Zero-Shot Stance Detection in the Wild: Dynamic Target Generation and Multi-Target Adaptation

Anonymous ACL submission

Abstract

Current stance detection research typically relies on predicting stance based on given targets and text. However, in real-world social media scenarios, targets are neither predefined nor static but rather complex and dynamic. To address this challenge, we propose a novel task: zero-shot stance detection in the wild with dynamic target generation and multi-target adaptation, which aims to automatically identify multiple target-stance pairs from text without prior target knowledge. We construct a Chinese social media stance detection dataset and design multi-dimensional evaluation metrics. We explore both integrated and two-stage fine-tuning strategies for large language models (LLMs) and evaluate various baseline models. Experimental results demonstrate that fine-tuned LLMs achieve superior performance on this task: the integrated fine-tuned Qwen2.5-7B attains the highest comprehensive target recognition score of 66.99%, while the two-stage fine-tuned DeepSeek-R1-Distill-Qwen-7B achieves a stance detection F1 score of 79.26%. The dataset and models are publicly available at: `https://anonymous.4open.science/r/DGTA-stance-detection-7299`.

Text:
With 13,120 units sold, xiaomi SU7 has surpassed the Tesla Model 3. Featuring exceptional exterior design and generous performance specifications, the SU7 has gained tremendous popularity both domestically and internationally - truly a remarkable achievement for Xiaomi!

DGTA-Output:
(Target: xiaomi SU7 , Stance: Support)
(Target: Tesla Model 3, Stance: Neutral)

Figure 1: Real-world example from the Chinese platform Weibo.The task involves automatically identifying two distinct targets and inferring corresponding stance labels by modeling the semantic relationship between the text and each target.

1 Introduction

Stance detection aims to identify an author's attitudinal tendency towards a specific target (AlDayel and Magdy, 2021; Mohammad et al., 2016), including support, against, or neutral (Li and Caragea, 2019; Küçük and Can, 2020). Most existing research has focused on known targets and achieved significant progress (Siddiqua et al., 2019; AlDayel and Magdy, 2021).

However, in open social media environments, due to topic diversity and the relatedness of discussion objects (Alturayeif et al., 2022), phenomena of unclear targets and multiple coexisting targets frequently emerge, resulting in single texts potentially containing multiple stance targets where stance labels are often associated with complex relationships between targets. Figure 1 provides a real examples from the Chinese platform Weibo. Although there have been studies on target adaptation, such as an unsupervised stance detection framework combining expert mixing, domain adversarial training, and target label embeddings to achieve cross-domain prediction for unseen targets (Hardalov et al., 2022), and the Target-Stance Extraction (TSE) task which only addresses single targets by jointly modeling target identification and stance detection (Li et al., 2023), these approaches rely on target candidate labels or only support single target identification, making them difficult to adapt to multi-target and unknown target real-world application scenarios (Putra et al., 2022; Sobhani et al., 2017).

To address these challenges, we propose a more open-ended task: Zero-Shot Stance Detection in the Wild with Dynamic Target Generation and Multi-Target Adaptation (DGTA), which aims to adaptively identify diverse targets and determine stances from input text without relying on predefined targets, thereby more effectively accommodating complex and dynamic real-world application scenarios. To support research on this task, we construct the first high-quality Chinese stance detection dataset covering multi-domain social media posts, comprising 70,931 annotated samples. We design multi-dimensional evaluation metrics for target identification and stance determination, where target identification assessment includes BERTScore (Zhang et al.), BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), Recall and a comprehensive score, while stance determination only evaluates samples whose target identification metrics reach a threshold. At the methodological level, we propose two strategies for fine-tuning large language models (LLMs): an integrated approach generating multiple target-stance label pairs and a two-stage method separately generating multiple targets and stance labels. We also implement various baseline models, including fine-tuned pre-trained models and differently prompted LLMs. Experimental results demonstrate that in the DGTA task, fine-tuned LLMs significantly outperform pre-trained and prompted models, with integrated and two-stage fine-tuning strategies each showing distinct advantages.

Our main contributions are summarized as follows:

- We propose a new task of Zero-Shot Stance Detection in the Wild with Dynamic Target Generation and Multi-Target Adaptation (DGTA), construct the first high-quality Chinese multi-domain social media stance detection dataset, and design unified and comprehensive evaluation metrics.
- We explore two strategies for fine-tuning LLMs, based on integrated and two-stage frameworks, providing a powerful baseline.
- We conduct baseline experiments including fine-tuned pre-trained models and various prompted LLMs, with detailed comparative analysis.

