Zero-Shot Conversational Stance Detection: Dataset and Approaches

Yuzhe Ding, Kang He, Bobo Li, Li Zheng, Haijun He, Fei Li,
Chong Teng, Donghong Ji[*]

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of
Education, School of Cyber Science and Engineering, Wuhan University
`{yuzheding,lifei_csnlp,dhji}@whu.edu.cn`

Abstract

Stance detection, which aims to identify public opinion towards specific targets using social media data, is an important yet challenging task. With the increasing number of online debates among social media users, conversational stance detection has become a crucial research area. However, existing conversational stance detection datasets are restricted to a limited set of specific targets, which constrains the effectiveness of stance detection models when encountering a large number of unseen targets in real-world applications. To bridge this gap, we manually curate a large-scale, high-quality zero-shot conversational stance detection dataset, named ZS-CSD, comprising 280 targets across two distinct target types. Leveraging the ZS-CSD dataset, we propose SITPCL, a speaker interaction and target-aware prototypical contrastive learning model, and establish the benchmark performance in the zero-shot setting. Experimental results demonstrate that our proposed SITPCL model achieves state-of-the-art performance in zero-shot conversational stance detection. Notably, the SITPCL model attains only an F1-macro score of 43.81%, highlighting the persistent challenges in zero-shot conversational stance detection. Our data and code are available at https://github.com/whu-yzding/ZS-CSD.

1 Introduction

With the rise of social networks, people increasingly express their opinions views online. The aim of stance detection is to determine people's opinionated standpoint or attitude (e.g. Favor, Against, or Neutral) expressed in text towards a specific target (Somasundaran and Wiebe, 2010; Mohammad et al., 2016; Sobhani et al., 2017; AlDayel and Magdy, 2021), such as the COVID-19 vaccine, LGBTQ rights, and educational philosophy. Stance detection has a wide range of potential applications, including in fields such as emotion recognition (Zhang et al., 2020; Upadhyaya et al., 2023; Zheng et al., 2023), argument mining (Sobhani et al., 2015; Rajendran et al., 2018; Sirrianni et al., 2020; Carnot et al., 2023; Zheng et al., 2025b), and rumor detection (Yu et al., 2020; Yang et al., 2024; Zheng et al., 2025a).

Existing research on stance detection is typically divided into in-target, cross-target, and zero-shot categories for training and evaluating models, with a primary focus on analyzing individual sentences (Zarrella and Marsh, 2016; Küçük and Can, 2020; Allaway and McKeown, 2020; Li et al., 2023b). However, on social media platforms, users not only express opinions through direct posts but also engage in conversations and express their views in the comment sections of tweets they find interesting. Multi-party conversation understanding involves participant diversity, interaction dynamics, context dependence, and rich emotions and tones, making stance detection in conversational scenarios particularly challenging.

Currently, five conversational stance detection datasets have been developed, including SRQ (Villa-Cox et al., 2020), CANT-CSD (Li et al., 2023c), CTSDT (Li et al., 2023d), MT-CSD (Niu et al., 2024b) and MmMtCSD (Niu et al., 2024a). Despite the diversity of conversational stance detection corpora, these datasets have several limitations. Firstly, current conversational stance detection datasets focus only on a few specific topics and only contain noun phrase type targets or the post type targets. The number of targets is relatively small, and the largest MT-CSD contains only five targets. Secondly, in the previous conversation stance detection dataset annotation and modeling process, only the reply relationship and historical context information were considered during the annotation and modeling process, without considering the speaker context information and

---

[*]Corresponding author.

potential speaker interaction information. Finally, current research in conversational stance detection is restricted to two types of tasks: in-target and cross-target stance detection, lacking exploration of zero-shot conversational stance detection tasks that better reflect real-world scenarios.

To address these issues, we introduce ZS-CSD, the first zero-shot multi-turn, multi-party conversational stance detection dataset. This dataset encompasses a substantial number of real-world human interactions on trending topics, events, and opinions from the Chinese online social media platform Weibo. Compared to existing conversational stance detection datasets, the ZS-CSD dataset captures a more diverse range of targets, including noun phrase type targets (i.e. entities, events, or topics) and claim type targets (i.e. phrases that express opinions). Furthermore, the stance annotation process in ZS-CSD thoroughly considers speaker context information and potential speaker interaction information, aspects that have been overlooked in previous studies. The specific details of ZS-CSD and other stance detection datasets can be seen in Table 1.

Based on the ZS-CSD dataset, we design a new conversational stance detection task: *target-based zero-shot conversational stance detection*, where a number of completely unseen targets are used to evaluate the conversational stance detection model. Our new task challenges in two main aspects. First, the model needs to learn how to judge the user's stance toward a given target from the training set and transfer the ability to unseen targets in the test set. Second, the model needs to fully understand the speaker context information and potential speaker interaction information in the conversation.

To solve these challenges, we present a SITPCL model. Specifically, we construct a speaker interaction network to model intra-speaker and inter-speaker dependencies. Additionally, we introduce a target-aware prototypical contrastive learning method to enable the model to better understand the targets. Experimental results demonstrate that our proposed SITPCL model achieves state-of-the-art performance in zero-shot conversational stance detection.

