# Is External Information Useful for Stance Detection with LLMs?

Anonymous ACL submission

## Abstract

In the stance detection task, a text is classified as either favorable, opposing, or neutral towards a target. Prior work suggests 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 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 15.9%. This degradation is even more pronounced at a 28.1% drop when stance biases are introduced in the external information, as LLMs tend to align their predictions with the stance of the provided information rather than the ground truth stance of the given text. We also find that fine-tuning mitigates bias but does not fully eliminate it. Our findings, in contrast to previous literature on BERT-based systems suggesting that external information enhances performance, highlight the risks of information biases in LLM-based stance classifiers.

## 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., 2022). 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 stance detection datasets and targets.

Surprisingly, we find that such information tends to *compromise* stance detection performance, with macro F1 dropping as much as **15.9%** and even further at **28.1%** when biases are synthetically introduced in the information. We also investigate the effects of how LLMs perceive the stance of external information and find that LLMs tend to align with it, which partially explains the performance decline. Finally, we find that fine-tuning mitigates but does not fully eliminate this effect. 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

### 2.1 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 formulation 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.

### 2.2 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 in this area.

## 3 Experimental Setup

### 3.1 Data and Models

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 8 popular LLMs, both open- and closed-source (see Table 1). 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. 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

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. Furthermore, in order to simulate biases that are inherent in open, non-static platforms like Wikipedia (Greenstein and Zhu, 2012; Hube, 2017), we generate three additional versions of each Wikipedia excerpt using GPT-4o mini, where the content is rewritten to portray either a Favor, Against, or Neutral stance towards the target. The exact prompt is included in Appendix B.

### 3.3 Research Questions and Experiments

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

**Effects on predictions** We examine the mechanism in which external information shapes the predictions of LLMs. Our hypothesis is that *a model is likely to align its prediction with the stance it detects in the given external information*. To examine this, we measure the proportions of Tweets associated with a given target that is classified as a stance $s$, where $s$ is the stance of an external information excerpt classified by the same model, relative to the same metric when no external information is given. Formally, we have our *tendency metric*

$$\tau(m, t, w) = P_m(s|t, w) - P_m(s|t, w_0)\,, \quad (1)$$

where $m$ denotes a model, $t$ denotes a target, $w$ is an external information excerpt, $w_0$ is an empty string, $s$ is the stance that $m$ predicts for $w$, and $P_m(s|t, w)$ stands for the proportion of Tweets for the target $t$ that is classified by $m$ as $s$, given the external information excerpt $w$.

**Fine-tuning** Finally, we examine the *effect of fine-tuning* as performance changes may vary when models are fine-tuned alongside with external information. We train low-rank adapters (LoRA) (Hu et al., 2022) of rank 16 for all Llama and Qwen models and use the fine-tuning API for GPT-4o mini and Gemini 1.5 Flash for 3 epochs with batch size 8 and learning rate $1e-4$. Note that, due to our compute budget, we only perform fine-tuning for the COVID-19-Stance dataset.