2 Related Work

2.1 Traditional Stance Detection

Traditional stance detection methods have evolved from manual feature engineering to contextualized pre-trained models (Glandt et al., 2021). Zarrella and Marsh (2016) integrates grammatical and syntactic information into RNNs and learns vector representations of input text, effectively enhancing stance detection performance on Twitter texts. Du et al. (2017) introduces attention mechanisms into LSTM, proposing a target-specific enhanced attention model. WS-BERT substantially improves performance in target-specific (He et al., 2022), cross-target, and zero/few-shot scenarios by integrating Wikipedia knowledge to enrich target representations. The GDA-CL model generates high-quality synthetic samples in embedding space through generative adversarial networks (GAN) and hybrid contrastive learning (Li and Yuan, 2022), using GPT-2 as the generator, RoBERTa as the discriminator, and BERT as the classifier within the GAN framework (Goodfellow et al., 2014), supplemented with a multilayer perceptron for contrastive learning, significantly improving zero-shot stance detection performance on unseen targets. Stance Reasoner aims to leverage explicit reasoning about background knowledge to guide models in inferring target stances (Taranukhin et al., 2024). LKI-BART introduces LLM knowledge to establish connections between text and unseen targets, achieving optimal performance on VAST and P-Stance datasets (Zhang et al., 2024; Allaway and McKeown, 2020; Li et al., 2021). However, these studies all rely on predefined targets, cannot adapt to real-world scenarios where targets are implicit or unknown, and fail to address the problem of dynamic target generation.

2.2 Target-Adaptive Stance Detection

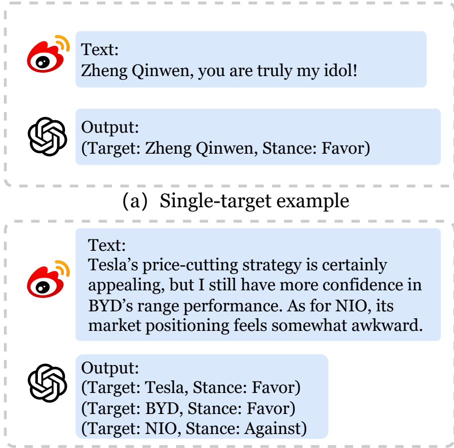
Recent studies have begun to focus on target-adaptive stance detection. TATA leverages contrastive learning to extract topic-agnostic and topic-aware embeddings from unlabeled news texts and applies them to downstream stance detection tasks (Hanley and Durumeric, 2023). TAPD enhances cross-target few-shot stance detection through target-aware prompt adaptation and multi-prompt distillation techniques (Wang and Pan, 2024; Wen and Hauptmann, 2023), mapping stance labels to continuous vectors. For cross-target domain adaptation, the target-aware domain adaptation method extracts key shared features through feature disentanglement and automatically identifies target relationships (Deng et al., 2022). Stanceformer introduces target-aware attention mechanisms (Garg and Caragea, 2024). OpenStance defines the open-domain zero-shot stance detection task (Xu et al., 2022), addressing stance detection without domain restrictions or specific topic focus. Wu et al. (2022) proposes a novel multi-source adaptive target detection method for Target-Related Knowledge Preservation. For the new task of cross-lingual cross-target stance detection, a dual-teacher knowledge distillation framework CCSD is designed (Zhang et al., 2023), utilizing cross-lingual and cross-target teachers to guide student model learning from source languages. Although these works have significantly advanced target-adaptive stance detection, they still rely on predefined target lists or domain-specific data and only address single-target adaptation, limiting model adaptability for stance detection in real scenarios with undefined, multiple targets requiring dynamic generation and adaptation.

3 Dynamic Target Generation and Multi-Target Adaptation for Stance Detection

To address the challenges of both dynamic targets and complex stances in real-world scenarios, we propose a new task of stance detection with dynamic target generation and multi-target adaptation. This task requires models to automatically identify (multiple) stance targets in text without predefined targets or domains, determine corresponding stances, and ultimately output pairs of targets and stances. We successively introduce the task definition, dataset construction and analysis, and LLM fine-tuning strategies.

3.1 Task Definition

The task of dynamic target generation and multi-target adaptation for stance detection (DGTA) is defined as follows: given input text posted by social media users, without any predefined targets, topics, or domains, the model is required to output all pairs of targets and their corresponding stances present in the text. Targets include both static entities (e.g., persons, organizations, institutions) and dynamic entities (e.g., actions, events, states), and their number may vary from single to multiple. The stance labels are categorized into three classes: support, against, and neutral. Figure 2 (a) illustrates a sample post where the model identifies a single target along with its associated stance, resulting in one target-stance pair as output; Figure 2 (b) presents a more complex scenario involving three distinct targets, each paired with a corresponding stance, yielding three target-stance pairs.



