MSME: Zero-Shot Stance Detection via Multi-Stage Multi-Expert Reasoning

Anonymous ACL submission

Abstract

In recent years, zero-shot stance detection based on LLMs has garnered increasing attention and demonstrated promising results. However, it continues to face three significant challenges: a heavy reliance on accurate event background knowledge, poor performance in reasoning about complex targets, and difficulties in handling rhetorical expressions such as irony and metaphor. To address these challenges,we design a **M**ulti-**S**tage **M**ulti-**E**xpert zero-shot stance detection framework(MSME). In the preparation stage, MSME automatically retrieves background knowledge related to the target and constructs explicit stance labels. In the analysis stage, the social media expert focuses on developing fine-grained stance labels, the knowledge reasoning expert emphasizes the logical connections between background information and the target, while the pragmatics expert analyzes the implicit influence of rhetorical devices on stance expression. In the decision-making stage, the decision-maker integrates the results of multi-dimensional analyses to produce a final, interpretable stance judgment. Experimental results show that the proposed MSME achieves higher F1 scores than current SOAT baselines across three public datasets, particularly for texts containing complex targets and rhetorical structures.

1 Introduction

Stance detection aims to automatically identify the stance(favor, against, or neutral) of a text towards a specific target (event, entity, or opinion)(Mohammad et al., 2016), as shown in Figure 1. It is a core technology in fields such as public opinion analysis and social media monitoring. With the rapid advancement of social media, an increasing number of individuals are expressing their views and opinions online. Traditional stance detection methods depend on domain-specific annotated data(Xu et al., 2016b). To tackle the continuous emergence of new topics and cross-domain scenarios(Hardalov et al., 2021) , zero-shot stance detection has emerged as a significant research focus(Allaway and McKeown, 2020a). Its essence lies in achieving stance inference on previously unseen targets through knowledge transfer and semantic decoupling, or by leveraging the zero-shot capabilities of models.

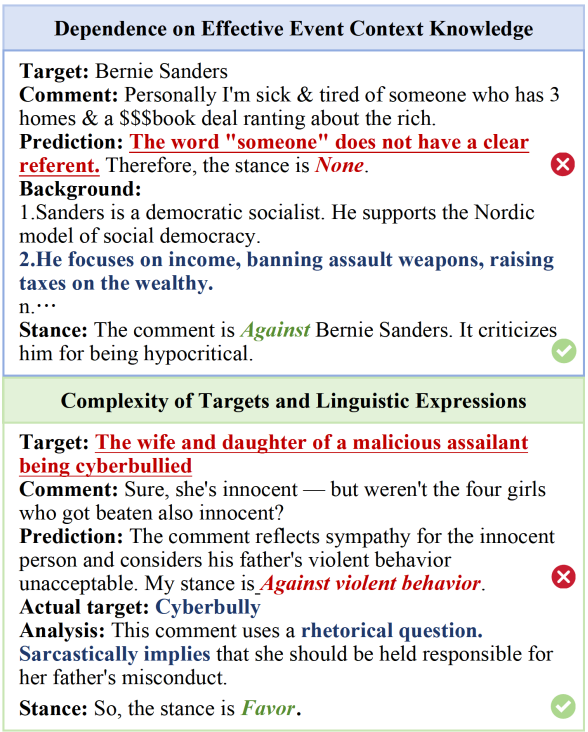

Figure 1: Two cases illustrating the complexity of social media stance detection.

In recent years, LLMs have exhibited remarkable capabilities in language understanding and generation(Touvron et al., 2023), achieved through extensive unsupervised pre-training. This advancement presents a novel technical pathway for zero-shot learning. Consequently, the academic community has begun to explore the potential of LLMs in stance detection, leading to the emergence of various innovative methods, including prompt-based

methods(Ding et al., 2021), knowledge injection techniques tailored for LLMs(Zhang et al., 2024), chain-of-thought reasoning strategies(Wei et al., 2022), and multi-agent frameworks(Chen et al., 2023), among others.

Although these methods have surpassed the supervised models, their inference mechanisms still exhibit flaws, making it challenging to comprehensively address the threefold challenges of zero-shot stance detection: (1) Background Knowledge Reliance: The internal knowledge of LLMs is generally insufficient for stance determination. As shown in the first case in Figure 1, if the model lacks background knowledge related to Bernie Sanders, it does not possess adequate contextual information to ascertain whether 'someone' refers to him; (2)Complex Logical Reasoning Deficiency: Targets in Chinese social media texts often involve intricate events containing multiple objects or subevents. As shown in the second case in Figure 1, the complex target 'the wife and daughter of a malicious assailant being cyberbullied' encompasses multiple objects, where the stance object should be 'cyberbullying,' rather than 'the act of assault' or 'the wife and daughter.' The model must recognize multiple objects and the complex target-stance relationships to infer the stance accurately; (3)Figurative Language Disability: In real-world scenarios, users' expressed stances are often implicit in short texts, cultural metaphors, and complex rhetoric, significantly complicating stance determination. As seen in the second case of Figure 1, a literal interpretation might mistakenly classify the assailant's daughter as innocent, leading to an erroneous stance determination.

