MSME: A Multi-Stage Multi-Expert Framework for Zero-Shot Stance Detection

Yuanshuo Zhang[1,2], **Aohua Li**[1,2], **Bo Chen**[1,2*], **Jingbo Sun**[1,2*], **Xiaobing Zhao**[1,2]

[1]School of Information Engineering, Minzu University of China
[2]National Language Resource Monitoring and Research Center of Minority Languages
23302158@muc.edu.cn, 23302161@muc.edu.cn, chenbomuc@muc.edu.cn,
sunjingbo@muc.edu.cn, nmzxb_cn@163.com

Abstract

LLM-based approaches have recently achieved impressive results in zero-shot stance detection. However, they still struggle in complex real-world scenarios, where stance understanding requires dynamic background knowledge, target definitions involve compound entities or events that must be explicitly linked to stance labels, and rhetorical devices such as irony often obscure the author's actual intent. To address these challenges, we propose MSME, a **M**ulti-**S**tage, **M**ulti-**E**xpert framework for zero-shot stance detection. MSME consists of three stages: (1) *Knowledge Preparation*, where relevant background knowledge is retrieved and stance labels are clarified; (2) *Expert Reasoning*, involving three specialized modules—Knowledge Expert distills salient facts and reasons from a knowledge perspective, Label Expert refines stance labels and reasons accordingly, and Pragmatic Expert detects rhetorical cues such as irony to infer intent from a pragmatic angle; (3) *Decision Aggregation*, where a Meta-Judge integrates all expert analyses to produce the final stance prediction. Experiments on three public datasets show that MSME achieves state-of-the-art performance across the board.

Code — https://github.com/zy-shuo/MSME

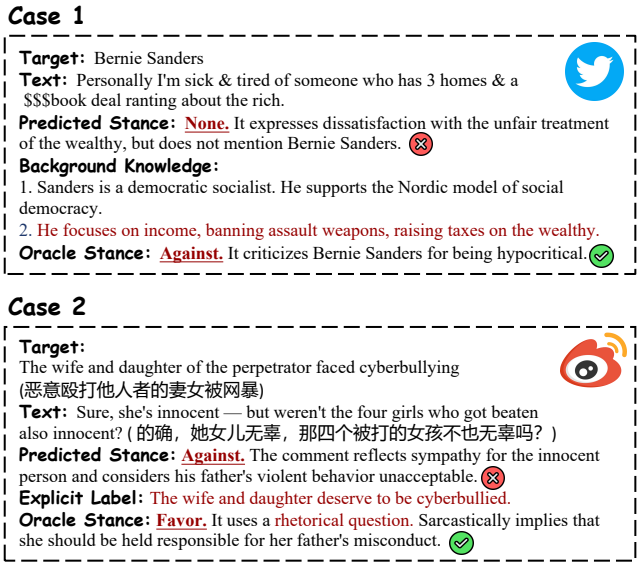

Figure 1: Examples from SEM16 (Mohammad et al. 2016) (Case 1) and Weibo-SD (Zhang et al. 2024a)(Case 2, translated from Chinese) illustrating real-world stance detection.

1 Introduction

Stance detection aims to identify whether a text expresses a *Favor*, *Against*, or *Neutral* perspective toward a specified target (e.g., an event or entity) (Mohammad et al. 2016). As social media continually spawns diverse and fast-evolving topics, traditional approaches relying on copious domain-specific annotations become impractical (Xu et al. 2016b; Hardalov et al. 2021). Zero-shot stance detection tackles this limitation by enabling stance reasoning on unseen or emerging targets (Liang et al. 2022a; Allaway and McKeown 2020a), leveraging either knowledge transfer or inherent zero-shot capabilities. However, even zero-shot paradigms require annotated data to learn stance patterns.

The advent of LLMs (Touvron et al. 2023) has inspired a range of new approaches for zero-shot stance detection, including prompt-based classification (Ding et al. 2021), leveraging LLMs as external knowledge sources (Zhang et al. 2024b), framing the task as logical reasoning (Wei et al. 2022; Taranukhin, Shwartz, and Milios 2024), and employing multi-agent collaboration frameworks (Chen et al. 2023).

Despite demonstrating strong performance on standard benchmarks, existing zero-shot stance detectors remain ill-equipped for the nuances of real-world discourse. They face three primary challenges: (1) Background Knowledge Dependency: Accurate stance interpretation often hinges on up-to-date world knowledge. For example, to recognize that the disparaging '*someone*' in Fig. 1 refers to Bernie Sanders and the author is criticizing his anti-wealthy policy stance, an external understanding of contemporary political debates is required. (2) Unclear Target-Label Mapping: Real-world targets frequently comprise compound entities or multi-faceted events. Consider the statement '*The wife and daughter of the perpetrator faced cyberbullying*' (Fig. 1). A model must discern whether the stance refers to

*Corresponding author

the act of cyberbullying itself, i.e., "*supporting cyberbullying*," rather than erroneously attributing support to the perpetrator or his family. Existing approaches struggle to decompose such compound targets and align them precisely with the intended stance label. (3) Pragmatic Complexity: Social media is rife with rhetorical expressions such as irony, sarcasm, and metaphor, which obscure literal sentiment. A comment like "*she's innocent*" may, in context, convey the opposite of its surface meaning (Fig. 1). Without specialized pragmatic analysis, LLMs tend to default to literal interpretations, leading to systematic misclassification of ironic or sarcastic stances.