(a) Single-target example

(b) Multi-target example

Figure 2: Examples of the new stance detection task

3.2 Chinese Social Media Personas Stance Dataset

3.2.1 Data Collection and Preprocessing

We select 240 users from diverse domains on the Weibo platform and collect their posts within the same time period. The reasons are as follows: (1) As a representative Chinese social media platform, Weibo features diverse users and topics, offering broad representativeness and applicability to various stance detection scenarios; (2) Selecting 240 users across entertainment, finance, law, education, and other domains helps evaluate the method's generalization capability in handling target and stance analysis in complex contexts. From the posts published by these 240 users, we collect a total of 125,176 textual entries. Due to the informal nature of user-generated content, we apply regular expressions and Unicode encoding techniques to remove non-standard text elements such as emojis, URLs, usernames, and special symbols, which often introduce noise and reduce stance classification accuracy. During this process, all collected posts undergo strict anonymization, with user IDs anonymized. No user identity information is used in any of the experiments reported here. This anonymized ID information supports future research focusing on user-centric stance detection. Finally, after preprocessing, 107,310 posts are retained for subsequent annotation.

3.2.2 Data Annotation and Validation

To ensure the standardization and reliability of dataset annotation, we construct an annotation workflow based on the combination of collaborative annotation by multiple LLMs, score-based cor-

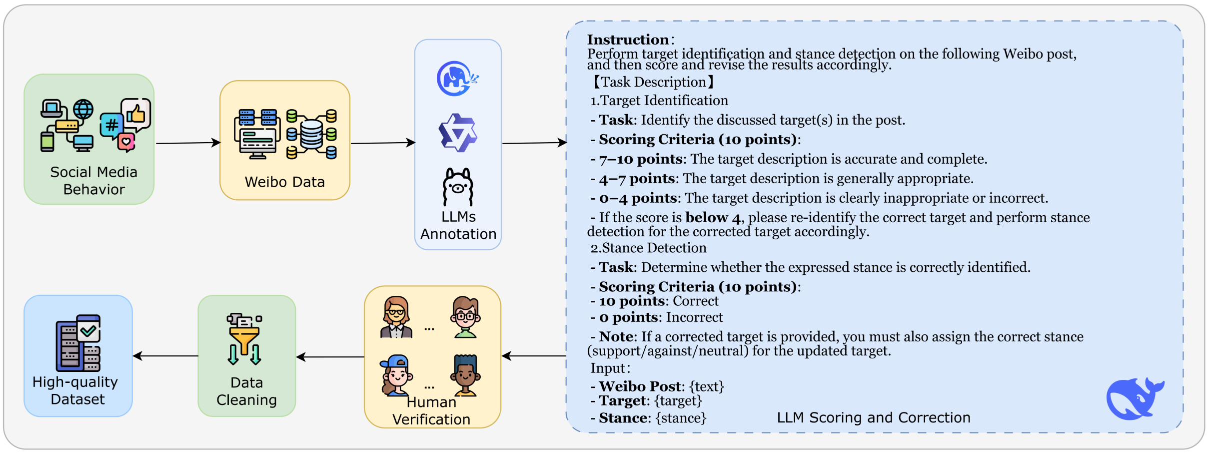


Figure 3: Workflow of dataset construction with collaboration between multiple LLMs and human verification

rection, and human verification, with the complete process illustrated in Figure 3.

Specifically, we select three mainstream LLMs (GLM4-9B, Qwen2.5-7B, and Llama3-8B) to independently perform the cascaded tasks of target identification and stance determination. For the annotation results from these three models, we establish a cross-validation mechanism: in the two-stage target-stance annotation, if at least two models produce identical target entity recognition results for the same text and reach consensus on stance judgment for that target, the sample is adopted as a valid annotation; if substantial disagreement exists at either stage, the sample is considered invalid and removed from the dataset. After completing the first round of cross-validation, we utilize prompt instructions to guide the DeepSeek-V3 model in conducting a secondary scoring evaluation of valid annotated samples, with low-scoring samples being modified and marked for review. Subsequently, eight professional annotators verify all automatically annotated samples. During the data cleaning phase, we eliminate low-quality texts containing logical contradictions, semantic ambiguities, or lacking clear target references. This process ultimately results in a high-quality annotated dataset with strict cross-validation constraints.

3.2.3 Dataset Statistics and Analysis

After the above processing steps, the final dataset comprises 70,931 textual entries, covering both single-target and multi-target scenarios. The detailed statistics of target quantity and stance distribution are provided in Appendix A Table 5. Table 6 in Appendix A shows the quantitative ranking of the top 10 most frequently discussed targets.

3.3 Evaluation Criteria

Due to the diversity and uncertainty in both expression and quantity of dynamically generated targets in this task, traditional evaluation metrics fail to adequately reflect model performance. To comprehensively assess model performance in this task, we design more targeted and comprehensive evaluation criteria.