The main contributions of this paper can be summarized as follows:

- We propose the first zero-shot multi-turn multi-party conversational stance detection dataset ZS-CSD, and propose a new target-based zero-shot conversational stance detection task.

- Compared to previous conversational stance detection datasets, our ZS-CSD incorporates a wider range of target types and an increased number of targets. Additionally, we take into account speaker context information and speaker interaction information. The final ZS-CSD dataset comprises 17,063 conversation samples from 8,667 users, including 113 noun phrase targets and 167 claim targets.

- We introduce a speaker interaction and target-aware prototypical contrastive learning model SITPCL to benchmark the ZS-CSD task. Our model effectively learns intra-speaker and inter-speaker dependencies and can better understand the unseen targets in the zero-shot setting.

2 Related Work

2.1 Stance Detection Datasets

Sentence-Level stance detection datasets Over the years, various datasets have been introduced as benchmarks for sentence-level stance detection, as summarized in Table 1. The SEM16 dataset (Mohammad et al., 2016) pioneered stance detection from X, while P-Stance (Li et al., 2021a) focuses on longer political tweets. The large-scale WT-WT corpus (Conforti et al., 2020) offers extensive labeled data, and specialized datasets like the COVID-19 Stance Detection dataset (Glandt et al., 2021) target specific discourse. The VAST dataset (Allaway and McKeown, 2020) is the most comprehensive, enabling zero-shot stance detection across thousands of topics. More recently, C-STANCE (Zhao et al., 2023) and EZ-STANCE (Zhao and Caragea, 2024) have expanded zero-shot stance detection using Weibo and X data.

Conversation-level stance detection datasets In real-world scenarios, users express perspectives conversationally, making conversational stance detection—identifying stances within discussion threads—increasingly relevant. The SRQ dataset (Villa-Cox et al., 2020) pioneered stance detection in comment data but was limited to single-turn replies and shallow conversations. CANT-CSD (Li et al., 2023c) addressed multi-turn interactions with a deeper commenting corpus, while CTSDT (Li et al., 2023d) relied primarily on automated annotations. Stanceosaurus (Zheng et al., 2022) constructed a conversational stance classification

Name	Source	Type	Task	# Target(s)	Target Type NP	C	Language	Size
SEM16	X	S	In-target	6	✓	×	English	4,870
P-stance	X	S	In-target	3	✓	×	English	21,574
WT-WT	X	S	In-target	5	✓	×	English	51,284
COVID19	X	S	In-target	4	✓	×	English	6,133
VAST	Comments	S	ZSSD	5,634	✓	×	English	18,545
C-STANCE	Weibo	S	ZSSD	40,204	✓	✓	Chinese	48,126
EZ-STANCE	X	S	ZSSD	40,678	✓	✓	English	47,316
SRQ	X	C	In-target	4	✓	×	English	5,220
CANT-CSD	HK SM	C	In-target	1	✓	×	Cantonese	5,876
CTSDT	X	C	In-target	1	✓	×	English	53,861
STANCEOSAURUS	X	C	Cross-target	251	×	✓	English,Hindi,Arabic	28,033
MT-CSD	Reddit	C	In-target Cross-target	5	✓	×	English	15,876
MmMtCSD	Reddit	C	In-target Cross-target	3	✓	×	English	21,340
ZS-CSD(Ours)	Weibo	C	ZS-CSD	280	✓	✓	Chinese	17,063

Table 1: Comparison of stance detection datasets: Type S denotes sentence-level data, while Type C indicates conversation-level data. Target Type NP refers to noun phrase targets, and Target Type C refers to claim targets.

dataset focused on misinformation-related claims. The dataset is designed to facilitate the identification of misinformation by analyzing the stances expressed within the surrounding discussion. MT-CSD (Niu et al., 2024b) expanded conversation depth across five targets, and MmMtCSD (Niu et al., 2024a) extended it to multimodal scenarios for broader applicability.

2.2 Stance Detection Approaches

In recent years, various approaches based on traditional machine learning and deep learning have been proposed to address stance detection for specific targets (Augenstein et al., 2016; Wei and Mao, 2019). The task settings can generally be categorized into in-target, cross-target, and zero-shot settings. Most previous work has focused on in-target stance detection, where a classifier is trained and evaluated on the same target (Mohammad et al., 2016; Li and Caragea, 2019, 2021). However, it is often challenging to obtain sufficient annotated data for each target, and conventional models tend to perform poorly when generalized to unseen targets. This limitation has motivated research on cross-target stance detection (Xu et al., 2018; Zhang et al., 2020; Li et al., 2021b), where a classifier is adapted to different but related targets. Furthermore, zero-shot stance detection (ZSSD), which aims to identify stances toward a large number of unseen targets, has garnered considerable attention in recent years (Liu et al., 2021; Liang et al., 2022;

Luo et al., 2022).