| Model | No Context | Default | Favor | Against | Neutral |
|---|---|---|---|---|---|
| COVID-19-Stance (Accuracy / Macro F1 in %) | | | | | |
| Qwen2.5-1.5B-Instruct | 46.3 / 36.9 | -13.1 / -15.9 | -8.8 / -7.9 | -15.0 / -19.5 | -13.0 / -15.5 |
| Qwen2.5-3B-Instruct | 59.7 / 54.4 | -6.2 / -2.4 | -10.7 / -10.3 | -10.1 / -8.9 | -10.4 / -8.9 |
| Llama-3.2-3B-Instruct | 49.5 / 44.9 | -12.7 / -15.3 | -9.8 / -16.0 | -17.5 / -26.3 | -12.1 / -14.5 |
| Llama-3.1-8B-Instruct | 64.2 / 63.5 | -0.4 / -1.3 | -1.8 / -3.1 | -4.1 / -5.8 | -0.5 / -3.2 |
| Gemini 1.5 Flash 8B | 69.8 / 68.1 | -5.0 / -6.3 | -3.8 / -5.2 | -9.2 / -10.1 | -1.4 / -2.6 |
| Gemini 1.5 Flash | 70.4 / 69.6 | -2.0 / -2.9 | -1.8 / -2.8 | -4.1 / -4.7 | -2.2 / -3.7 |
| GPT-4o Mini | 64.1 / 62.9 | -9.1 / -9.4 | -11.9 / -12.3 | -12.0 / -14.9 | -8.5 / -8.6 |
| Claude 3 Haiku | 64.4 / 63.7 | -7.5 / -9.0 | -7.9 / -12.4 | -14.0 / -16.9 | -9.0 / -10.6 |
| WS-BERT | 83.9 / 82.7 | +0.2 / +0.2 | +0.2 / +0.2 | +0.4 / +0.4 | +0.2 / +0.2 |
| PStance (Accuracy / Macro F1 in %) | | | | | |
| Qwen2.5-1.5B-Instruct | 71.2 / 46.7 | -3.2 / +11.5 | -1.4 / +20.8 | -13.4 / -1.4 | -3.2 / +15.5 |
| Qwen2.5-3B-Instruct | 65.4 / 64.4 | -9.1 / -14.0 | -3.7 / -6.7 | -1.7 / -19.1 | -4.5 / -9.4 |
| Llama-3.2-3B-Instruct | 77.4 / 59.1 | -12.5 / -15.7 | -20.9 / -22.5 | -19.8 / -21.6 | -13.9 / -16.7 |
| Llama-3.1-8B-Instruct | 81.6 / 62.9 | -5.3 / +12.1 | -9.7 / +8.0 | -13.4 / -0.1 | -3.9 / +13.8 |
| Gemini 1.5 Flash 8B | 82.3 / 65.2 | -3.1 / +3.7 | -2.1 / -12.1 | -6.9 / -9.0 | -3.0 / -6.1 |
| Gemini 1.5 Flash | 84.7 / 74.6 | -2.8 / +6.6 | -3.8 / +5.4 | -4.2 / +5.1 | -3.2 / +6.2 |
| GPT-4o Mini | 80.9 / 53.8 | -3.8 / +4.5 | -3.0 / -2.6 | -6.9 / +1.3 | -3.6 / -3.0 |
| Claude 3 Haiku | 83.6 / 73.5 | -7.3 / -9.6 | -2.4 / +6.5 | -18.1 / -15.5 | -6.5 / -8.1 |
| WS-BERT | 80.9 / 80.3 | +1.0 / +1.1 | +1.0 / +1.1 | +1.0 / +1.1 | +1.0 / +1.1 |
| SemEval 2016 Task 6 (Accuracy / Macro F1 in %) | | | | | |
| Qwen2.5-1.5B-Instruct | 60.9 / 43.3 | +1.9 / -10.8 | -1.6 / -15.5 | -3.3 / -20.0 | -0.5 / -12.3 |
| Qwen2.5-3B-Instruct | 45.8 / 43.4 | -2.7 / -7.1 | +0.0 / -4.6 | +4.5 / -0.9 | -0.8 / -4.2 |
| Llama-3.2-3B-Instruct | 63.6 / 46.5 | -2.4 / -1.2 | -23.1 / -18.1 | -16.5 / -28.1 | -2.0 / -3.3 |
| Llama-3.1-8B-Instruct | 69.0 / 63.3 | -5.8 / -5.3 | -9.7 / -8.6 | -12.8 / -17.1 | -6.1 / -6.6 |
| Gemini 1.5 Flash 8B | 68.3 / 62.8 | -2.0 / -3.8 | -5.8 / -8.3 | -4.2 / -8.7 | -3.2 / -4.5 |
| Gemini 1.5 Flash | 71.7 / 64.6 | -0.1 / -0.0 | -0.5 / +0.1 | +0.6 / -0.8 | -0.3 / -0.1 |
| GPT-4o Mini | 74.1 / 67.6 | -2.5 / -5.4 | -2.5 / -6.3 | -4.6 / -9.6 | -2.5 / -5.1 |
| Claude 3 Haiku | 71.8 / 67.6 | -1.7 / -5.1 | +0.4 / -4.7 | -8.6 / -20.5 | +0.5 / -5.1 |
| WS-BERT | 70.9 / 57.2 | -1.2 / -2.3 | -1.4 / -2.4 | -1.4 / -2.5 | -1.1 / -2.2 |

Table 1: The effect of external information. In each row, 'No Context' stands for the accuracy when there is no external information; other columns contain the accuracy/macro F1, relative to 'No Context', when a corresponding type of external information relevant to the target is included. We color positive and negative changes in blue and red, respectively.