To address these challenges, we propose MSME (**M**ulti-**S**tage **M**ulti-**E**xpert), a zero-shot stance detection framework, which consists of three stages: In the *Knowledge Preparation* stage, we retrieve target-specific background knowledge to compensate for the static nature of LLMs' internal knowledge. We also clarify stance labels by explicitly defining what constitutes Favor or Against for the given target, thereby narrowing the semantic space and reducing ambiguity in target-label mapping. The *Expert Reasoning* stage engages three specialized experts. The **Knowledge Expert** filters retrieved knowledge, removing irrelevant content based on the target text. For each retained item, the expert reasons from a knowledge perspective and derives a clear conclusion. The refined knowledge then serves as shared context for all experts. The **Label Expert** constructs a fine-grained stance taxonomy based on clarified labels, reasoning from these detailed labels to reduce uncertainty in the mapping between targets and labels. The **Pragmatic Expert** identifies pragmatic patterns in the text such as irony or sarcasm and infers the author's actual intent beyond literal interpretation. Finally, in the *Decision Aggregation* stage, a **Meta-Judge** integrates the analyses from all experts, weighing knowledge-based reasoning, label mapping, and pragmatic interpretation to produce the final stance prediction. Our main contributions are:

- We propose the MSME, the first framework to specifically address zero-shot stance detection in complex real-world scenarios.
- Extensive experiments on three datasets: SEM16, P-Stance, and Weibo-SD, demonstrate that MSME achieves state-of-the-art results, with ablation studies validating the necessity of each expert.
- We find that the label expert performs exceptionally well on complex targets (Weibo-SD and *Climate Change Is Real Concern* in SEM16). This success is attributed to the fine-grained stance label system, which clarifies the mapping between targets and labels.

2 Related Work

2.1 In-target Stance Detection

Stance detection has evolved from machine learning (Xu et al. 2016a) to neural networks (Igarashi et al. 2016) and further to pre-trained language models (Hosseinia, Dragut, and Mukherjee 2020). Early work (Zhang and Lan 2016) combined multiple features with ensemble classifiers (SVM/RF/GBDT) for single-target detection, while later approaches leveraged CNN and LSTM (Taulé et al. 2018; Dey, Shrivastava, and Kaushik 2018) to model texts and targets. He, Mokhberian, and Lerman (2022) proposed WS-BERT, which inject Wikipedia-derived target knowledge into BERT to enhance accuracy. However, scarce annotated data and domain divergence limit traditional methods' adaptability, driving increased focus on zero-shot stance detection.

2.2 Zero-shot Stance Detection

Zero-shot stance detection refers to inferring stance toward unseen targets or in the absence of annotated data (Allaway and McKeown 2020a), confronting three key challenges: data scarcity, implicit expressions (e.g., irony/rhetorical questions), and cross-domain semantic differences. Current solutions focus on knowledge enhancement and transfer learning. For instance, Zhang et al. (2023b) proposed a self-supervised data augmentation method based on coreference resolution for zero-shot and few-shot stance detection. Liu et al. (2021) introduced CKE-Net, a model integrates commonsense knowledge graphs. By using ConceptNet to construct relational subgraphs, it enhances reasoning over implicit expressions. In transfer learning, the TOAD model employs adversarial training to learn domain-invariant features (Allaway, Srikanth, and McKeown 2021), reducing dependence on specific targets. However, these still require labeled data, whereas our approach enables parameter-free inference by leveraging LLMs' inherent reasoning and generation capabilities (Chang et al. 2024).

2.3 LLMs-based Stance Detection

LLMs have shown strong zero-shot capabilities across various tasks, motivating researchers to explore their applications in stance detection. Zhang et al. (2024a) systematically studied LLMs' stance detection performance using prompt learning, showing that explicit stance labels and brief background information can improve accuracy. Li et al. (2023) proposed the KASD framework, which leverages situational and discourse knowledge for stance detection via ChatGPT, resulting in notable performance gains for both fine-tuned models and LLMs. Taranukhin, Shwartz, and Milios (2024) introduced Stance Reasoner, modeling reasoning as an explicit inference from premises to conclusions. It guides stance inference using background knowledge generated by LLMs. Zhang et al. (2024b) employed LLMs to extract relationships between texts and targets as contextual knowledge, which was then injected into the generative model BART to enhance stance detection with richer context and semantics. Lan et al. (2024) proposed COLA, a multi-agent collaborative framework for stance reasoning, demonstrating high accuracy, interpretability, and generalization. Weinzierl and Harabagiu (2024) constructed counterfactual tree prompts to guide LLMs in generating explanations for each of the three stance labels, based on which the final stance is determined. Unlike these methods, our MSME is specifically designed to adapt to stance detection in real-world scenarios.

