LLMs, LoRA rank 16 (in % performance)				
Accuracy Mean	-3.28	0.61	-0.59	-0.07
Accuracy Std	6.59	4.58	2.93	2.30
Macro F1 Mean	-2.21	-1.02	-1.30	-0.43
Macro F1 Std	10.46	3.90	5.49	2.84
LLMs, LoRA rank 32 (in % performance)				
Accuracy Mean	-3.28	0.34	-0.62	-0.41
Accuracy Std	6.59	2.47	2.29	1.86
Macro F1 Mean	-2.21	-0.99	-1.25	-0.59
Macro F1 Std	10.46	3.53	3.75	2.36
WS-BERT (in % performance)				
Accuracy Mean	-0.02	2.17	1.75	0.20
Accuracy Std	1.46	15.08	2.48	1.72
Macro F1 Mean	-0.15	-1.29	3.56	0.13
Macro F1 Std	1.32	5.95	5.57	2.47

Table A3: The mean and standard deviation of relative performance in each fine-tuning epoch.

For CoT settings, we follow suggestions by Kojima et al. (2022) and conduct a second round of prompting for answer extraction. The answer extraction prompt is a concatenation of the original prompt, the generated CoT, and a trigger sentence. We use `"Therefore, among FAVOR, AGAINST, and NONE, the final answer is "` for COVID-19-Stance and SemEval 2016 Task 6 and `"Therefore, between FAVOR and AGAINST, the final answer is "` for P-Stance as trigger sentences. Again, following (Kojima et al., 2022), we parse the final answer by getting the first match after "the final answer is" that is within the set {`favor`, `favour`, `against`, `none`, `neutral`} (non-case-sensitive). For P-Stance, `none` and `neutral` are considered invalid answers.

Sentiment outputs are similarly validated, with `favor` (and variations) and `against` replaced with `positive` and `negative`, respectively.

D Notes on Length Correlations

In Figure 2, note that FAVOR prediction rates are calculated after excluding NONE predictions for a fair comparison among the three datasets. For NONE prediction rates, we exclude P-Stance, which has no such label. There is no AGAINST prediction rate visualization since its correlation with information length is precisely the additive inverse of the FAVOR counterpart: $\rho(1-X,Y) = -\rho(X,Y)$, where X is the FAVOR prediction rate, $1-X$ is the AGAINST prediction rate, and Y denotes information length.

E Performance Changes with Fine-tuning

Tables A2 and A3 show the number of instances with negative relative performance as well as relative performance in each fine-tuning epoch.