3.3.1 Target Identification Evaluation Criteria

For the open-ended characteristics of the target identification phase, we propose a multi-dimensional evaluation approach that integrates semantic similarity, surface form matching, and quantity alignment to construct a comprehensive target identification score (C-Score). This metric comprises BERTScore, BLEU, ROUGE-L, and the Recall of target quantity.

$$\begin{aligned} C\text{-Score} = (&\alpha \times \text{BERTScore} \\ &+ \beta \times \text{BLEU} + \gamma \times \text{ROUGE-L}) \times \text{Recall} \end{aligned} \quad (1)$$

Where α, β, and γ control the weighting proportions of the three metrics. Considering the semantic and structural differences between predicted targets and reference targets, we first align the metrics. Experiments show that setting $\alpha = 0.6$, $\beta = 0.2$, and $\gamma = 0.2$ emphasizes semantic consistency while balancing lexical and structural matching.

3.3.2 Stance Detection Evaluation Criteria

Considering that stance classification evaluation is only meaningful when based on accurate target identification, we first establish a threshold-driven mechanism for determining target correctness: through experimental analysis, we set thresholds of 0.7, 0.2, 0.4, 0.8, and 0.3 for BERTScore,

Instruction:
You are an expert in target identification and stance detection. Based on the given Weibo comment, identify the main targets being discussed (such as people, events, or actions), and determine the stance toward each target (Support / Against / Neutral). If the comment contains multiple targets, evaluate the stance for each one separately.
Output format:(Target: [Target1]; [Target2], Stance: [Stance1]; [Stance2])

Input:
That shot by Fan Zhendong was truly amazing—his backhand is so powerful! The team members all stood up to applaud and cheer for him!

Output:
(Target: Fan Zhendong, Stance: Support)

Figure 4: Prompt template and example for the integrated fine-tuning strategy

BLEU, ROUGE-L, Recall, and the comprehensive score, respectively. Samples exceeding these thresholds are deemed to have correct target identification. On this foundation, we employ the classic metrics of Precision, Recall, and F1 score.

3.4 Fine-tuning Large Language Models

We fine-tune LLMs to provide powerful baselines for this task. We propose two fine-tuning strategies—integrated and two-stage, and use each strategy to construct instruction fine-tuning data. All fine-tuning is conducted using LoRA (Hu et al., 2022). The integrated fine-tuning strategy adopts an end-to-end approach, modeling the "target identification + stance detection" task as an instruction-driven sequence generation process. Model input consists of task instructions in natural language concatenated with the original text (as shown in Figure 4), explicitly prompting the task intent, guiding the model to simultaneously complete target extraction and stance classification, ultimately outputting (multiple) target-stance pairs, achieving task coordination.

The two-stage fine-tuning strategy decouples target identification and stance determination into two independent subtasks, each undergoing separate instruction fine-tuning. In the first stage, the model receives input text with task instructions (as shown in Appendix B Figure 5), focusing exclusively on extracting potential targets from the text. In the second stage, the identified targets and original text serve as input (as shown in Appendix B Figure 5), accompanied by stance determination instructions, guiding the model to classify stance for specific targets. Through independent fine-tuning, models can focus on a single task. It is worth noting that we use different models for our two-stage fine-tuning approach, rather than the same model.

4 Experiments and Analysis

4.1 Experimental Setup

4.1.1 Dataset

We divide our constructed dataset into training, validation, and test sets in an 8:1:1 ratio for fine-tuning and baseline evaluation experiments. Considering the high computational resources required for evaluating the full dataset, we adopt a random sampling strategy, extracting 1,000 samples from the test set as a subsequent model testing subset.

4.1.2 Comparison Models

We compare three categories of models: fine-tuned pre-trained models, instruction-prompted LLMs, and instruction-tuned LLMs. For the fine-tuned pre-trained models, we employ fine-tuned mT5 for target identification and fine-tuned BERT for stance detection (Xue et al., 2021; Devlin et al., 2019). For the instruction-prompted LLMs, we experiment with current mainstream models including DeepSeek-V3 (Liu et al., 2024), GLM4-9B (GLM et al., 2024), GPT-4o (Hurst et al., 2024), and Llama3-8B (Grattafiori et al., 2024) using instruction prompting. For the instruction-tuned LLMs, we fine-tune Qwen2.5-7B-Instruct and DeepSeek-R1-Distill-Qwen-7B using an integrated approach (Qwen et al., 2024; Guo et al., 2025), and also independently fine-tune Qwen2.5-7B-Instruct for target identification and DeepSeek-R1-Distill-Qwen-7B for stance detection in a two-stage process.