All current methods for stance detection can generally be classified into fine-tuning based approaches (Zhu et al., 2022; Liang et al., 2024; Wang and Pan, 2024) and LLM-based approaches (Li et al., 2023a; Weinzierl and Harabagiu, 2024; Zhang et al., 2024). Fine-tuning based methods involve adding a fully connected layer to the [CLS] token of a pre-trained model and fine-tuning the model specifically for stance detection. Besides, large language models (LLMs) have demonstrated remarkable capabilities across diverse applications due to their inherent semantic understanding (Zhang et al., 2023; Li et al., 2024; Zheng et al., 2024). This semantic understanding presents exciting opportunities for stance detection. Most LLMs can effectively perform stance prediction via zero-shot prompting, significantly enhancing usability (Ding et al., 2024).

3 Dataset Construction

Data Collection To obtain authentic multi-party conversations with diverse targets, we collected the original corpus of our ZS-CSD dataset from Weibo[1], China's largest social media platform. We utilized the Weibo API[2] to collect 2 million posts along with the online discussions under each post from 80 verified and widely followed news bloggers. After obtaining the raw data, we selected a

[1]https://weibo.com/
[2]https://open.weibo.com/

	# Targets		# Conversations		Avg. Length
	N	C	N	C	
Train	79	111	5,604	6,299	37.84
Val	17	28	961	1615	37.28
Test	17	28	967	1,617	36.98
Overall	113	167	7,532	9,531	37.63

Table 2: The dataset split statistics for the ZS-CSD are presented, where N and C denote noun phrase type targets and claim type targets, respectively. Avg. Length refers to the average number of characters per utterance.

	Depth of Conversation					
	1	2	3	4	5	>=6
Train	210	1,905	3,732	2,548	1,636	1,872
Val	45	399	832	530	345	425
Test	45	445	877	585	351	281
Overall	300	2,749	5,441	3,663	2,332	2,578

Table 3: Statistics of samples at various depths in the ZS-CSD dataset are presented.

series of keywords representing controversial topics for data screening and classified them into six domains. The list of domains and keywords are provided in Appendix A. To ensure the high quality of the ZS-CSD dataset, we implemented two rigorous preprocessing steps. Firstly, we observed that certain comments, which directly respond to the original post, can incite heated discussions. To address this, we selected all first-level comments that directly reply to the original post as the root node of the discussion tree. Subsequently, we gradually constructed multiple discussion tree as candidates based on the reply relationship. Secondly, to balance the number of conversation samples across different depths in the dataset, we collected samples from the discussion tree based on their depth. Ultimately, we obtained a diverse set of samples spanning six levels $1, 2, 3, 4, 5, \geq 6$ for subsequent annotation.

Human Annotation We built our annotation platform using Doccano[3], with each thread sample as a separate annotation task. Three expert annotators and five NLP researchers, all native Chinese speakers, participated in the process. Annotation was conducted in two steps: target annotation followed by stance annotation. In Step 1, three expert annotators collaboratively identify one or two targets per conversation after thoroughly reviewing the discussion. The selected targets must: (i) allow for clear stance expression, (ii) avoid near-unanimous stances, such as those violating public values, and (iii) be either noun phrase targets (e.g., entities, concepts, events) or claim targets (e.g., opinions). This process yielded 17,063 conversation-target pairs for stance annotation in Step 2. In step 2, given 17,063 pairs, five annotators were asked to combine conversation history, speaker context, and

speaker interaction information to annotate the current utterance's stance toward the given target: *favor*, *against*, or *neutral*. Neutrality to the target can occur in two cases: i) The user neither supports nor opposes the target. ii) The comment is irrelevant to the target. To ensure data quality and reliability, each data instance is annotated by at least two annotators. In cases of disagreement between the annotators, a panel of at least two expert annotators will discuss the issue to reach a final judgment. After obtaining the annotation results, we calculated a Cohen's Kappa score of 0.83, indicating that our annotated corpus is of high quality.

Dataset Statistics Following the VAST dataset guidelines (Allaway and McKeown, 2020), we split the annotated data into training(70%), validation(15%), and test(15%) sets for target-based zero-shot conversational stance detection. Each set contains distinct conversations and targets with no overlap. The dataset distribution is detailed in Table 2 and Table 3.