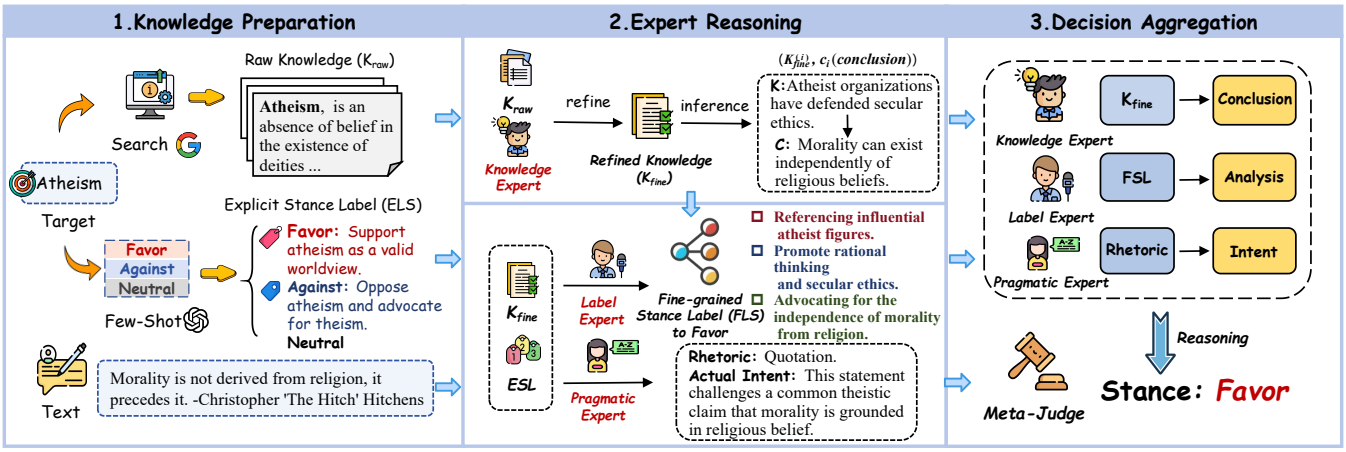

Figure 2: Architecture of our proposed MSME, illustrated with a sample from SEM16.

3 Multi-Stage Multi-Expert Reasoning

Given the complexity of the three challenges faced by stance detection in real-world scenarios, an end-to-end approach struggles to effectively address all aspects. Therefore, we decompose our solution into three steps. The MSME framework consists of three stages (Fig. 2). In the *Knowledge Preparation* stage, relevant background is retrieved and stance labels are clarified. The *Expert Reasoning* stage includes three experts, each reasoning from a different perspective. In the *Decision Aggregation* stage, the analyses from all experts are integrated to produce a final stance.

3.1 Stage 1: Knowledge Preparation

In this stage, MSME retrieves raw background knowledge K_{raw} and constructs explicit stance labels (ESL). For a target t, we query related topics using a Search API, extract core texts from the results automatically, and segment it into chunks $k_1, k_2, \ldots, k_n$. Redundant chunks are removed based on embedding similarity, and the top-3 most relevant segments are concatenated to form K_{raw}. To generate ESL, we employ LLMs with few-shot prompting (Fig. 3).

Instruction: Please design Explicit Stance Labels for the target "The wife and daughter of the perpetrator faced cyberbullying"
Input: [Examples]*4
Example 1:
Target: College professor dismissed due to controversial remarks
ESL_{Favor}: The professor deserved to be dismissed
$ESL_{Against}$: The professor should not have been dismissed
Example 2: ...
Output:
ESL_{Favor}: The wife and daughter deserve to be cyberbullied
$ESL_{Against}$: The wife and daughter should not be cyberbullied

Figure 3: Few-shot prompt to generate explicit stance labels.

3.2 Stage 2: Expert Reasoning

In this stage, we use prompts to instruct LLMs to assume the roles of three specialized experts, reasoning about the stance from knowledge, label, and pragmatic perspectives. Simplified prompt templates are shown in Fig. 4 .

You are a rigorous **knowledge expert**.
Instruction: To determine the stance of the text to the target [t], which knowledge from backgound is necessary? Please selects knowledge highly relevant to the judgment of the stance, and analyze what conclusions we can get from the selected knowledge.
Input: [K_{raw}], [T], [ESL]
Output:
Refined Knowledge: [K_{fine}]
Reasoning: 1. [$K_{fine}^{(1)}$] → [c_1] 2. [$K_{fine}^{(2)}$] → [c_2] ⋯ **Stance:** [ESL_j]

You are a skilled **label expert**.
Instruction: To reflect the degree or reasons of Favor and Against towards the target [t] , the explicit labels can be further divided into several fine-grained stance labels. Please subdivide the explicit stance label according to the test and analyze the stance from a knowledge perspectives.
Input: [K_{fine}], [T], [ESL]
Output:
Fine-grained Stance Labels:
A. $FSL_{Fav} \in \{FSL_{Fav}^{(1)}, FSL_{Fav}^{(2)}, \cdots\}$ B. $FSL_{Aga} \in \{FSL_{Aga}^{(1)}, \cdots\}$ C. Neutral/None
Analysis: [Analysis] **Stance:** [FSL_j^m]

You are a careful **pragmatic expert**.
Instruction:
Please analyze the rhetorical devices contained in the text regarding the target [t], to understand the actual intent of the text.
Input: [K_{fine}], [T], [ESL]
Output:
Rhetoric: [R] **Intent:** [I] **Stance:** [ESL_j]

Figure 4: Simplified prompt templates for the three experts: Knowledge Expert, Label Expert, and Pragmatic Expert.

Knowledge Expert To reason about stance from a knowledge perspective, it is necessary to reduce noise in K_{raw}. The knowledge expert extracts salient information to enhance reasoning accuracy. Given a text T and target t, it selects segments from K_{raw} most relevant to T, yielding refined knowledge $K_{fine} = k_{fine}^{(1)}, k_{fine}^{(2)}, \ldots$. The K_{fine} is shared across experts. For each piece of knowledge $k_{fine}^{(i)}$, reasoning from a knowledge perspective generates a conclusion c_i, forming the pair $(k_{fine}^{(i)}, c_i)$. This chain-of-thought-

like process can enhance reasoning capability.