4.2 Experimental Results and Analysis

The experimental results are shown in Table 1, and we can find that:

- Fine-tuned LLMs significantly outperform both fine-tuned pre-trained models and instruction-prompted LLMs on the DGTA task. Both integrated and two-stage fine-tuned LLMs achieve comprehensive scores exceeding 66% in target identification, with Qwen2.5-7B demonstrating optimal performance (66.99%). In target identification tasks, both fine-tuned and pre-trained models exhibit BERTScore metrics above 84%, indicating that fine-tuning enhances target semantic comprehension capabilities. For stance detection, fine-tuned DeepSeek-R1 models achieve F1 scores (79.26% and 75.37%) that surpass Llama3-8B by over 20 percentage

Model	Target Identification					Stance Detection		
	BERT	BLEU	ROUGE	Recall	C-Score	P	R	F1
mT5‡	84.29	28.82	65.71	86.59	60.16	-	-	-
Bert§	-	-	-	-	-	67.89	67.11	67.51
Qwen2.5-7B	82.47	28.26	63.69	91.16	61.87	64.22	67.54	65.05
DeepSeek-V3	76.87	25.65	50.38	92.05	56.45	69.25	72.60	70.64
GLM4-9B	77.98	24.16	51.99	94.14	58.38	68.50	69.20	66.90
GPT-4o	73.72	21.51	43.99	94.34	54.09	74.22	74.75	74.45
Llama3-8B	77.69	27.94	56.98	85.90	54.63	58.45	65.03	59.52
Qwen2.5-7B†	**85.09**	31.12	**67.14**	94.16	66.58	65.31	65.33	64.16
DeepSeek-R1-Qwen†	84.94	30.96	66.99	94.62	66.76	**87.46**	**77.08**	**79.26**
Qwen2.5-7B‡	84.69	**31.64**	66.17	**95.19**	**66.99**	-	-	-
DeepSeek-R1-Qwen§	-	-	-	-	-	83.25	74.52	75.37

Table 1: Overall experimental results on the DGTA task (Unit: %, best results are in bold. † indicates integrated fine-tuning; ‡ indicates the target identification stage in two-stage fine-tuning; § indicates the stance determination stage in two-stage fine-tuning. DeepSeek-R1-Distill-Qwen-7B is abbreviated as DeepSeek-R1-Qwen. BERTScore is abbreviated as BERT, and ROUGE-L is abbreviated as ROUGE.)

points, demonstrating that fine-tuning substantially improves stance reasoning abilities in complex semantic contexts.

- Integrated and two-stage strategies each have advantages in target identification and stance detection subtasks. For target identification, phased fine-tuning enables greater focus, with Qwen2.5-7B achieving the optimal score of 66.99%. For stance detection, integrated fine-tuning exhibits superior performance, with DeepSeek-R1-Distill-Qwen-7B outperforming two-stage models across all evaluation metrics, likely due to its ability to simultaneously model inter-target relationships and stance associations.

- Models with reasoning capabilities generally perform better than those without. Between the two integrated fine-tuned models—Qwen2.5-7B and DeepSeek-R1-Distill-Qwen-7B—the latter underwent reasoning distillation. Comparison reveals that the reasoning-capable DeepSeek-R1 model achieves a comprehensive score of 66.76% in target identification and an F1 score of 79.26% in stance detection, outperforming the Qwen2.5-7B model overall. This indicates that reasoning capabilities contribute to more precise target identification and stance determination in complex scenarios.

4.3 Dynamic Target Difference Analysis

4.3.1 Target-Oriented Difference Analysis

Target	BERT	BLEU	ROUGE	Recall	C-Score
Single	82.04	28.79	61.55	**99.36**	**67.28**
Dual	**82.47**	**29.47**	**62.72**	93.59	63.76
Triple	80.52	27.89	58.01	80.43	53.02
Multi	79.29	26.80	55.34	67.67	45.42

Table 2: Overall experimental results categorized by the number of targets (Unit: %. Values are the average results across all models)

We conduct a comprehensive comparison of all models based on target quantity(Table 2), categorizing samples into single-target, dual-target, triple-target, and multi-target (more than three targets).

Dual-target samples perform optimally on semantic evaluation metrics. These samples achieve the highest scores across three semantic-related metrics: BERTScore (82.47%), BLEU (29.47%), and Rouge-L (62.72%). This superior performance can be attributed to two primary factors: First, dual-target texts typically involve comparative, parallel, or opposing relationships, which strengthen target boundaries and semantic contrasts, making targets more distinguishable for the model. Second, compared to single-target texts with limited information density (such as reference ambiguity in "That person looks familiar, didn't expect them to be a fan too") and texts with three or more targets that suffer from excessive semantic density and referential confusion, dual-target texts maintain an optimal balance between length and semantic content, facilitating better model comprehension and extraction.

Single-target samples significantly outperform others in recall at 99.36%. This is because such texts revolve around a single discussion object or topic, allowing the model to focus and achieve comprehensive coverage. However, their slightly lower performance on semantic relevance metrics suggests that models may overfit to explicit information while still having limited capability in processing implicit or ambiguous expressions.