To address these challenges, we propose MSME, which achieves high precision and explainable stance detection in zero-shot scenarios through a collaborative mechanism involving multiple experts across multiple stages. The framework consists of three stages: in the knowledge preparation stage, MSME automatically retrieves background knowledge related to the target and constructs clear stance labels. Yin et al. (2024) have demonstrated that explicit stance labels can enhance stance detection. In the multi-expert analysis stage, social media experts focus on constructing fine-grained stance labels. Compared to clear stance labels, fine-grained stance labels further refine stance labels, better addressing the challenge of weak reasoning caused by target complexity. The knowledge reasoning expert focuses on the logical connections between background information and targets, extracting relevant information from raw background knowledge to minimize the interference of irrelevant information on the model. The pragmatics expert concentrates on parsing the implicit influence of rhetorical devices on stance expression, identifying the true stance behind these rhetorical devices. In the final decision-making stage, the decision-maker coordinates the analysis results from multiple experts and ultimately outputs an explainable stance judgment. The code is publicly available at: https://anonymous.4open.science/r/E770124

Our primary contributions are as follows:

- We introduce MSME, a multi-stage, multi-expert zero-shot stance detection framework characterized by a three-stage architecture comprising "preparation, analysis, and decision".
- MSME demonstrates a significant performance improvement over baseline models on the Weibo-SD, SEM16, and P-Stance datasets, achieving an average enhancement of 12% compared to the best baseline. Moreover, the more complex the dataset, the more significant the improvement.
- Experimental results show that collaborative analysis by multiple experts is essential; the removal of any single expert results in a decrease ranging from 2.1% to 7.6%. The social media expert demonstrates more contribution when dealing with complex samples, while the knowledge reasoning expert shows the greater impact in simpler cases.

2 Related Work

2.1 In-target Stance Detection

The evolution of stance detection has progressed from methods based on traditional machine learning(Xu et al., 2016a), to neural networks(Igarashi et al., 2016), and further to pre-trained language models(Hosseinia et al., 2020). For instance, Zhang and Lan (2016) integrated multiple features and employed machine learning such as SVM, RF, and GBDT to achieve single-target stance detection tasks. Taulé et al. (2018) utilized CNN to perform multimodal modeling on text, context, and image information, resulting in more accurate stance detection outcomes. He et al. (2022) proposed WS-BERT, which includes two variants that

leverage background knowledge about targets from Wikipedia to enhance stance detection. However, in real-world scenarios, annotated data is scarce, and the domains are highly variable, complicating the adaptation of traditional methods to the complex needs across domains and targets. This challenge has prompted research to shift towards the more demanding task of zero-shot stance detection.

2.2 Zero-shot Stance Detection

Zero-shot stance detection refers to the model directly inferring the stance of the text toward unseen targets or in the absence of annotated data(Allaway and McKeown, 2020a). Its core challenges include data scarcity, the implicit nature of stance expression (such as irony or rhetorical questions), and cross-domain semantic differences. To address these challenges, existing approaches primarily achieve breakthroughs through knowledge enhancement, transfer learning, and generative models. Zhang et al. (2023b) proposed a self-supervised data augmentation method based on coreference resolution for zero-shot and few-shot stance detection tasks.Liu et al. (2021) proposed a stance detection model, CKE-Net, which integrates commonsense knowledge graphs. By incorporating the external commonsense knowledge graph ConceptNet to construct relational subgraphs, the model enhances its reasoning capabilities for implicit stance expressions. In the realm of transfer learning, the TOAD model employs an adversarial training strategy that compels the model to learn domain-invariant features(Allaway et al., 2021), thereby reducing its reliance on specific targets. However, these methods still necessitate labeled data for model training. In contrast, our approach does not involve training any model parameters and relies entirely on the reasoning and generative capabilities of LLMs.

2.3 LLMs-driven Stance Detection

LLMs have demonstrated exceptional zero-shot capabilities across various tasks, prompting researchers to explore their stance detection abilities. Yuanshuo et al. (2024) conducted a comprehensive investigation into the stance detection capabilities of LLMs based on prompt learning, confirming that explicit stance labels and brief background can enhance stance detection. Li et al. (2023) proposed KASD framework that incorporates situational and discourse knowledge into the task of stance detection on social media. It leverages ChatGPT to extract and integrate these two types of knowledge, resulting in significant improvements for both fine-tuned and LLMs. Taranukhin et al. (2024) introduced Stance Reasoner, which conceptualizes reasoning as explicit inference from premises to conclusions, guiding the model's stance inference through the background knowledge derived from LLMs. Zhang et al. (2024) employed LLMs to extract the relationship between paired texts and targets as contextual knowledge, injecting this LLM-driven knowledge into the generative model BART to enhance stance detection with rich context and semantics. Lan et al. (2024) proposed COLA, a multi-agent collaborative stance reasoning framework, which demonstrates high accuracy, interpretability, and generalizability. In contrast to these approaches, our MSME integrates knowledge, labels, and rhetorical analysis, proposing a unified reasoning framework that exhibits superior effectiveness and interpretability.