Label Expert To enable more precise stance reasoning from a labeling perspective, it is essential to reduce the ambiguity of explicit labels. Therefore, the label expert constructs a fine-grained stance label system (FSL) derived from ESL. While ESL reduces coarse label–target mismatches, its labels may still cover multiple nuanced positions and reflecting varying degrees of stance or underlying motivations. For example, an ESL_{Favor} label like '*The wife and daughter deserve to be cyberbullied*' can imply justified punishment of harm, sympathy for victims, or seeking victim justice. Formally, given a text T, target t, and refined knowledge K_{fine}, the expert refines each ESL_j ($j \in$ Favor, Against) into sub-label sets FSL_j^m ($m = 1, \ldots, n$) and infers the stance by selecting the most appropriate FSL.

Pragmatic Expert To mitigate the impact of rhetorical complexity on stance inference, the pragmatic expert uncovers the author's actual intent behind figurative language. Specifically, given a text T, target t, and knowledge K_{fine}, it first detects rhetorical patterns R in T. If R is present, the expert extracts R and infers the underlying actual intent I; otherwise, it directly analyzes the literal intent.

3.3 Stage 3: Decision Aggregation

In this stage, the meta-judge integrates the outputs of all experts to produce the final stance. Specifically, given text T, target t, explicit labels ESL and refined knowledge K_{fine}, it synthesizes: knowledge–conclusion pairs $(k_{fine}^{(i)}, c_i)$ from the knowledge expert, fine-grained labels and analyses from the label expert, and rhetorical pattern R with inferred actual intent I from the pragmatic expert. The meta-judge then outputs the final stance s along with a transparent reasoning process (Fig. 5).

You are a **fair judge.**

Instruction: Please combine the analysis of three experts, make your own analysis, and determine the stance of the comment on the target [t].

Input: [K_{fine}], [T], [ESL]

Knowledge Expert: ($K_{fine}^{(i)}$, C_i)

Label Expert: [FSL_j^m] , [Analysis]

Pragmatic Expert: [R] , [I]

Output:

Reasoning Process: [Reasoning Process] **Stance:** [s]

Figure 5: Simplified prompt for the Meta-Judge in the Decision Aggregation stage.

4 Experiments

4.1 Setup

Datasets We evaluate MSME on two widely used English benchmarks and one complex Chinese dataset to comprehensively validate the framework's effectiveness.

SEM16 (SemEval-2016 Task 6A) (Mohammad et al. 2016) focuses on stance classification in social media tweets, categorized into *Favor*, *Against*, and *Neutral*. It contains tweets annotated for stance toward five targets: *Atheism* (A), *Climate Change Is Real Concern* (CC), *Feminist Movement* (FM), *Hillary Clinton* (HC), and *Legalization of Abortion* (LA), with 1,249 test instances.

P-Stance (Li et al. 2021) focuses on stance toward political figures: *Donald Trump* (DT), *Joe Biden* (JB), and *Bernie Sanders* (BS), with 2,156 test instances. It includes only two stance categories, *Favor* and *Against*, and is commonly used for zero-shot and cross-target stance detection tasks.

Weibo-SD (Zhang et al. 2024a) consists of 1,698 comments from Weibo, focusing on hot social media events. The stance in the dataset is categorized into *Favor*, *Against*, and *Neutral*. The five compound targets are: *The wife and daughter of the perpetrator faced cyberbullying* (CB), *Woman refused to let a 6-year-old boy enter the female restroom and was criticized* (FR), *Police confirmed Hu Xinyu's suicide* (HS), *The movie Manjianghong's official sues prominent influencers of weibo* (MM), and *Water Splash Festival woman forgives the offender* (FO). Each target is accompanied by brief background knowledge and explicit stance labels to provide context for stance classification.

To enable comparison with existing work, we evaluate only on the official test sets for SEM16 and P-Stance.

Evaluation Metrics Metrics For the SEM16 and P-Stance datasets, we follow prior work and report the average F1 score (F_{avg}) for the Favor and Against labels (Allaway and McKeown 2020b). For the Weibo-SD dataset, we use the commonly adopted Macro-F1 metric (Conforti et al. 2020).

Implementation Details We use SerpAPI[1] for retrieval and BGE model (Xiao et al. 2023) for text embeddings. In our experiments, we employ four models: GPT-3.5 (Ye et al. 2023) (standard model), and three inference models—GPT-4o (Hurst et al. 2024), QWQ-32B (Zheng et al. 2024), and DeepSeek-R1 (Guo et al. 2025), all accessed via API. To ensure result stability and reproducibility, the model temperature is set to 0. Results are reported as the average of three experimental runs.

Baselines We compare MSME with state-of-the-art methods, including supervised models for in-target tasks and zero-shot approaches (supervised and unsupervised).

In-Target Models: These models are trained and evaluated on the same target. They include CrossNet (Xu et al. 2018) with enhanced attention, BERT (Koroteev 2021) fine-tuned directly, and graph-based models such as ASGCN (Zhang, Li, and Song 2019) and TPDG (Liang et al. 2021).