## 4 Results and Discussion

### 4.1 Effects on Performance

Table 1 shows the performance of all models with different types of external information, relative to their performance without it.[1] While there are variations depending on the model, information type, and dataset, we see an overall trend of performance degradation. Using the default information, the most extreme changes are with Qwen-2.5-1.5B-Instruct for COVID-19-Stance, where accuracy and macro F1 drop by 13.1% and 15.9%, respectively. With GPT-modified information, The steepest drop in both accuracy and macro F1 reaches 23.1% for Favor and 28.1% for Against, which correspond to Llama-3.2-3B-Instruct for SemEval. This behavior of LLMs contrasts with the fine-tuned WS-BERT, which stays robust against different information types. For WS-BERT, the performance generally increases, consistent with He et al. (2022).[2] Together, these results suggest that *external information decreases LLMs' stance detection performance* on average and that *this performance decrease tends to be more pronounced when the external information contains a biased stance*.

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

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

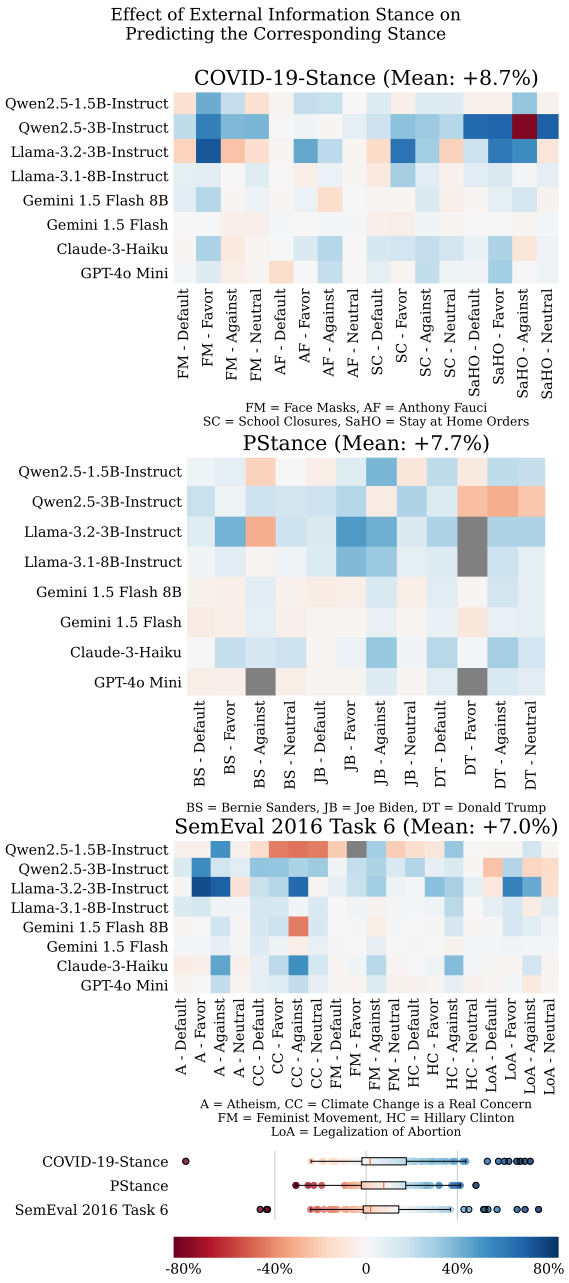


Figure 1: The effect of external information stance on predicting the corresponding stance. For each dataset, LLM, and piece of external information, we plot the *tendency metric*, defined by Equation 1. Combinations where the LLM outputs an invalid information stance are colored gray. The box plot shows the distribution of plotted values for each dataset.

### 4.2 Effects on Predictions

To gain insight into why external information often reduces performance, we quantify the tendency of LLMs to follow the stance of external information (Equation 1). In Figure 1, we observe that *LLMs tend to be biased towards the stance of external information*, with a mean tendency metric of +8.7%, +7.7%, and +7.0% for the COVID-19-Stance, PStance, and SemEval datasets, respectively. This bias inevitably lowers the performance when external information has a detected stance different from the ground truth stance.