Increasing target quantity leads to overall performance degradation. As the number of targets increases, all metrics show a declining trend, with multi-target samples scoring lowest at 45.42% overall. Analysis reveals several contributing factors: elevated semantic complexity, blurred boundaries due to cross-referenced and nested target expressions, and the tendency of models to generalize multiple targets as one, negatively affecting recall and overall performance.

4.3.2 Model-Oriented Difference Analysis

From a model perspective, we conduct a systematic analysis of Qwen2.5-7B and its four variants on target identification tasks in real-world scenarios, categorized by target quantity (Table 7 in Appendix D).

Overall model performance degrades as target quantity increases, reflecting the impact of task complexity. All models show declining trends across five metrics, particularly CoT-Qwen2.5-7B, whose C-Score drops from 70.81% to 47.14%, highlighting the challenges current models face when handling texts with multiple semantic targets.

CoT enhancement excels in single-target tasks but degrades significantly in multi-target settings. The CoT-augmented model achieves the highest score (70.81%) in single-target tasks, demonstrating strong reasoning capabilities in simple contexts. However, its performance drops sharply in multi-target scenarios to 47.14%, suggesting that reasoning chains become unstable when balancing multiple semantic focal points in complex contexts.

DeepSeek-R1-Distill-Qwen-7B demonstrates greater robustness in complex tasks. This model shows the strongest resilience in triple-target and multi-target scenarios, achieving scores of 60.69% and 55.09%, respectively. This indicates better generalization when processing semantically complex texts, likely due to the model's exposure to richer multi-target alignment corpora during distillation and fine-tuning.

4.4 Impact of Chain-of-Thought on Prompted LLMs

We investigate whether introducing chain-of-thought (CoT) improves LLM performance on the DGTA task. The results are presented in Table 3.

After introducing CoT, all LLMs show significant improvements in both target identification and stance determination. GLM4-9B's target identification score increases by 7 percentage points, indicating that step-by-step reasoning more effectively guides the model to capture key targets. Qwen2.5-7B's stance detection score improves by 4 percentage points, as the reasoning chain encourages the model to analyze systematically, reducing inferential leaps and incorrect judgments, thereby significantly enhancing stance classification performance.

4.5 Target Significance Difference Analysis

Considering different topic backgrounds and expression styles, targets in texts exhibit varying degrees of salience. We employ DeepSeek-V3 to classify annotated targets in the extracted test set as either "explicit" or "implicit", where the former refers to directly mentioned specific entities and the latter to abstract concepts requiring semantic understanding.

Based on experimental results, we select the high-performing integrated fine-tuned models DeepSeek-R1-Distill-Qwen-7B and Qwen2.5-7B for statistical analysis of target salience. Table 4 analysis reveals.

In target identification tasks, models perform significantly better when processing explicit targets compared to implicit ones. For instance, Qwen2.5-7B scores notably higher across multiple metrics, indicating that explicit targets have clearer semantic boundaries, facilitating extraction and matching.

In target identification tasks, implicit targets demonstrate more prominent performance in terms of recall. DeepSeek-R1 achieves a recall rate of 96.63% for implicit targets. Due to the abstract nature of implicit targets, models tend to generate multiple related expressions for coverage, enhancing recall but potentially reducing precision.

In stance detection tasks, explicit targets similarly demonstrate superior detection performance. Explicit targets help models more accurately grasp user attitudes, improving F1 scores, while implicit

Model	Target Identification					Stance Detection		
	BERT	BLEU	ROUGE	Recall	C-Score	P	R	F1
Qwen2.5-7B	82.47	28.26	63.69	91.16	61.87	64.22	67.54	65.05
DeepSeek-V3	76.87	25.65	50.38	92.05	56.45	69.25	**72.60**	70.64
GLM4-9B	77.98	24.16	51.99	94.14	58.38	68.50	69.20	66.90
CoT-Qwen2.5-7B	84.97	31.02	67.77	**94.23**	**66.70**	69.07	70.84	69.25
CoT-DeepSeek-V3	82.92	**32.33**	62.62	94.18	64.75	**73.68**	71.42	**71.06**
CoT-GLM4-9B	**85.56**	31.46	**68.38**	92.03	65.63	69.38	69.22	67.91

Table 3: Experimental results with chain-of-thought incorporated in the prompt (Unit: %)

Target	Model	Target Identification					Stance Detection		
		BERT	BLEU	ROUGE	Recall	C-Score	P	R	F1
Explicit (80.51%)	DeepSeek-R1-Qwen†	87.57	35.06	73.14	94.15	70.17	**69.40**	**72.95**	**70.88**
	Qwen2.5-7B†	**87.79**	**35.25**	**73.39**	93.85	**70.24**	65.34	65.54	64.09
Implicit (19.48%)	DeepSeek-R1-Qwen†	73.35	12.88	39.93	**96.63**	52.34	63.21	64.50	63.54
	Qwen2.5-7B†	73.22	12.93	39.56	95.51	51.62	65.15	63.29	63.90

Table 4: Performance comparison on explicit and implicit targets. (Unit: %. Explicit targets cover 80.51% of the data, while implicit targets cover 19.48%.)

targets increase judgment difficulty due to semantic ambiguity. Models equipped with reasoning capabilities can further enhance performance in these scenarios.