3 Multi-Stage Multi-Expert Reasoning

To address the challenges in stance detection, we propose the MSME for zero-shot stance detection. This framework comprises three stages, as illustrated in Figure 2: the preparation stage, the analysis stage, and the decision-making stage.

3.1 Preparation Stage

The preparation stage involves acquiring relevant background information and generating explicit stance labels. Research by Yin et al. (2024) shows that brief background and explicit stance labels can enhance stance detection. We first utilized search APIs and web crawlers to gather relevant background information for each target,as illustrated in the preparation stage of Figure 2. Given the significant impact of explicit stance labels, for complex targets like 'the wife and daughter of a malicious assaulter being cyberbullied,' the favor label should be explicitly defined as 'the wife and daughter of the assaulter deserve to be cyberbullied,' rather than support for the assaulter or his wife and daughter. This approach facilitates the model's reasoning and judgment. Unlike their heuristic approach, we employ LLMs with few-shot prompting to achieve this goal, maintaining accuracy while being more flexible in generating explicit stance labels for arbitrary targets. The inputs and outputs of the prompt templates for all stages are illustrated in Appendix A.

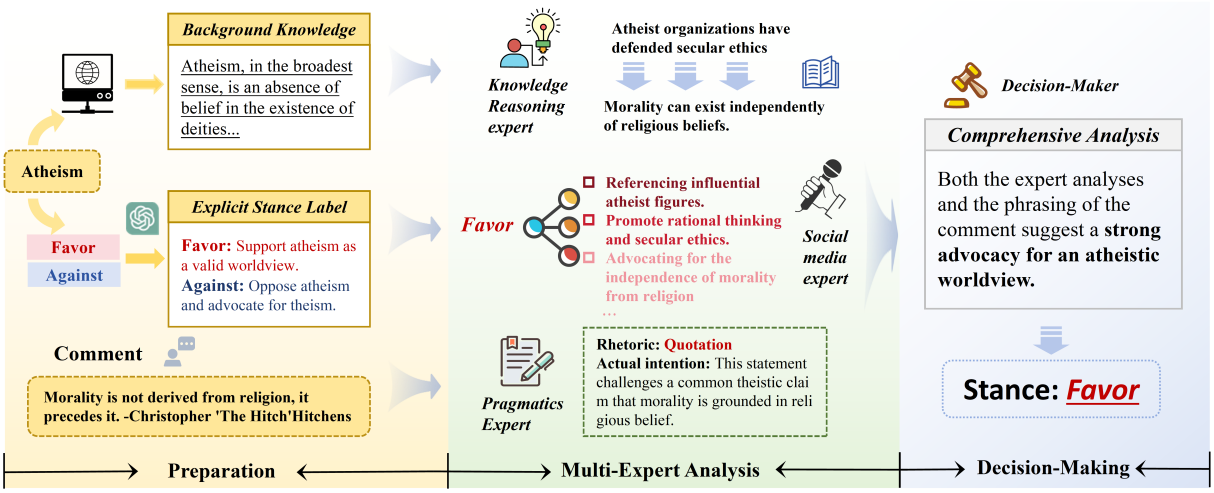

Figure 2: Architecture of our proposed MSME. From left to right are the three stages of our MSME.

3.2 Multi-Expert Analysis Stage

In the analysis stage, we employ three experts to analyze the text from distinct perspectives.

Knowledge Reasoning Expert The knowledge reasoning expert extracts relevant knowledge points from the background information gathered during the preparation stage and conducts a stance analysis based on knowledge reasoning. The background information collected in the preparation stage contains a significant amount of irrelevant data, and incorporating all of it would introduce additional noise. To mitigate this issue, as shown in Figure 2, we introduce the knowledge reasoning expert in a role-playing capacity, requiring it to distill the necessary knowledge points for reasoning the text's stance from the raw background information based on the provided input sample. Subsequently, it analyzes the stance label of the input sample through the lens of knowledge reasoning.

Social Media Expert The social media expert refine the explicit stance labels obtained during the preparation stage to derive more precise, fine-grained stance labels, which are then utilized for stance analysis. Our further analysis reveals that in complex target-stance relationships, explicit stance labels still present ambiguity. For instance, in the target of previously mentioned "cyberbully" , the generated explicit labels include: "the wife and daughter of the assaulter deserve to be cyberbullied," "the wife and daughter of the assailant do not deserve to be cyberbullied," and "neutral." However, the favor stance label encompasses multiple scenarios, such as "believing cyberbully is a justified punishment," "showing more sympathy for the woman who was beaten," and "seeking justice for the victim." To address this complexity, we introduce a social media expert, requiring it to generate fine-grained stance labels based on background information, explicit stance labels, and the comments themselves. This expert will also analyze the stance of given samples based on relevant knowledge points and fine-grained stance labels.

Pragmatics Expert The pragmatics expert address the complex linguistic phenomena present in texts and conduct stance analysis from a pragmatic perspective. In real social media environments, comments often exhibit intricate linguistic phenomena such as irony, metaphor, and sarcasm. These phenomena present significant challenges to stance detection tasks, particularly because the use of rhetorical devices complicates reliance on literal semantics for judgment. To address this, we have introduced pragmatics experts in a role-playing capacity, requiring them to identify and parse the rhetorical devices utilized in the samples, and further analyze their semantic roles and potential intentions within the context, thereby restoring the true stance of the samples.