4 Our Approach

4.1 Task Definition

The primary objective of the zero-shot conversational stance detection task is to determine the stance of the current utterance toward a given target, utilizing both the conversation history and speaker context. Notably, the targets in the test set are unseen in both the training and validation sets. Formally, given a target t, a multi-turn conversation C, a collection of speakers S and a set of stance labels $\mathcal{E}$, if conversation C consists of n utterances, it can be denoted as $C = (s_i, u_i)_{i=1}^n$, where u_i is the utterance of i-th turn and $s_i \in S$ is the corresponding speaker of u_i. If we denote u_n as the current utterance, the model can only take (s_n, u_n, t) and the previous utterances $\{(s_i, u_i)\}_{i=1}^{n-1}$ as input to predict the stance label y_n of u_n toward t, where $y_n \in \mathcal{E}$.
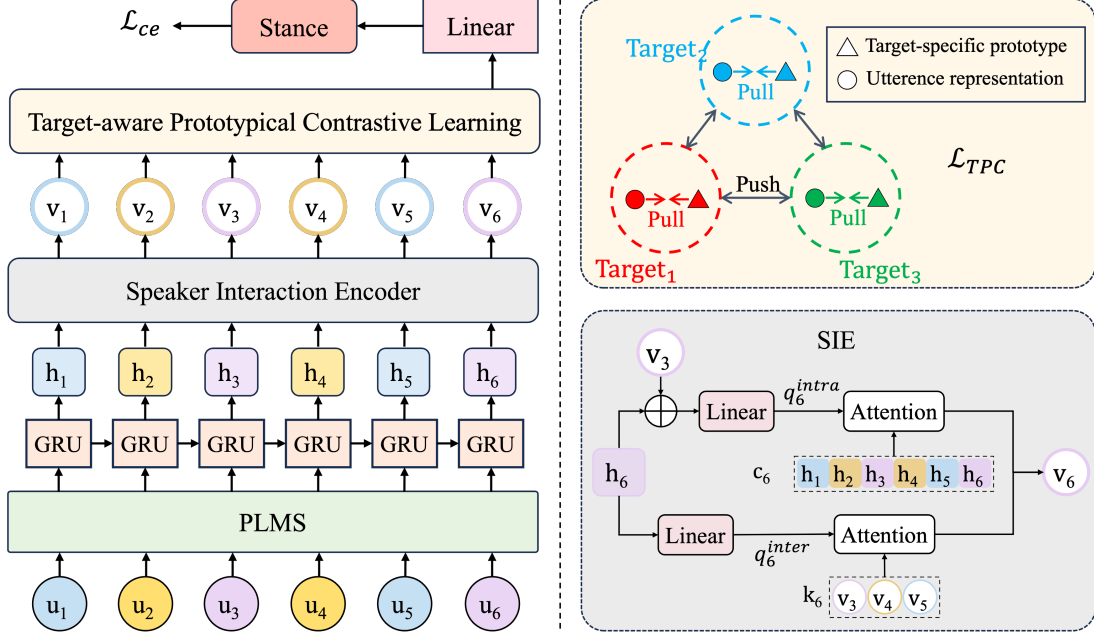
Figure 1: The architecture of our SITPCL framework. Initially, we use the pre-trained language model and GRU network to obtain a contextualized representation of the entire conversation. Next, we capture speaker interactions using the Speaker Interaction Encoder. To enhance representation learning, we incorporate target-aware prototypical contrastive learning. Finally, the enriched utterance representation is leveraged for stance detection.

4.2 Utterance Encoder

We adopt a pre-trained language model (PLM), such as BERT (Devlin et al., 2019), to encode utterance-target pairs within the conversation. Since the entire conversation may exceed BERT's maximum input length, we encode each utterance-target pair individually using separate PLMs. We concatenate each utterance u_i with its speaker s_i and add a template with a placeholder to implicitly express the speaker's stance:

$$m_i = PLM([s_i, \mathbf{u}_i, \langle /s \rangle, template, \langle /s \rangle]) \quad (1)$$

$$template = \; s_i \text{ expresses } \langle mask \rangle \text{ towards target} \quad (2)$$

where $</s>$ denotes the separator used to distinguish different contents, $<mask>$ is the placeholder (known as the masked token) used for learning the stance label toward given target expressed in this utterance. To get the contextual representations, $m_1, m_2, ..., m_n$ are passed to a GRU layer (Cho et al., 2014). The final representations are defined by:

$$[h_0, h_1, \ldots, h_n] = GRU([m_0, m_1, \ldots, m_n]) \quad (3)$$

where $h_i \in \mathbb{R}^d$. Here, d is the dimension of each utterance vector.

4.3 Speaker Interaction Encoder

To accurately capture the dynamics of speaker interactions, we propose an encoder network specifically designed to model both inter-speaker and intra-speaker dependencies.

Intra-Speaker Dependency Modeling. For a given utterance u_i which is uttered by speaker s_{u_i}, we take the last previous speaker enhanced representation of s_{u_i} and previous context c_i = $\{h_1, h_2, ..., h_i\}$ as input, modeling the intra-speaker dependency implied in the conversation.

First, we utilize the last previous speaker enhanced representation of s_{u_i} and the representation vector of u_i to obtain the query vector q_i^{intra} as follows:

$$q_i^{\text{intra}} = W_q^{\text{intra}} \left[v_{s_{u_i}} \oplus h_i \right] + b_q^{\text{intra}} \quad (4)$$

where $v_{s_{(u_i)}} \in \mathbb{R}^{h \times 1}$ means the last previous speaker enhanced representation of s_{u_i}; $\oplus$ means the concatenation operation; $W_q^{\text{intra}} \in \mathbb{R}^{h \times 2h}$ and $b_q^{\text{intra}} \in \mathbb{R}^1$ are model parameters.