4.6 Case Analysis

We randomly sample cases from the integrated fine-tuned model DeepSeek-R1-Distill-Qwen-7B's prediction results for case analysis (Figure 6 in Appendix D).

Target identification performs well overall, but semantic fragmentation and stance judgment biases remain. In case (a), "the issue of Syrian women wearing black robes" is decomposed into "Syria", "women wearing black robes" and "Middle East" resulting in a loss of semantic integrity. Simultaneously, the model fails to identify the implicit critical attitude in "women wearing black robes" incorrectly judging it as neutral, reflecting its insufficient ability to reason about irony or implicit semantics.

Inconsistent target granularity and insufficient understanding of sarcastic expressions are observed. In case (b), although the model can extract multiple targets, it exhibits problems with mixed usage of different expressions for the same object, such as "Baidu's AI LLM" and "Wenxin Yiyan" both referring to "Baidu". Additionally, the stance judgment toward "Apple" as neutral fails to identify the metaphorical expression "a lady from a good family marrying a cowherd" reflecting the model's difficulty in recognizing stance under non-straightforward expressions like sarcasm and metaphor.

The model demonstrates optimal performance in scenarios with clearly defined targets and explicitly expressed stances. Case (c) revolves around the single target "Black Myth: Wukong" with direct textual expressions and distinct emotions, such as "stunning" and "holding back for so long" clearly conveying a positive stance. The model accurately identifies the target and correctly judges the stance, indicating high predictability in such samples.

5 Conclusion

Addressing the complexity and diversity of user stance expressions in real social contexts, we propose a new task: Zero-Shot Stance Detection in the Wild with Dynamic Target Generation and Multi-Target Adaptation. We construct a high-quality Chinese stance detection dataset covering multiple social scenarios. To accommodate the new characteristics of this task, we design an evaluation metric system that considers both target identification and stance determination. We propose two approaches for fine-tuning LLMs and compare them with pre-trained models and LLMs under various prompting methods. The experimental results clearly demonstrate that fine-tuned LLMs exhibit significant advantages in target extraction accuracy, stance classification robustness, and reasoning capability in complex linguistic contexts.

Limitations

We construct a dataset based on 240 users, each with approximately 300-400 expressions. Our current modeling approach does not incorporate user IDs and treats each stance expression as an independent sample, ignoring potential stance correlations between users. However, users with similar viewpoints often demonstrate consistent attitudes when facing the same targets, particularly evident in groups with shared interests. In future work, we will further utilize user ID information to develop user relationship-based stance modeling methods to capture stance consistency between users.

References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

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

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Ruofan Deng, Li Panl, and Chloé Clavel. 2022. Domain adaptation for stance detection towards unseen target on social media. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4969–4984.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

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*.

Hans Hanley and Zakir Durumeric. 2023. Tata: Stance detection via topic-agnostic and topic-aware embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11280–11294.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10729–10737.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Yang Li and Jiawei Yuan. 2022. Generative data augmentation with contrastive learning for zero-shot stance detection. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6985–6995.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *International Workshop on Semantic Evaluation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cornelius Bagus Purnama Putra, Diana Purwitasari, and Agus Budi Raharjo. 2022. Stance detection on tweets with multi-task aspect-based sentiment: A case study of covid-19 vaccination. *International Journal of Intelligent Engineering & Systems*, 15(5).

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, and Fei Huang. 2024. Qwen2.5 technical report.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 1868–1873.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15257–15272.

Shangkang Wang and Li Pan. 2024. Target-adaptive consistency enhanced prompt-tuning for multi-domain stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15585–15594.

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.

Jiaxi Wu, Jiaxin Chen, Mengzhe He, Yiru Wang, Bo Li, Bingqi Ma, Weihao Gan, Wei Wu, Yali Wang, and Di Huang. 2022. Target-relevant knowledge preservation for multi-source domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5301–5310.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. *arXiv preprint arXiv:2210.14299*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection.