3.3 Decision-making Stage

Although three experts have conducted in-depth analyses of the text from their respective professional perspectives and arrived at preliminary judgments, relying solely on these individual analyses cannot comprehensively capture the complexity and diversity inherent in the stance detection task. Therefore, in the decision-making stage, we introduced the decision maker, a comprehensive role required to derive the final stance judgment based

Category	Model	A	CC	FM	HC	LA	Avg
In-target	BERT	60.7	38.8	59.0	61.3	63.1	56.6
	CrossNet	56.4	40.1	55.7	60.2	61.3	54.7
	ASGCN	59.5	40.6	58.7	61.0	63.2	56.6
	TPDG	64.7	42.3	67.3	73.4	74.7	64.5
Zero-shot	TOAD	46.1	30.9	54.1	51.2	46.2	45.7
	TGA Net	52.7	36.6	46.6	49.3	45.2	46.1
	BERT-GCN	53.6	35.5	44.3	50.0	44.2	45.5
	JointCL	54.5	39.7	53.8	54.8	49.5	50.5
Zero-shot based on LLMs	Base	58.3	54.1	62.3	72.0	60.8	61.5
	CoT	64.1	59.7	65.4	73.7	61.9	65.0
	BKSL	71.5	66.0	63.1	76.5	64.2	68.3
	Stance reasoner	69.7	62.5	73.9	67.0	64.4	67.8
	COLA	62.3	64.0	69.1	75.9	**71.0**	68.5
	MSME	**76.2**	**73.9**	**75.5**	**79.1**	66.9	**74.8**

Table 1: Comparison of MSME with baselines on SEM16, using GPT-3.5.

on the detailed analyses provided by the three experts(excluding their respective final stance judgments). This approach avoids the model to simply favor the majority expert opinion and ensures that it can independently think and judge based on comprehensive information.

4 Experiment and Analysis

4.1 Experimental Setup

Datasets We conduct experiments on three distinct datasets:

Weibo-SD(Yin et al., 2024): This is a publicly available Chinese stance detection dataset focusing on highly controversial social events. It includes 5 targets from Weibo, totaling 2500 data entries. The dataset is characterized by complex target-stance relationships and each target is a complex event, such as 'the wife and daughter of a malicious assailant being cyberbullied' and 'woman was scolded for not letting 6-year-old boy use girls' restroom'

SEM16:(Mohammad et al., 2016) SemEval-2016 Task 6A focuses on the stance classification task in social media tweets, covering five controversial topics including atheism and climate change. As our experiments are based on zero-shot learning, we conduct experiments only on the test set.

P-Stance:(Li et al., 2021) This is a cross-target stance detection dataset focused on political figures, including social media texts targeting politicians such as Joe Biden and Donald Trump. In the zero-shot setting, experiments are conducted solely on the test set.

Evaluation Metric For the Weibo-SD and P-Stance datasets, we adopt the commonly used Macro-F1 as the evaluation metric(Conforti et al., 2020). For the SEM16 dataset, we follow previous research conventions and report F_{avg}(Allaway and McKeown, 2020b), which is the average of the F1 scores for the Favor and Against labels. We report the average of three experimental runs.

Comparison Methods We compare MSME with SOTA methods, including zero-shot stance detection approaches in both supervised and unsupervised models, as well as supervised models specifically trained for target-specific settings.

Supervised Models: These are typically divided into in-target and zero-shot methods. In-target methods involve training and evaluating the model on the same target. This includes BERT fine-tuned directly(Koroteev, 2021), CrossNet with enhanced attention mechanisms(Xu et al., 2018), and graph learning-based models such as ASGCN(Zhang et al., 2019) and TPDG(Liang et al., 2021). Zero-shot methods refer to training models on data involving unseen targets and then evaluating them on specified targets. These include TGA-Net based on attention(Liang et al., 2022a), TOAD utilizing adversarial learning(Allaway et al., 2021), BERT-GCN based on graph neural networks(Jeong et al., 2020), and JointCL which integrates contrastive learning(Liang et al., 2022b).

Unsupervised Models: These generally refer to methods that leverage LLMs for zero-shot stance detection. This includes approaches like directly

judging stance by inputting the target and text (Base), inferring stance using chain-of-thought (CoT)(Zhang et al., 2023a), determining stance by brief background knowledge and explicit stance labels (BKSL)(Yuanshuo et al., 2024), utilizing logical chains for stance reasoning (Stance Reasoner)(Taranukhin et al., 2024), and employing collaborative frameworks among multiple agents (COLA)(Lan et al., 2024).

Our Model: We implement the MSME framework using GPT-3.5(Ye et al., 2023), GPT-4o(Hurst et al., 2024), QwQ-32B(Zheng et al., 2024), and Deepseek-r1(Guo et al., 2025). All models are accessed via API calls.