Zero-Shot Models: These models are trained on data with specified targets and evaluated on unseen targets. Notable methods include TGA-Net (Liang et al. 2022a) based on attention, TOAD (Allaway, Srikanth, and McKeown 2021) utilizing adversarial learning, BERT-GCN (Jeong et al. 2020) based on graph neural networks, and JointCL (Liang et al. 2022b) integrating contrastive learning.

Zero-Shot Based on LLMs: These methods leverage LLMs for zero-shot stance detection. Approaches include direct stance inference by inputting the target and text (Base), stance reasoning using chain-of-thought (CoT) (Zhang et al. 2023a), stance determination with brief background knowledge and explicit stance labels (BKEL)

[1] https://serpapi.com/

Category	Model	SEM16 (%)						P-Stance (%)				Weibo-SD (%)					
		A	CC	FM	HC	LA	Avg	DT	JB	BS	Avg	CB	FR	HS	MM	FO	Avg
In-target	CrossNet	56.4	40.1	55.7	60.2	61.3	54.7	58.0	65.0	53.0	58.7	–	–	–	–	–	–
	BERT	60.7	38.8	59.0	61.3	63.1	56.6	67.7	73.1	68.2	69.7	50.7	46.9	53.2	50.4	62.4	52.7
	ASGCN	59.5	40.6	58.7	61.0	63.2	56.6	77.0	78.4	70.8	75.4	–	–	–	–	–	–
	TPDG	64.7	42.3	67.3	73.4	74.7	64.5	76.8	78.1	71.0	75.3	–	–	–	–	–	–
Zero-shot	TGA Net	56.1	52.9	61.2	60.2	55.7	57.2	–	–	–	–	–	–	–	–	–	–
	TOAD	46.1	30.9	54.1	51.2	46.2	45.7	53.0	68.4	62.9	61.4	–	–	–	–	–	–
	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	62.0	59.0	73.0	64.7	–	–	–	–	–	–
Zero-shot based on LLMs	Base*	58.3	51.1	62.3	65.0	60.8	59.5	67.3	78.2	71.6	72.4	40.1	45.2	56.9	52.2	46.9	48.3
	CoT*	64.1	55.7	62.4	70.7	61.9	63.0	71.4	80.5	74.1	75.3	36.2	53.9	58.1	58.2	50.7	51.4
	BKEL*	71.5	66.0	63.1	76.5	64.2	68.3	80.3	78.3	79.6	79.4	58.0	51.3	65.6	67.6	60.4	60.6
	Stance Reasoner*	69.7	62.5	73.9	67.7	60.3	66.8	79.5	81.0	79.6	80.0	52.5	46.1	51.5	55.7	48.3	50.8
	COLA	70.8	65.5	63.4	81.7	71.0	70.5	86.6	84.0	79.7	83.4	55.8*	44.6*	59.6*	52.5*	59.2*	54.3*
	ToC	–	–	–	–	–	69.4	75.7*	83.1*	80.4*	79.7*	47.8*	48.3*	69.7*	54.4*	57.5*	55.7*
MSME (ours)	GPT-3.5*	75.2	74.9	72.5	81.1	69.9	74.7	87.7	84.9	82.8	85.1	62.6	67.1	71.4	75.3	66.3	68.5
	GPT-4o*	80.3	76.2	75.5	**81.9**	71.9	77.2	**88.6**	**85.6**	84.1	**86.1**	68.3	71.8	72.6	**76.3**	70.7	72.0
	DeepSeek-r1*	**81.5**	**78.5**	74.8	80.6	**73.5**	**77.8**	87.1	84.7	**84.5**	85.4	**69.8**	**75.9**	72.2	75.5	**74.0**	**73.5**
	QwQ-32b*	79.5	77.1	**76.3**	76.9	68.4	75.6	85.1	84.3	83.5	84.3	66.1	70.5	**73.2**	72.8	69.6	70.4

Table 1: Comparison of MSME with baselines across three datasets. * indicates results from our own experiments. Results without * are taken from the original papers. Bold and underline refer to the best and 2nd-best performance. All results are statistically significant with paired t-tests, $p < 0.05$.

(Zhang et al. 2024a), logical chain-based stance reasoning (Stance Reasoner) (Taranukhin, Shwartz, and Milios 2024), collaborative frameworks with multiple agents (COLA) (Lan et al. 2024), and counterfactual tree prompts to guide reasoning (ToC) (Weinzierl and Harabagiu 2024).

4.2 Main Result

Table 1 presents a comparison of MSME with various baselines across three datasets. We report the experimental results for each target individually. These results demonstrate the strong performance of our MSME:

MSME achieves significant improvements over state-of-the-art methods across all three datasets. On SEM16 and P-Stance, it achieves F1 scores of 74.7 and 85.1—improvements of 4.2 and 1.7 points over the best baseline (COLA). On Weibo-SD, MSME raises F1 from 60.6 to 68.5 (+7.9) compared to BKEL. These results confirm the effectiveness of our approach. Unlike COLA's generic multi-agent framework, MSME's three-stage design directly addresses dependencies on background knowledge, label ambiguity, and pragmatic cues. Compared to BKEL, MSME refines background information through the knowledge expert, extracting more relevant facts to enhance reasoning, while the label expert further clarifies and refines labels, mitigating the ambiguity in target–label mapping.