Ruike Zhang, Hanxuan Yang, and Wenji Mao. 2023. Cross-lingual cross-target stance detection with dual knowledge distillation framework. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10804–10819.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

A Dataset Statistics and Analysis

Target	Number	Stance Distribution		
		Support	Against	Neutral
Single	27,148	12,554	5,232	9,362
Dual	25,312	25,122	10,223	15,279
Multi	18,471	35,245	14,880	27,222
Total	**70,931**	**72,921**	**30,335**	**51,863**

Table 5: Dynamic target quantity and stance distribution

Target	Number	Stance Distribution		
		Support	Against	Neutral
USA	907	112	560	235
Trump	816	262	287	267
Cheng Yi	765	710	1	54
China	726	380	158	188
iPhone	567	150	127	290
Israel	562	105	348	109
Wang Chuqin	414	286	11	117
Russia	413	120	146	147
Sun Yingsha	410	318	10	82
Huawei	350	256	9	85

Table 6: Top 10 targets by discussion frequency

B Prompt Template and Example for the Integrated Fine-tuning Strategy

Instruction:
You are a target identification expert. Based on the given Weibo comment, identify the main targets being discussed (such as people, events, or actions). If the comment involves multiple targets, list them all.
Output format:Target: [Target1]; [Target2]

Input:
They say stand-up comedy is the art of offense, but when a female comedian insults men, it's fine, while a male comedian gets punished for joking about women. I think we should just call it the art of offending men.

Output:
Target: stand-up comedy; gender double standards in comedic expression

Instruction:
You are a stance detection expert. Based on the given Weibo comment, determine the stance (Support / Against / Neutral) toward each of the provided targets. If multiple targets are given, assess the stance for each one in order.
Output format: Stance: [Stance1]; [Stance2]

Input:
They say stand-up comedy is the art of offense, but when a female comedian insults men, it's fine, while a male comedian gets punished for joking about women. I think we should just call it the art of offending men.
Target: stand-up comedy; gender double standards in comedic expression

Output:
Stance: Against; Against

Figure 5: Prompt template and example for the two-stage fine-tuning strategy

C Case Analysis

(a)
Text:
#Syrian Prime Minister Agrees to Transfer Power to the Opposition# Syria is already in such chaos—does whether women wear black robes even matter for the country or the Middle East? In war-torn regions, just staying alive is already fortunate. Life and death are beyond one's control, and it's just a matter of which faction ends up killing you.
Label:
Target: The issue of Syrian women wearing black robes
Stance: Against
Predict:
Target: Syria; women wearing black robes; the Middle East
Stance: Against; Neutral; Neutral

(b)
Text:
Although Baidu was the first major internet company to go all in on AI, its achievements in the AI field have become rather mediocre as its experts have gradually left. In today's fierce large model competition, I personally feel that Wenxin Yiyan's performance is just so-so. If Apple were to use Baidu's AI LLM, it would be like a lady from a good family marrying a cowherd. What do you think?
Label:
Target: Apple; Baidu
Stance: Against; Against
Predict:
Target: Apple; Baidu's AI LLM; Baidu; Wenxin Yiyan
Stance: Neutral; Against; Against; Against

(c)
Text:
#Black Myth: Wukong surpasses 4.5 million units sold across all platforms# The sales are truly stunning! I honestly didn't expect it to sell this much at first. I actually wanted to buy a copy myself, but unfortunately, I can't afford it. Today, Black Myth: Wukong is all over the internet. Everyone has been holding back for so long — finally, Black Wukong is here.
Label:
target: Black Myth: Wukong
stance: Support
Predict:
target: Black Myth: Wukong
stance: Support

Figure 6: Three representative cases

D Model-Oriented Difference Analysis

Model	Target	BERT	BLEU	ROUGE	Recall	C-Score
Qwen	Single	83.29	28.53	64.51	99.77	68.57
	Dual	86.86	34.07	71.29	95.15	69.90
	Triple	85.38	31.71	67.17	87.58	**61.62**
	Multi	82.84	31.41	64.98	73.46	53.30
Qwen†	Single	83.90	29.14	64.77	99.99	69.11
	Dual	86.89	33.46	71.22	94.26	68.98
	Triple	85.78	33.03	68.65	85.45	61.45
	Multi	83.64	30.11	62.08	75.17	53.58
Qwen‡	Single	83.59	30.21	63.61	99.99	68.91
	Dual	87.68	34.06	73.11	94.71	**70.28**
	Triple	83.35	30.38	63.66	85.15	59.59
	Multi	83.74	31.36	63.09	83.49	59.48
CoT -Qwen	Single	84.42	34.93	65.91	99.77	**70.81**
	Dual	82.78	30.89	62.63	94.71	64.75
	Triple	80.74	30.95	57.28	86.97	56.76
	Multi	78.66	26.37	52.53	73.05	47.14
Deep -Seek†	Single	84.03	29.32	65.37	99.99	69.35
	Dual	86.49	33.17	70.26	94.56	68.72
	Triple	85.64	32.46	67.95	85.45	60.69
	Multi	82.97	29.25	61.90	78.98	**55.09**

Table 7: Overall experimental results categorized by model (Unit: %. Qwen2.5-7B is abbreviated as Qwen, CoT-Qwen2.5-7B as CoT-Qwen, and DeepSeek-R1-Distill-Qwen-7B as DeepSeek.Bolded values indicate the highest C-Score within each target quantity category.)