### 4.3 Fine-Tuning

To examine how fine-tuning shapes the role of external information, we visualize the performance of models through 3 epochs of fine-tuning in Figure 2. Overall, we observe a monotonic increase with variances contracting with more epochs. Nevertheless, even at the third epoch, we still observe 51 and 54 of the 96 instances falling below zero for relative accuracy and macro F1, respectively, and the standard deviation of relative macro F1 among the LLMs is 3.2% compared to 1.1% for WS-BERT (see Appendix E for more details). This means that fine-tuned LLMs are still *not robust and do not benefit from external information* in many cases. We draw similar conclusions from another dimension of analysis, model sizes, in Appendix D.

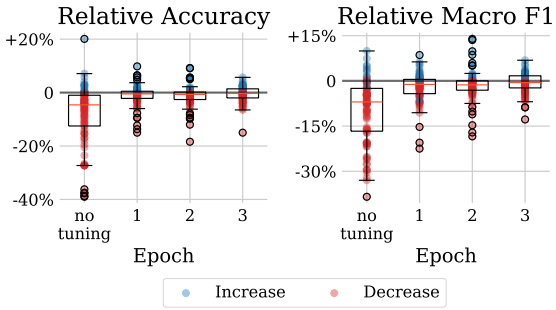


Figure 2: The effect of fine-tuning. Each point represents a combination of the target, model, and type of external information. Values are relative to the same target and model without external information.

## 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 they perceive in external information. Furthermore, 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 effect the performance of LLM stance detection systems. This research can serve as a foundation for a number of crucial future research directions.

First, due to the lack of computational resources, our analysis of open-source models is limited to LLMs with 8B or fewer parameters; we also did not experiment with prompt variations. It is thus an important avenue for future research to determine whether our reported observations can hold against further parameter scaling and prompt variations.

Second, our analysis of LLMs' tendency to align with the stance of external information represents one of many possible perspectives and levels of depth for interpreting the main results. One of such perspectives can be to probe the internal activations of models (Liu et al., 2024). We look forward to future work that can investigate such perspectives.

Finally, for a fair comparison with WS-BERT, we did not employ any knowledge filtering or inference schemes (e.g., chain-of-thoughts) for LLM stance detectors, which may plausibly improve the stance detection performance of LLMs. Investigations on how such techniques can alleviate our reported biases are left open.

## 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.

## 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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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.

Shane Greenstein and Feng Zhu. 2012. Is wikipedia biased? *American Economic Review*, 102(3):343–348.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 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*.

Christoph Hube. 2017. Bias in wikipedia. In *Proceedings of the 26th international conference on world wide web companion*, pages 717–721.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 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.

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.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Bing Liu, Haonan Lu, and Wenliang Chen. 2024. Probing language models for pre-training data detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1587, Bangkok, Thailand. Association for Computational Linguistics.

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.

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, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

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, et al. 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). 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), Gemini 1.5 Flash (version `001`) and Gemini 1.5 Flash 8B (version `001`) (Georgiev et al., 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-{1.5B, 3B}-Instruct (Yang et al., 2024). Due to our compute budget, we used 4-bit quantizations for all Llama and Qwen models provided by Unsloth (Daniel Han and team, 2023) and also made use of its fine-tuning library.

All of our inference temperatures are set to zero, meaning there is no stochasticity, hence no report of error margins in our paper. Our hardware was 4× NVIDIA RTX A5000, with which our training and evaluation (for both LLMs and BERT models) take approximately 50 GPU hours at maximum utilization. Our training and evaluation through the OpenAI API (for GPT-4o mini), Gemini API (for Gemini 1.5 Flash and Gemini 1.5 Flash 8B), and Anthropic API (for Claude 3 Haiku) cost approximately 100 USD in total.

## B Prompts

For stance detection, all models are prompted with the following instruction:

```
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.
```

| 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 |
| PStance | | | | | | | | | |
| 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"

Here {wiki} stands for the external information excerpt that can also be empty, in which case the sentence in blue is omitted. Meanwhile, {options} is `[FAVOR, AGAINST]` for P-Stance and `[FAVOR, AGAINST, NONE]` for COVID-19-Stance and SemEval 2016 Task 6.

For creating biased Wikipedia excerpts, we prompt GPT-4o mini as follows:

```
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.
```

Here {wiki} stands for a Wikipedia excerpt retrieved using the Wikipedia API and {stance} can be either `favor`, `against`, or `neutral`.