4.2 Main Result

Model	Wb-SD	SEM16	P-Stance
Base	51.6	61.5	62.4
CoT	59.8	65.0	68.1
BKSL	67.2	68.3	67.4
SR	61.9	67.8	68.5
COLA	63.1	68.5	69.6
MSME	**75.2**	**74.8**	**76.5**
GPT-4o	78.3	**78.6**	78.5
QwQ-32B	75.7	75.3	78.8
Ds-r1	**78.6**	76.7	**79.1**

Table 2: Results of MSME and baselines base on LLMs on three datasets. The last three rows report the results of MSME on other LLMs. Here, Wb-SD, SR, and Ds-r1 are abbreviations for Weibo-SD, Stance Reasoner, and Deepseek-r1, respectively.

Table 1 presents a comparison of the MSME with various baselines on SEMl16, utilizing the GPT-3.5. In Table 2, we compile a comparison of the MSME against LLMs methods across three datasets, while also displaying the F1 scores of MSME when employing three other LLMs. Our analysis reveals that:

MSME achieves significant improvements over the current SOTA on three datasets. Specifically, it shows an enhancement of 8.0 on the Weibo-SD, 6.3 on the SEM16, and 6.9 on the P-Stance. Furthermore, MSME demonstrates the best performance across the four targets of the SEM16 , with a notable improvement of 12.4 compared to the model operating under the in-target setting. This is attributed to the effective knowledge and fine-grained stance labels that support the reasoning process, as well as the coordinated decision-making facilitated by multiple experts and decision-maker.

The most significant improvement is observed in the Weibo-SD dataset, while the least noticeable enhancement occurs in the SEM16 dataset. This discrepancy can be attributed to the relative simplicity of the SEM16, which also suffers from certain data annotation issues. Conversely, the complexity of the targets in the Weibo-SD necessitates more explicit reasoning.

MSME performs better on LLMs with stronger reasoning capabilities. On Weibo-SD and P-Stance, Deepseek-r1 achieves the highest results, with F1 scores of 78.6 and 79.1, respectively. In contrast, on SEM16, GPT-4o outperforms others with an F1 score of 78.6. Meanwhile, on the more cost-effective GPT-3.5, MSME continues to exhibit stable and competitive performance, underscoring its strong applicability.

4.3 Ablation Study

To evaluate the influence of various experts and decision-maker, we conduct ablation experiments utilizing the GPT-3.5, with the results presented in Table 3, while results from other models are detailed in the Appendix B. Our analysis reveals that:

Compared to the Integration and Vote settings, the F1 score of MSME increased by 3.8 and 1.3 points, respectively. While the Integration setting can streamline the process, it requires the model to manage complex information from multiple perspectives within the single prompt, which complicates the effective balancing of various factors. The Vote setting, fundamentally a simple majority voting mechanism, lacks the capacity to deeply understand and reason with complex contexts and multi-layered information.

In the Weibo-SD, the most significant decline is observed following the removal of the social media expert, resulting in a decrease of 5.3 points. This decline can be attributed to the inherent complexity of the target-stance relationships present in this dataset, where the fine-grained labels provided by the social media expert are essential for effectively addressing this complexity. Conversely, in the Semeval16 and P-Stance , the most notable decline occurs after the removal of the knowledge reasoning expert, with an average drop of 3.6 points. We believe that in these two datasets, the targets are relatively straightforward, consisting of single entities, which diminishes the impact of fine-grained labels while amplifying the significance of knowledge-

Model	Weibo-SD	SEM16	P-Stance
MSME	**75.2**	**74.8**	**76.5**
Integration	70.8	71.5	72.7
Vote	**73.7**	**74.1**	**74.9**
w/o Knowledge Reasoning Expert	70.9	72.0	72.1
w/o Social Media Expert	69.9	72.6	**73.2**
w/o Pragmatics Expert	**71.4**	**73.2**	72.3
Knowledge Reasoning Expert Only	70.1	**72.8**	**73.0**
Social Media Expert Only	**70.6**	71.5	71.9
Pragmatics Expert Only	69.5	70.8	71.2

Table 3: Experimental results of ablation study. In the Integration setting, we combine multi-expert analysis stage and decision-making stage into a unified process, completing all tasks through a single prompt. In the Vote setting, we aggregate the independent judgments of each expert and employ a majority voting mechanism to determine the final stance. In the Expert Only setting, only a single expert is retained for independent judgment.

based reasoning.

When retaining only one expert, Weibo-SD demonstrates optimal performance with the social media expert, whereas SEM16 and P-Stance achieve their best results with the knowledge reasoning expert. This observation further corroborates the previous conclusion.

The decline resulting from the removal of the pragmatics expert is relatively minor, while the performance is at its lowest when only the pragmatics expert is retained. However, this does not imply that pragmatics analysis lacks significance. This is due to the fact that only a portion of the dataset contains rhetorical elements, which consequently limits its overall impact. In the subsequent experiment, we further substantiate this point.

4.4 Analysis of Neutral Stance

From the confusion matrix of our preliminary experiments(Appendix B), we observed that existing methods perform determining in identifying neutral stances. To investigate the efficacy of our method in determining neutral stances, we conducted a dedicated analysis. Figure 4 illustrates the F1 scores for neutral stances across three datasets. The results show a significant enhancement in the assessment of neutral stances under the Social Media Expert and MSME settings, with an average increase of 10.6 compared to the setting without the Social Media Expert. We attribute this improvement to the development of fine-grained stance labels. In contrast to ternary classification labels, the fine-grained division offers a more interpretable sub-label system, which can more accurately reflect the degree of variation in samples' favor or against stances.