An attention mechanism is then introduced to model the intra-speaker context, with the query vector q_i^{intra} acting as the query and the previous context c_i serving as the key and value in the attention framework. This mechanism generates the

intra-speaker state vector v_i^{intra} as follows:

$$\alpha_i^{\text{intra}} = \text{softmax}\left(\mathbf{W}_1\left(\mathbf{q}_i^{\text{intra}} \odot \mathbf{c}_i\right) + \mathbf{b}_1\right) \quad (5)$$

$$\mathbf{v}_i^{\text{intra}} = \alpha_i^{\text{intra}} \circ \mathbf{c}_i^T \quad (6)$$

Inter-Speaker Dependency Modeling. To model inter-speaker dependencies, we use the context vector h_i. The key vector k_i for the attention mechanism is calculated prior local information $\{v_j \mid j < i\}$ linked to speaker s_{u_i}. This approach captures the latent inter-speaker dependency. To derive the inter-speaker state vector for u_i, we apply the following procedure:

$$q_i^{\text{inter}} = W_q^{\text{inter}} h_i + b_q^{\text{inter}} \quad (7)$$

$$\alpha_i^{\text{inter}} = \text{softmax}\left(\mathbf{W}_2\left(\mathbf{q}_i^{\text{inter}} \odot \mathbf{k}_i\right) + \mathbf{b}_2\right) \quad (8)$$

$$\mathbf{v}_i^{\text{inter}} = \alpha_i^{\text{inter}} \circ \mathbf{k}_i^T \quad (9)$$

After obtaining the intra-speaker and inter-speaker state vectors, v_i^{intra} and v_i^{inter}, we combine them to obtain the current speaker enhanced representation v_i, where $W_3 \in \mathbb{R}^{h \times 2h}$ and $b_3 \in \mathbb{R}^1$ are model parameters.

$$v_i = W_3\left[v_i^{\text{intra}} \oplus v_i^{\text{inter}}\right] + b_3 \quad (10)$$

4.4 Target-aware Prototypical Contrastive Learning

Contrastive learning aims to learn an effective representation space by making samples closer to their semantically similar positive samples and away from other samples. Based on this, we design a target-aware prototype-level contrastive learning objective in a mini-batch.

Taking target t as an example, its prototype representation p_t is defined as the average of all utterances representation associated with t. The p_t is iteratively refined as the corresponding utterances representation are optimized. The target-aware prototypical contrastive loss is:

$$p_t = \frac{1}{|C_t|} \sum_{i \in C_t} \mathbf{v}_i \quad (11)$$

$$\mathcal{L}_{\text{TPC}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{v}_i, p_{y_i})/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(\mathbf{x}_i, p_k)/\tau)} \quad (12)$$

where C_t is the number of utterances for the target t in a mini-batch, p_t is the prototype representation of the target t, $sim(.,.)$ is the similarity metric, τ is the temperature parameter, K is the number of targets, and N is the number of all utterances.

Therefore, the representation of any utterance in the mini-batch will move closer to its corresponding target prototype while distancing itself from the prototype representations of other targets. Simultaneously, the prototype representations of different targets will progressively diverge from one another.

4.5 Stance Detection

The final output representation of the current utterance v_i is then input into a classifier with a softmax function to produce a predicted stance distribution $\hat{y}_i \in \mathbb{R}^{d_y}$.

$$\hat{y}_i = \text{softmax}(W \cdot v_i + b) \quad (13)$$

where d_y is the dimensionality of stance labels. $W \in \mathbb{R}^{d_y \times h}$ and $b \in \mathbb{R}^{d_y}$ are trainable parameters. We adopt a cross-entropy loss between predicted distribution $\hat{y}_i$ and ground-truth distribution y_i to train the classifier:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N_b} \sum_{j=1}^{d_y} y_i^j \log \hat{y}_i^j \quad (14)$$

4.6 Learning Objective

The learning objective of our proposed model is to train the model by jointly optimizing a supervised loss of stance detection $\mathcal{L}_{\text{CE}}$ with a target-aware prototypical contrastive loss. The overall loss $\mathcal{L}$ is formulated by summing up two losses together:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \gamma \mathcal{L}_{\text{TPC}} \quad (15)$$

where γ are the hyper-parameter for weights balance.

5 Experimental Setup

5.1 Baseline Methods

Fine-tuning with PLMs. (1) RoBERTa (Liu, 2019) represents an enhancement over BERT, utilizing larger batch sizes and more data for training; (2) Branch-BERT (Li et al., 2023c) utilizes a CNN to extract important ngrams features incorporating contextual information in conversation threads. (3) GLAN (Niu et al., 2024b) architecture adopts a three-branch structure to address the intricacies of conversational dynamics comprehensively.

LLM-based methods. We use current state-of-the-art LLMs with various parameter sizes, including the closed-source GPT3.5[4], GPT4o-mini[5]

---

[4]https://platform.openai.com/docs/models/gpt-3-5
[5]https://platform.openai.com/docs/models/gpt-4o-mini