The greatest improvement appears on Weibo-SD, with the smallest gain on P-Stance. We attribute this to three key factors: First, Weibo-SD targets are complex events with multiple sub-events and rich rhetorical devices, demanding advanced reasoning, whereas P-Stance involves single political figures, requiring no extra reasoning to align targets and labels. Second, Weibo-SD covers recent, emerging events require injected background knowledge, while P-Stance's older data is likely encoded in the LLMs' internal knowledge. Third, COLA's strong performance on P-Stance (83.4 F1) leaves less room for gain. Superior gains on the most challenging dataset further validate MSME's robustness.

MSME demonstrates superior capability on compound targets. On SEM16's only compound target, *Climate Change Is Real Concern* (CC), the best baseline (COLA) achieves an F1 of 65.5, while MSME attains 74.9, a 9.4-point improvement. This further highlights that by introducing the label expert, MSME effectively addresses the challenge of ambiguous target–label mapping.

MSME also performs robustly across different LLMs. On SEM16 and Weibo-SD, DeepSeek-R1 yields the best results, with F1 scores of 77.8 and 73.5, representing improvements of 3.1 and 5.0 compared to GPT-3.5. On P-Stance, GPT-4o achieves the highest F1 of 86.1, which is 1.0 points higher than GPT-3.5. QWQ-32B, while not the best performer, still improves by 0.9 on SEM16 and 1.9 on Weibo-SD relative to GPT-3.5. This demonstrates the strong performance of inference models, with QWQ-32B—despite having only 32B parameters—still delivering excellent results.

4.3 Ablation Study

In the ablation study, we tested three settings: removing one expert, retaining only one expert, and using no experts (i.e., relying solely on the ESL and the K_{raw} obtained in the knowledge preparation stage and for stance reasoning). Table 2 presents our results on three LLMs:

The ablation study demonstrates the necessity of each expert. When all experts are retained, the model achieves the best performance across all three datasets. Removing any expert results in a decline in performance. For example, with GPT-3.5, the performance drops by 1.2 to 2.9 points on

Model	KE	LE	PE	SEM16	P-Stance	Weibo-SD
GPT-3.5 Turbo	✓	✓	✓	**74.7**	**85.1**	**68.5**
	×	✓	✓	71.8	81.3	65.1
	✓	×	✓	73.1	**83.5**	64.0
	✓	✓	×	**73.5**	83.3	**66.1**
	✓	×	×	**72.0**	**82.8**	62.8
	×	✓	×	71.1	81.1	**64.5**
	×	×	✓	69.9	80.7	61.6
	×	×	×	68.0	79.2	59.9
GPT-4o	✓	✓	✓	**77.2**	**86.1**	**72.0**
	×	✓	✓	74.9	82.9	70.1
	✓	×	✓	75.1	84.3	69.6
	✓	✓	×	**76.1**	**84.6**	**70.4**
	✓	×	×	**75.0**	**83.7**	**68.5**
	×	✓	×	74.4	81.5	68.2
	×	×	✓	73.5	80.8	66.7
	×	×	×	70.7	81.2	64.9
DeepSeek -R1	✓	✓	✓	**77.8**	**85.4**	**73.5**
	×	✓	✓	75.4	83.2	71.4
	✓	×	✓	76.5	**84.1**	70.7
	✓	✓	×	**77.1**	83.9	**71.7**
	✓	×	×	**76.1**	**83.5**	69.3
	×	✓	×	75.5	81.7	**69.5**
	×	×	✓	74.6	82.0	68.0
	×	×	×	72.8	81.6	66.1

Table 2: Ablation study results with GPT-3.5 Turbo, GPT-4o, and DeepSeek-R1. KE denotes the Knowledge Expert, LE the Label Expert, and PE the Pragmatic Expert.

SEM16, 1.8 to 3.8 points on P-Stance, and 2.4 to 4.5 points on Weibo-SD, showing a significant decline.

The knowledge expert is essential for the other expert modules. For example, with GPT-3.5, removing the knowledge expert causes the largest performance decline on SEM16 and P-Stance, with F1 dropping by 2.9 and 3.8, respectively, and by 3.4 on Weibo-SD. DeepSeek-R1 shows identical results to GPT-3.5, while GPT-4o exhibits the largest drop across all datasets, further confirming this conclusion. In the expert reasoning stage, the knowledge expert refines raw knowledge and shares it with the other experts, highlighting its critical role in knowledge refinement.

Knowledge-based reasoning yields the greatest improvements in general scenarios. On SEM16 and P-Stance, all models perform best when only the knowledge expert is retained. For instance, GPT-3.5 achieves F1 scores of 72.0 and 82.8 on SEM16 and P-Stance, respectively, representing the largest gains. This is likely due to the simpler target-label mapping in these datasets—SEM16 contains only one complex target (CC), while P-Stance involves single-entity targets. Reasoning solely from relevant knowledge leads to the greatest improvements on these datasets.

The label expert is key to handling complex targets in Chinese scenarios. Removing it has the greatest impact on Weibo-SD, with F1 decreasing by 4.5, 2.4, and 2.8 across the three models. In the setup where only one expert is retained, the best results are achieved when the label expert alone is kept. For instance, GPT-3.5 and DeepSeek-R1 reach F1 scores of 64.5 and 69.5, respectively, improving by 4.6 and 3.4 compared to no experts. This is attributed to the fine-grained stance label system, which simplifies the mapping between targets and labels.