Target: The wife and daughter of an individual who has maliciously assaulted others were
Comment: I just wanna know the kids' info of the other criminals too.

Fine-Grained Stance Label:
A. Favor:
a. Believes that cyberbullying the wife and daughter is a deserved punishment for the assailant
b. Advocates exposing the assailant's family members to exert pressure on relatives
c. Views cyberbullying as a justified act of retaliation
B. Against:
a. Considers the wife and daughter of the assailant to be innocent
b. Believes that cyberbullying unrelated individuals is unethical
c. Argues that cyberbullying does not resolve issues but instead causes further harm
C. Neutral/Irrelevant
Analysis: The comment expresses a desire to obtain information about the children of other offenders, **aligning with the stance that supports extending the consequences of violent actions to family members**. Specifically, this indicates support for exerting pressure on the assailant's family by exposing their personal information.
Stance: A.b--Advocates exposing the assailant's family members to exert pressure on relatives

Figure 3: Cases of explainations generated by social media expert.

This not only improves the accuracy of stance expression but also clarifies the definition of a neutral stance. When the model is unable to categorize a comment into any fine-grained favor or against stance, it is classified as neutral.

In the case illustrated in Figure 3, the sample "I just wanna know the kids' info of the other criminals too" was mistakenly classified by the model as neutral, as it neither directly supports nor explicitly opposes cyberbullying. However, among the generated fine-grained stance labels, one supporting label "Advocates exposing the assailant's family members to exert" aligns with the comment's intent, thereby inferring that the stance of this sample is favor.

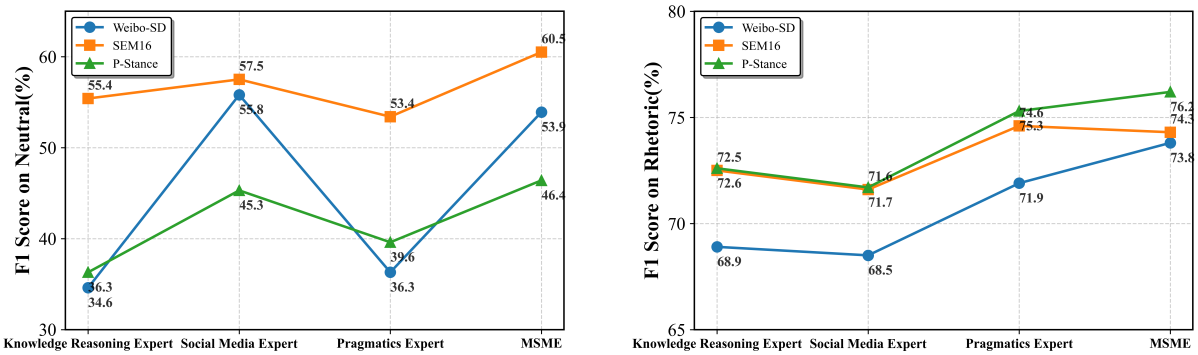

Figure 4: On the left is the result of neutral stance analysis, and on the right is the result of rhetorical analysis.

4.5 Judgment of Rhetorical Expressions

To evaluate the capability of MSME in processing complex rhetorical texts, we employed three LLMs with strong reasoning abilities(GPT-4o, QwQ-32B, and Deepseek-r1). These models are utilized to determine the presence of rhetoric within the texts. We employ a majority voting mechanism to select the final set of comments containing rhetorical expressions. Statistical data show that the proportions of texts containing rhetoric in the Weibo-SD, SEM16, and P-Stance datasets were 80.6%, 52.2%, and 69.6%, respectively. Figure 4 illustrates the F1 scores for texts containing rhetoric across different settings. The results reveal that the introduction of pragmatic expert significantly improves the models' accuracy in identifying rhetorical texts, with an average increase of 3.4 in F1 scores. Particularly within the context of the MSME framework, the evaluation of complex rhetoric surpasses that achieved by relying solely on pragmatics expert. This finding further demonstrates the advantages of the MSME in the realm of information fusion.

Target: Atheism
Comment: Imagine a species that had split the atom and ventured into space yet most still believe a magic man created everything with magic

Analysis:
The comment employs **metaphor** by juxtaposing the advanced technological achievements of a species (splitting the atom and venturing into space) with the belief in a "magic man" who created everything. This contrast highlights a perceived absurdity in holding onto a belief in a deity while simultaneously achieving significant scientific progress. Overall, the comment **critiques theism and implies a preference for atheism as a more rational worldview.**
Stance: A. Favor: Support atheism as a valid worldview.

Figure 5: Cases of explainations generated by pragmatics expert.

In the case illustrated in Figure 5, the comment metaphorically compares God to a 'magic man' who creates everything and likens the splitting of atoms and venturing into space to modern science. It satirizes those who prefer belief in God over scientific understanding, implying that atheism represents a more rational worldview, thereby inferring that the stance of this sample is favor.