METHOD	Mixed targets				Noun-phrase targets				Claim targets			
	Favor	Against	Neutral	All	Favor	Against	Neutral	All	Favor	Against	Neutral	All
Llama3-8B	35.09	39.34	33.29	35.91	38.31	42.68	37.13	39.37	33.26	37.44	30.68	33.79
GPT4o-mini	24.75	38.63	51.10	38.16	20.78	46.86	52.99	40.21	26.87	32.77	50.07	36.57
GPT3.5	28.04	38.89	**53.84**	40.25	38.41	48.79	54.84	47.35	20.99	32.29	**53.32**	35.53
Llama3-70B	40.81	38.04	44.37	41.07	**48.77**	50.40	47.00	**48.72**	35.96	30.98	42.75	36.56
Qwen2.5-14B	33.84	**44.01**	50.48	42.78	27.67	**50.56**	**55.58**	44.60	36.86	**40.23**	46.93	41.34
Roberta	37.37	38.94	44.50	40.27	37.50	42.57	48.31	42.79	37.30	36.64	42.20	38.71
Branch-BERT	43.11	38.19	41.97	41.09	46.29	43.73	43.41	44.48	41.43	34.34	41.13	38.96
GLAN	**43.68**	34.91	46.79	41.79	45.91	37.65	51.58	45.04	**42.43**	33.18	43.88	39.83
SITPCL(Ours)	41.03	41.34	49.07	**43.81**	40.99	47.14	54.49	47.54	41.06	37.34	46.02	**41.47**

Table 4: We compare different models on the target-based zero-shot conversational stance detection task. The performance is evaluated using the F1 score for the categories: favor, against, neutral, and the overall F1-macro.

METHOD	Mixed targets			
	Favor	Against	Neutral	All
SITPCL(Ours)	41.03	41.34	49.07	43.81
W/o SIE	40.11	43.46	45.15	42.91
W/o TPCL	31.50	47.26	49.02	42.59
W/o BOTH	37.80	42.14	44.28	41.40

Table 5: Experimental results of ablation study.

METHOD	DEPTH					
	1	2	3	4	5	≥ 6
Llama3-8B	33.42	40.25	35.91	36.38	27.19	35.04
GPT4o-mini	34.03	30.94	37.65	38.83	38.01	43.59
GPT3.5	22.51	36.67	35.54	42.84	39.31	43.54
Llama3-70B	54.12	40.90	39.25	38.69	44.40	39.50
Qwen2.5-14B	40.23	39.42	42.49	**43.99**	42.34	38.70
Roberta	25.49	36.26	39.02	38.40	37.99	46.55
Branch-BERT	26.90	**43.76**	36.73	37.90	38.59	46.11
GLAN	29.23	36.21	39.39	38.11	38.37	**48.93**
SITPCL(Ours)	40.19	38.37	42.84	40.62	44.73	45.25

Table 6: Evaluation of the performance of each model on instances with different conversation depths. The performance is reported using the F1-macro score, which encompasses the favor, against, and neutral classes.

and open-source LLMs such as Qwen2.5-14B[6], Llama3-70B[7] and Llama3-8B[8]. We evaluate these models exclusively under the zero-shot prompt setting. The prompt details are provided in Appendix B.

5.2 Training Settings and Evaluation Metrics

We perform experiments using two NVIDIA RTX 3090 GPUs. We initialize all the pre-trained language models using Chinese-Roberta-wwm-ext (Cui et al., 2021). The hidden dimension of the GRU layer is set to 768 to match the output dimension of the PLM. During training, we use a batch size of 16 and run for 20 epochs. The total training time is less than 3 hours. The optimizer is AdamW, with a learning rate of 1e-5 along with a weight decay of 1e-6. We employ five different random seeds for all experiments, and the final score is the average of these five runs. Consistent with previous studies (Zhao et al., 2023), we use the F1 score for each class and the macro-average F1 score across all classes as evaluation metrics.

6 Experimental Results

6.1 Main Results

We present the experimental results of various models within the target-based zero-shot conversational stance detection task, as shown in Table 4. The mixed targets refer to the overall result on the test set without distinguishing the target type, while the noun phrase targets and the claim targets only calculate the test results of the corresponding target examples. From the results, we make the following observations: 1) Our proposed method, SITPCL, outperforms all baselines in the mixed-target setting, demonstrating its effectiveness. 2) The best F1-macro score for all test examples (mixed targets) is only 43.81%, underscoring the challenge of the target-based zero-shot conversational stance

---

[6]https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
[7]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
[8]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