Model	SEM16	P-Stance	Weibo-SD
Noise Injection			
No KE + Noise	70.1	79.8	60.7
No KE	71.8	81.3	65.1
MSME + Noise	74.3	84.5	66.9
MSME	**74.7**	**85.1**	**68.5**
Neutral Detection			
Base	46.7	–	38.1
No Expert	52.3	–	45.2
Label Expert	59.7	–	52.9
MSME	**60.5**	–	**54.6**
Rhetoric Handling			
Base	55.2	68.1	43.3
No Expert	63.2	75.5	50.6
Pragmatic Expert	67.5	80.1	58.7
MSME	**71.4**	**82.4**	**64.0**
Decision Capability			
Integrated	71.8	79.8	63.4
Vote	74.1	83.9	66.8
MSME	**74.7**	**85.1**	**68.5**

Table 3: Supplementary experiments analyzing MSME's performance in noise injection, neutral detection, rhetoric handling and decision capability.

The impact of the pragmatic expert is relatively small. For all three models, removing it results in the smallest performance decline across the three datasets. For example, with GPT-3.5, F1 drops by 1.2, 1.8, and 2.4 on SEM16, P-Stance, and Weibo-SD, respectively. When only the pragmatic expert is retained, the performance improvement is also minimal, with F1 increasing by 1.9, 1.5, and 1.7 compared to using no experts. This is likely because not every text contains rich rhetorical devices, making the it's role more supplementary. This will be further explained in the supplementary experiments.

4.4 Supplementary Experiments

To further investigate the role of the three experts and the decision aggregation stage, we conducted four supplementary experiments with the following setups:

Experiment 1: We added noise to K_{raw} by injecting target-irrelevant knowledge of the same scale. We compared the performance of MSME without the knowledge expert but with noise injection (No KE + Noise), MSME without the knowledge expert (No KE), MSME with noise injection (MSME + Noise) , and MSME.

Experiment 2: We analyzed the F1 scores for the neutral label, comparing the Base, results without any experts (No Expert), the results when only the label expert was used (Label Expert), and MSME.

Experiment 3: We used three inference models (GPT-4o, QWQ-32B, and DeepSeek-R1) to detect rhetoric in texts with intuitive prompts, employing majority voting to select texts containing rhetoric (details in Appendix). Re-

sults showed the proportion of texts containing rhetoric was 52.2% in SEM16, 69.6% in P-Stance, and 80.6% in Weibo-SD. We then compared results on texts containing rhetoric using Base, No Expert, Pragmatic Expert Only, and MSME.

Experiment 4: We compared integrating the expert reasoning and decision aggregation stages into a single process with one prompt (Integrated), summarizing each expert's independent judgment and applying majority voting for the final stance (Vote), and MSME.

All results are shown in Table 3. We find that:

MSME effectively refines background knowledge. In the No KE + Noise setup, performance is poor, with F1 dropping significantly across all three datasets—by 4.6, 5.3, and 7.8 points compared to MSME. In the MSME + Noise setup, performance drops only by 0.4, 0.6, and 1. This highlights the importance of the knowledge expert in mitigating noise in the knowledge.

MSME excels in handling neutral text. Existing methods struggle with neutral stance classification, with the Base setup achieving F1 scores of only 46.7 and 38.1 on SEM16 and Weibo-SD. Compared to the No Expert setup, the label expert improves F1 by 7.4 and 7.7, respectively. This is attributed to the fine-grained stance label system, which simplifies target-label mapping, enabling better detection for neutral stance when finer labels are not applicable.

MSME effectively handles rhetorical texts. In the Pragmatic Expert setup, accuracy in detecting rhetorical cues significantly improves, with F1 increasing by 4.3, 4.6, and 8.1 points compared to the No Expert setup. This is due to the specialized analysis of rhetorical devices in the text.

MSME enables effective integration of multi-dimensional reasoning. Compared to Integrated and Vote setups, MSME achieves the best performance. The Integrated setup shows a significant drop in F1, with decreases of 2.9, 5.3, and 5.1, as end-to-end methods struggle to balance multiple factors. The Vote setup, while closer, still shows a drop in F1 of 0.6, 1.2, and 1.7, demonstrating that MSME goes beyond simple voting, effectively integrating knowledge from different experts.

4.5 Case Study

To clearly demonstrate the workings of MSME, Figure 6 presents a complete example. For the target *Climate Change is a Real Concern* (CC), the three experts provide their individual reasoning:

Knowledge Expert refines the background knowledge and, from a knowledge perspective, identifies the text's downplaying of human impact on climate change. This leads to the stance of opposing human-induced climate change.

Label Expert applies fine-grained stance labels, reasoning from the label perspective, and categorizes the stance as Against, specifically: '*Argues that human activity has a negligible impact compared to natural phenomena.*'

Pragmatic Expert detects the use of exaggeration and irony in the text, reasoning from the pragmatic perspective, revealing the actual intent: human impact on climate change is minimal.

Finally, the meta-judge integrates these analyses, confirming that the stance expressed is Against climate change, implying that it is not a real concern and does not warrant significant attention or action.

Figure 6: Cases of explainations generated by MSME.

5 Conclusion

In this work, we propose MSME, a multi-stage, multi-expert framework for zero-shot stance detection, addressing challenges in real-world scenarios including background knowledge dependency, unclear target-label mapping, and pragmatic complexity. MSME consists of three stages: in the *Knowledge Preparation* stage, necessary background knowledge is retrieved and stance labels are clarified; in the *Expert Reasoning* stage, three experts reason from knowledge, label, and pragmatic perspectives; and in the *Decision Aggregation* stage, the final stance is determined by integrating all insights. Through extensive experiments on three datasets and four LLMs, we demonstrate that MSME significantly outperforms existing state-of-the-art methods. Ablation and supplementary experiments emphasize the key role of each expert, confirming MSME's robustness in handling noisy knowledge, neutral stances, and rhetorical devices, as well as its excellent decision capability.