5 Conclusion

In this work, we propose MSME to address the challenges faced by current zero-shot stance detection based on LLMs, including 1) reliance on background knowledge, 2) deficiency in complex logical reasoning, and 3) disability in handling figurative language. The MSME encompasses three stages: preparation, analysis, and decision-making. During the analysis stage, multiple experts are employed to conduct knowledge analysis, stance label analysis, and rhetorical strategy analysis, ultimately collaborating to infer the stance. We conducted a series of experiments using MSME on Weibo-SD, SEMl16, and P-Stance, and the results indicate that MSME achieved optimal performance across all three datasets. The more complex the dataset, the more significant the performance improvement observed. Additionally, MSME demonstrates strong applicability, delivering commendable results in both low-cost and reasoning models. Ablation studies reveal that MSME achieves a fusion of multi-expert analysis results, and the removal of any expert results in a decline. Furthermore, the social media expert plays a more significant role when dealing with samples that exhibit complex target-stance relationships, while the knowledge reasoning expert contributes more effectively when addressing simpler targets. Additional experiments indicate that MSME is better equipped to make judgments on neutral samples and those with complex rhetorical structures.

Limitations

Cost The design of MSME incurs significant computational and invocation costs. Each sample processed necessitates invoking the LLMs API at least four times, although the preparation stage for the same target can be reused. This requirement poses challenges related to resource consumption and response latency.

Real Scenario Existing stance detection tasks typically involve predefined targets, often limited to a single target, which renders them inadequate for scenarios lacking preset targets or involving multiple targets. Our method is also applicable to such scenarios. To accommodate situations without predefined targets and those involving multiple targets, we need to extract potential target objects from the comments during the preparation stage. Subsequently, we need to analyze and make decisions for each target during the analysis and decision-making stage.

References

Emily Allaway and Kathleen McKeown. 2020a. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Emily Allaway and Kathleen McKeown. 2020b. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection. *arXiv preprint arXiv:2204.03839*.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40.

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*.

Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2016. Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 401–407.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

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.

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.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM web conference 2022*, pages 2738–2747.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the web conference 2021*, pages 3453–3464.

Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, volume 1, pages 81–91. Association for Computational Linguistics.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

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, San Diego, California. Association for Computational Linguistics.

Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. *arXiv preprint arXiv:2403.14895*.

Mariona Taulé, Francisco M Rangel Pardo, M Antònia Martí, and Paolo Rosso. 2018. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In *IberEval@ SEPLN*, pages 149–166. Sevilla.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. *arXiv preprint arXiv:1805.06593*.

Jiaming Xu, Suncong Zheng, Jing Shi, Yiqun Yao, and Bo Xu. 2016a. Ensemble of feature sets and classification methods for stance detection. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 679–688. Springer.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016b. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Proceedings of the ICCPOL 2016*, pages 907–916.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, and 1 others. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Zhaoning Yin, Yuanshuo Zhang, Aohua Li, Bo Chen, and Xiaobing Zhao. 2024. Fusion of stance labels and event knowledge for effective stance detection. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 203–207. IEEE.

Zhang Yuanshuo, Li Aohua, Yin Zhaoning, Wang Panyi, Chen Bo, and Zhao Xiaobing. 2024. Research on stance detection with generative language model. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 481–491.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li, and Liwen Jing. 2023a. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Jiarui Zhang, Shaojuan Wu, Xiaowang Zhang, and Zhiyong Feng. 2023b. Task-specific data augmentation for zero-shot and few-shot stance detection. In *Companion proceedings of the ACM web conference 2023*, pages 160–163.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. Llm-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378.

Zhihua Zhang and Man Lan. 2016. Ecnu at semeval 2016 task 6: Relevant or not? supportive or not? a two-step learning system for automatic detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 451–457.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

A Prompt Template

Construction of Explicit Stance Labels

input:
Please design Explicit Stance Labels for the target "<target>".
<example>*4
output:
A. Favor: <Explicit Stance Label for Favor>
B. Against: <Explicit Stance Label for Against>
C. Neutral/None

Figure 6: Input and output examples of the prompt template for generating explicit stance labels.

Knowledge Reasoning Expert

input:
To determine the stance of the comment to the target "**<target>**", which information is necessary? Please retain useful background information, analyze how retained information influences the judgment of the stance.
Background: **<Background>**
Comment: **<Comment>**
Explicit Stance Labels: **<Explicit Stance Labels>**
output:
1.<Information 1>--><Analysis 1>
2.<Information 2>--><Analysis 2>
…
Stance: **<Choose from Explicit stance labels>**

Figure 7: Input and output examples of the prompt template for knowledge reasoning expert.

B Other Results

Social Media Expert

input:
To reflect the degree or reasons of Favor and Against towards the target "**<target>**" , the clear labels can be further divided into several
fine-grained stance labels. Please subdivide the clear stance label according to the comment and analyze the stance of the comment.
Background: **<Background>**
Comment: **<Comment>**
Explicit Stance Labels: **<Explicit Stance Labels>**
output:
Fine-grained Stance Labels:
A. Favor: B. Against: C. Neutral/None
a. b. c. ... a. b. c. ...
Analysis: **<Analysis>**
Stance: **<Choose from Explicit stance labels>**

Figure 8: Input and output examples of the prompt template for knowledge reasoning expert.