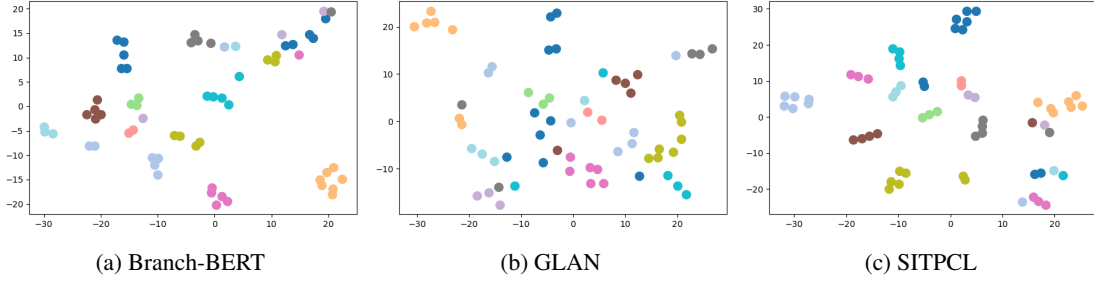(a) Branch-BERT   (b) GLAN   (c) SITPCL

Figure 2: Visualization of intermediate embeddings from three models. Dots of different colors represent utterances of different targets.
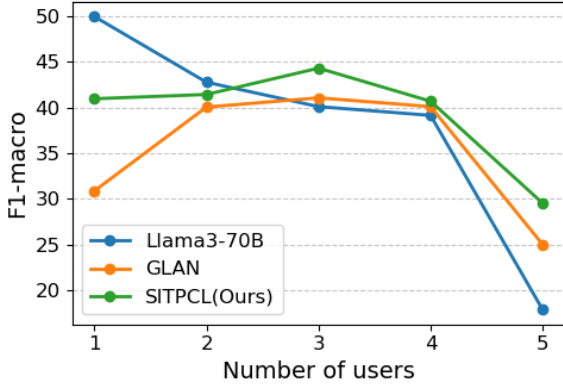

Figure 3: Evaluation of the performance of each model on instances with different user couts.

6.2 Ablation Study

To rigorously assess the contribution of each module in SITPCL, we conduct ablation studies, as shown in Table 5. The results indicate that removing either the SIE or TPCL module individually reduces model performance across all settings. Specifically, the F1-macro score drops by 0.90% and 1.22% for mixed targets when removing SIE and TPCL, respectively, demonstrating their effectiveness. Removing both modules results in a larger performance decline of 2.41%, emphasizing that their integration enhances model optimization.

6.3 Further Analysis

Impact of Conversation Depth This analysis evaluates the performance of various conversational stance detection models across different conversation depths. The results are presented in Table 6. Compared to other PLM-based fine-tuning methods, our model demonstrates significant superiority across all samples except those with depths of 2 and ≥ 6, confirming the effectiveness of SITPCL. Additionally, except for Llama3-8B, Llama3-70B and Qwen2.5-14B, all other methods achieve their best performance on samples with depth ≥ 6, likely because a longer conversation history provides more contextual clues for stance expression.

Impact of user count Figure 3 presents the evaluation results across different user counts. Our SITPCL model performs best when there are more than two users. However, as the number of users increases, all models exhibit a performance decline, highlighting the growing challenges of understanding complex user interactions.

6.4 Error Analysis

We randomly selected 300 error samples from the test set for detailed analysis and categorized the errors into four types: 1) Failure to understand

detection task and dataset. 3) All methods perform significantly better on noun phrase type targets than on claim type targets, indicating the latter is more challenging in zero-shot conversational stance detection. 4) While Llama3-70B achieves the best performance on noun phrase type targets, it falls behind PLM-based fine-tuning methods on claim type targets. This suggests that while LLMs leverage extensive knowledge for noun phrases, they struggle with directly understanding claim type targets. 5) In the baseline evaluation of large language models, we observed significant performance disparities across different stance categories. For instance, in the mixed-target setting, GPT-3.5 achieved its highest performance on the neutral class with an accuracy of 53.84%, while its performance on the favor class was considerably lower at only 28.04%. We attribute this discrepancy to potential stance biases acquired during the model's pretraining phase or to alignment mechanisms implemented for safety, which may steer the model toward producing responses that are consistently neutral or objective.

the target (48%): The model lacks the necessary knowledge to comprehend the target or loses track of the target during integration with the conversational context. 2) Failure to consider user interaction context (61%): The model fails to incorporate the user's interaction history, leading to incorrect stance interpretation. 3) Over-inference (29%): The model excessively interprets the user's stance based on limited clues, resulting in misjudgment of the user's intent. 3) Inability to recognize sarcasm (21%): The model struggles with identifying implicit expressions such as sarcasm, reflecting its limited pragmatic understanding. In Appendix C, we provide an error prediction example and model improvement analysis for each error type.

6.5 Visualization

To qualitatively demonstrate the SITPCL model's ability to better capture target information, we use t-SNE to visualize intermediate utterance representations generated by different methods. As shown in Figure 2, dots of the same color represent utterances associated with the same target. In both Branch-BERT and GLAN, utterance representations for different targets exhibit significant overlap. In contrast, SITPCL produces more distinct representations, with utterances related to the same target forming well-defined clusters. This validates that SITPCL's target-aware prototypical contrastive learning effectively separates utterance representations, enhancing model performance.

7   Conclusion

In this paper, we present ZS-CSD, the first zero-shot multi-turn multi-party conversational stance detection dataset, and introduce a novel and challenging task: target-based zero-shot conversational stance detection. This task evaluates models on a diverse set of previously unseen targets. Compared to existing datasets, ZS-CSD contain 113 noun phrase targets and 167 claim targets. Additionally, it integrates both speaker context and speaker interaction information. To benchmark the ZS-CSD task, we propose SITPCL, a speaker interaction and target-aware prototypical contrastive learning model. Experimental results indicate that ZS-CSD presents a significant challenge, particularly for claim type targets.

Acknowledgments

Limitations

Our ZS-CSD dataset is primarily derived from Chinese social media platforms and contains a limited number of targets. In the future, we aim to develop more complex zero-shot conversational stance detection datasets, incorporating multilingual and multimodal scenarios. This expansion will facilitate deeper research in the field and further advance zero-shot conversational stance detection.