Acknowledgments

We sincerely thank all reviewers for their valuable feedback and constructive comments. This work is supported by the National Social Science Fund of China (Grant No. 22&ZD035).

References

Allaway, E.; and McKeown, K. 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)*, 8913–8931.

Allaway, E.; and McKeown, K. 2020b. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Allaway, E.; Srikanth, M.; and McKeown, K. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Qian, C.; Chan, C.-M.; Qin, Y.; Lu, Y.; Xie, R.; et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4): 6.

Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. *arXiv preprint arXiv:2005.00388*.

Dey, K.; Shrivastava, R.; and Kaushik, S. 2018. Topical stance detection for Twitter: A two-phase LSTM model using attention. In *European Conference on Information Retrieval*, 529–536. Springer.

Ding, N.; Hu, S.; Zhao, W.; Chen, Y.; Liu, Z.; Zheng, H.-T.; and Sun, M. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. Cross-Domain Label-Adaptive Stance Detection. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9011–9028. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

He, Z.; Mokhberian, N.; and Lerman, K. 2022. Infusing knowledge from wikipedia to enhance stance detection. *arXiv preprint arXiv:2204.03839*.

Hosseinia, M.; Dragut, E.; and Mukherjee, A. 2020. Stance Prediction for Contemporary Issues: Data and Experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, 32–40.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Igarashi, Y.; Komatsu, H.; Kobayashi, S.; Okazaki, N.; and Inui, K. 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)*, 401–407.

Jeong, C.; Jang, S.; Park, E.; and Choi, S. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics*, 124: 1907–1922.

Koroteev, M. V. 2021. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Lan, X.; Gao, C.; Jin, D.; and Li, Y. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, 891–903.

Li, A.; Liang, B.; Zhao, J.; Zhang, B.; Yang, M.; and Xu, R. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 15703–15717.

Li, Y.; Sosea, T.; Sawant, A.; Nair, A. J.; Inkpen, D.; and Caragea, C. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2355–2365.

Liang, B.; Chen, Z.; Gui, L.; He, Y.; Yang, M.; and Xu, R. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM web conference 2022*, 2738–2747.

Liang, B.; Fu, Y.; Gui, L.; Yang, M.; Du, J.; He, Y.; and Xu, R. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the web conference 2021*, 3453–3464.

Liang, B.; Zhu, Q.; Li, X.; Yang, M.; Gui, L.; He, Y.; and Xu, R. 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, 81–91. Association for Computational Linguistics.

Liu, R.; Lin, Z.; Tan, Y.; and Wang, W. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 3152–3157.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In Bethard, S.; Carpuat, M.; Cer, D.; Jurgens, D.; Nakov, P.; and Zesch, T., eds., *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.

Pick, R. K.; Kozhukhov, V.; Vilenchik, D.; and Tsur, O. 2022. Stem: unsupervised structural embedding for stance

detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11174–11182.

Taranukhin, M.; Shwartz, V.; and Milios, E. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. *arXiv preprint arXiv:2403.14895*.

Taulé, M.; Pardo, F. M. R.; Martí, M. A.; and Rosso, P. 2018. Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In *IberEval@ SEPLN*, 149–166. Sevilla.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Weinzierl, M.; and Harabagiu, S. 2024. Tree-of-Counterfactual Prompting for Zero-Shot Stance Detection. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–880. Bangkok, Thailand: Association for Computational Linguistics.

Xiao, S.; Liu, Z.; Zhang, P.; and Muennighoff, N. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597.

Xu, C.; Paris, C.; Nepal, S.; and Sparks, R. 2018. Cross-target stance classification with self-attention networks. *arXiv preprint arXiv:1805.06593*.

Xu, J.; Zheng, S.; Shi, J.; Yao, Y.; and Xu, B. 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*, 679–688. Springer.

Xu, R.; Zhou, Y.; Wu, D.; Gui, L.; Du, J.; and Xue, Y. 2016b. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *Proceedings of the ICCPOL 2016*, 907–916.

Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Zhang, B.; Fu, X.; Ding, D.; Huang, H.; Dai, G.; Yin, N.; Li, Y.; and Jing, L. 2023a. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Zhang, C.; Li, Q.; and Song, D. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Zhang, J.; Wu, S.; Zhang, X.; and Feng, Z. 2023b. Task-specific data augmentation for zero-shot and few-shot stance detection. In *Companion proceedings of the ACM web conference 2023*, 160–163.

Zhang, L., Yuanshuoand Aohua; Zhaoning, Y.; Panyi, W.; Bo, C.; and Xiaobing, Z. 2024a. Research on Stance Detection with Generative Language Model. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, 481–491.

Zhang, Z.; and Lan, M. 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)*, 451–457.

Zhang, Z.; Li, Y.; Zhang, J.; and Xu, H. 2024b. 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)*, 371–378.

Zheng, C.; Zhang, Z.; Zhang, B.; Lin, R.; Lu, K.; Yu, B.; Liu, D.; Zhou, J.; and Lin, J. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.