Pragmatics Expert

input:
Please analyze the rhetorical devices contained in the comments regarding the target "**<target>**" to understand the actual intention of the comment.
Background: **<Background>**
Comment: **<Comment>**
Explicit Stance Labels: **<Explicit Stance Labels>**
output:
Analysis: **<Rhetoric>** --> **<Actual intention>**
Stance: **<Choose from Explicit stance labels>**

Figure 9: Input and output examples of the prompt template for pragmatics expert.

Decision-Maker

input:
Please combine the analysis of three experts, make your own analysis, and determine the stance of the comment on the target "**<target>**".
Comment: **<Comment>**
<Analysis of the three experts>
output:
Analysis: **<Analysis>**
Stance: **<label>**

Figure 10: Input and output examples of the prompt template for decision-maker.

Model	Weibo-SD	SEM16	P-Stance
MSME	**78.3**	**78.6**	**78.5**
Integration	76.3	76.7	73.8
Vote	**77.5**	**78.1**	**76.3**
w/o Knowledge Reasoning Expert	**76.1**	75.0	74.1
w/o Social Media Expert	73.9	**76.1**	74.7
w/o Pragmatics Expert	75.7	75.9	**75.3**
Knowledge Reasoning Expert Only	74.6	**76.9**	**75.0**
Social Media Expert Only	**77.1**	75.5	73.9
Pragmatics Expert Only	74.7	73.3	74.2

Table 4: Ablation study results on GPT-4o.

Model	Weibo-SD	SEM16	P-Stance
MSME	**75.7**	**75.3**	**78.8**
Integration	70.5	71.2	74.7
Vote	**74.1**	**74.0**	**76.9**
w/o Knowledge Reasoning Expert	**72.5**	72.2	74.3
w/o Social Media Expert	70.6	73.3	**75.1**
w/o Pragmatics Expert	72.3	**73.5**	73.3
Knowledge Reasoning Expert Only	70.1	**72.2**	**74.4**
Social Media Expert Only	**73.0**	71.3	73.8
Pragmatics Expert Only	71.5	69.8	72.2

Table 5: Ablation study results on QwQ-32B.

Model	Weibo-SD	SEM16	P-Stance
MSME	**78.6**	**76.7**	**79.1**
Integration	74.7	72.5	75.2
Vote	**77.3**	**74.8**	**78.3**
w/o Knowledge Reasoning Expert	**75.9**	72.0	75.2
w/o Social Media Expert	74.5	72.6	75.1
w/o Pragmatics Expert	75.3	**73.2**	**76.3**
Knowledge Reasoning Expert Only	75.2	**73.8**	**76.1**
Social Media Expert Only	**76.1**	72.1	75.3
Pragmatics Expert Only	73.5	71.7	73.4

Table 6: Ablation study results on Deepseek-r1.

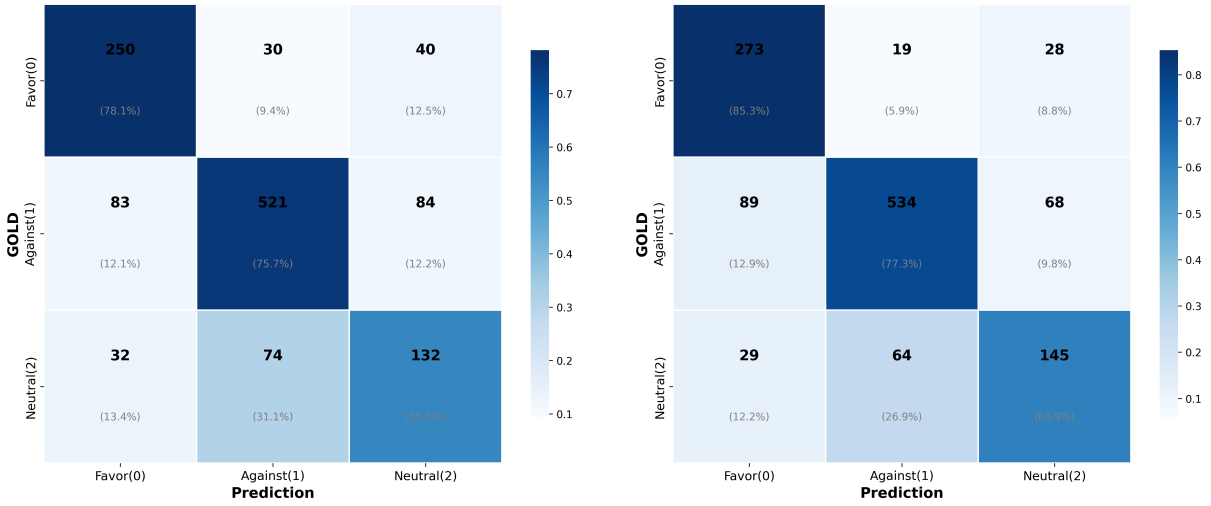

Figure 11: From left to right: the confusion matrices of the Pragmatics Expert Only and MSME on the SEM16 dataset using GPT-3.5.