## C LLM Output Validation

Given an LLM output, we first remove all non-alphabetical characters and replace all upper-case characters with their lower-case versions. If the output is within {`favor`, `favour`, `favorable`, `favourable`}, 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 PStance among our datasets, `none` and `neutral` outputs are invalid.

## D Differences by model size

We additionally investigate *how the number of parameters of models affects performance*. We consider 4 classes of model sizes: 1B (Qwen2.5-1.5B), 3B (Llama-3.2-3B, Qwen2.5-3B), 8B (Llama-3.1-8B, Gemini 1.5 Flash 8B), and greater than 8B (Claude 3 Haiku, Gemini 1.5 Flash, GPT-4o mini). The number of parameters for closed-source models are estimates by the authors and not precise.

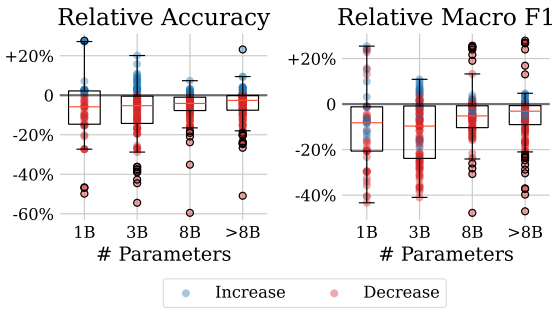


Figure A1: The effect of model sizes on relative accuracy and macro F1. Each point represents a combination of target, model, and type of external information. Values are relative to the same target and model without external information.

Figure A1 illustrates the distribution of the accuracy and macro F1 across all targets, models, and types of external information, relative to predictions without external information. Across the two metrics, the mean performance loss and variance

tends to decrease but does not completely vanish as the number of parameters increases. Specifically, for the class '>8B', mean accuracy and macro F1 changes reduce to -4.8% and -5.2% while standard deviations reduce to 7.8% and 10.6%, respectively (Table A2). This means that external information, on average, is not helpful for zero-shot LLM stance detection, even with larger models. Furthermore, larger models are not completely robust against different types of information. This result aligns with existing literature on the impact of prompt formatting for LLMs (Sclar et al., 2024), where prompt variations can still affect the performance of models having up to 70B parameters.

| Metric (%) | 1B | 3B | 8B | >8B |
|---|---|---|---|---|
| Accuracy Mean | -6.6 | -8.7 | -5.5 | -4.8 |
| Accuracy Std. | 16.7 | 14.3 | 8.4 | 7.8 |
| Macro F1 Mean | -8.1 | -11.7 | -4.5 | -5.2 |
| Macro F1 Std. | 17.1 | 13.6 | 11.1 | 10.6 |

Table A2: The mean and standard deviation of relative performance for each class of model sizes.

## E Performance Changes with Fine-tuning

Tables A3 and A4 show the performance changes in accordance with each fine-tuning epoch.

| Metric | No Tuning | Ep. 1 | Ep. 2 | Ep. 3 |
|---|---|---|---|---|
| LLMs (out of 96 instances) | | | | |
| Accuracy | 79 | 60 | 64 | 51 |
| Macro F1 | 84 | 59 | 70 | 54 |
| WS-BERT (out of 16 instances) | | | | |
| Accuracy | 5 | 10 | 3 | 4 |
| Macro F1 | 9 | 14 | 8 | 8 |

Table A3: The number of combinations of target, model, and 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 (in % performance) | | | | |
| Accuracy Mean | -8.2 | -1.1 | -1.2 | -0.5 |
| Accuracy Std. | 11.1 | 3.7 | 3.8 | 2.9 |
| Macro F1 Mean | -9.7 | -2.0 | -1.6 | -0.7 |
| Macro F1 Std. | 10.4 | 4.7 | 4.7 | 3.2 |
| WS-BERT (in % performance) | | | | |
| Accuracy Mean | -0.1 | -0.0 | 0.1 | 0.5 |
| Accuracy Std. | 3.4 | 2.2 | 0.5 | 0.8 |
| Macro F1 Mean | -0.7 | -4.3 | -0.1 | 0.5 |
| Macro F1 Std. | 2.5 | 3.4 | 1.1 | 1.1 |

Table A4: The mean and standard deviation of relative performance in each fine-tuning epoch.