Ethical Considerations

Our dataset does not provide any personally identifiable information. Microblogs are collected using generic keywords instead of user information as queries, therefore our dataset does not have a large collection of microblogs from an individual user. Thus our dataset complies with Sina Weibo's information privacy policy. To annotate the data, we recruited 8 senior researchers working in natural language processing, each of whom was paid $6.5 per hour (higher than the average salary for similar jobs in the local area). The entire annotation process lasted 2 months, and the average annotation time of the 8 researchers was 100 hours.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Miriam Louise Carnot, Lorenz Heinemann, Jan Braker, Tobias Schreieder, Johannes Kiesel, Maik Fröbe, Martin Potthast, and Benno Stein. 2023. On stance detection in image retrieval for argumentation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2562–2571.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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.

Daijun Ding, Li Dong, Zhichao Huang, Guangning Xu, Xu Huang, Bo Liu, Liwen Jing, and Bowen Zhang. 2024. EDDA: An encoder-decoder data augmentation framework for zero-shot stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5484–5494.

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 (Volume 1: Long Papers)*, pages 1596–1611.

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

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

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

Yingjie Li 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.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. 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. 2021a. 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.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023c. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*.

Yupeng Li, Dacheng Wen, Haorui He, Jianxiong Guo, Xuan Ning, and Francis CM Lau. 2023d. Contextual target-specific stance detection on twitter: Dataset and method. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 359–367.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. 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)*, pages 81–91.

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

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024a. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3867–3876.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024b. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2018. Is something better than nothing? automatically predicting stance-based arguments using deep learning and small labelled dataset. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 28–34.

Joseph Sirrianni, Xiaoqing Liu, and Douglas Adams. 2020. Agreement prediction of arguments in cyber argumentation for detecting stance polarity and intensity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5746–5758.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.

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.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Intensity-valued emotions help stance detection of climate change twitter data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6246–6254.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Shaokang 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.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1173–1176.

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

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13423–13439.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th An-*

nual Meeting of the Association for Computational Linguistics, pages 3188–3197.

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.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2132–2151.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025a. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26057–26065.

Li Zheng, Hao Fei, Fei Li, Bobo Li, Lizi Liao, Donghong Ji, and Chong Teng. 2024. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19688–19696.

Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2023. Ecqed: emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969*.

Li Zheng, Sihang Wang, Hao Fei, Zuquan Peng, Fei Li, Jianming Fu, Chong Teng, and Donghong Ji. 2025b. Enhancing hyperbole and metaphor detection with their bidirectional dynamic interaction and emotion knowledge. In *Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2070–2075.

A Query Keywords and Domains

We selected 28 keywords covering controversial topics. We grouped the 28 keywords into 6 areas: "Culture and Education", "Policy", "Science and Technology", "Healthcare", "Entertainment and Consumption" and "Social Issues" which can be seen in Table 7.

B Prompt for LLMs

We use current state-of-the-art LLMs with various parameter sizes, including the closed-source GPT3.5 and open-source LLMs such as Llama3-70B and Llama3-8B. We evaluate these models exclusively under the zero-shot prompt setting. Figure 4 shows the prompt we use.

C Error case Analysis

We categorize the errors into four major types: 1) Failure to Understand the Target (the model lacks the knowledge to understand the target or loses the understanding of the target when incorporating conversation context); 2) Failure to Consider the Context of User Interaction (the model fails to take the user's interaction history into account); 3) Over-Reasoning (the model over-interprets the user's stance based on a few cues, failing to accurately grasp the user's intent); 4) Failure to Recognize Sarcasm (the model lacks the ability to understand implicit expressions such as sarcasm). Below, we provide one error prediction example and model improvement analysis for each type.

As shown in Figure 5, the SITPCL model did not accurately align with the target to understand User 2's stance, instead relying solely on emotional cues to classify the stance as opposing the target. In the future, the model architecture needs to be further designed to enhance its ability to be guided by and perceive the targets more effectively. As shown in Figure 6, the SITPCL model in this example failed to understand the context of the interaction between users. Although User 2's last sentence alone does not clearly indicate their stance, it becomes evident when considering their previous statements. In future work, pretraining tasks could be designed to enable the model to better comprehend multi-turn conversations involving multiple participants. Additionally, the model could be improved to track each user's current stance, which could then be used to assist in future judgments when the same user appears again. As shown in Figure 7, from the overall context of the conversation, it is clear that

Domain	Query Keywords
Cultural and Education	ACG, Chinese Traditional Culture , Teacher, College Entrance Examination, Tutoring
Policy	Epidemic Prevention, Fireworks Regulations, Family Planning, Civil Service Examination
Science and Technology	Artificial Intelligence, Electric Vehicles, Smartphone, Autonomous Driving Technology
Healthcare	Centralized Medical Procurement, COVID-19 Vaccine, Traditional Chinese Medicine
Entertainment and Consumption	E-commerce, Live Streaming, Mobile Application, Housing Prices, Star Chaser
Social Issues	LGBTQ, Surrogacy, XYY Syndrome, Stray Dog, Bride-price, Elderly care, Employment

Table 7: The domains used in our dataset and the selected query keywords for each domain.

You are an excellent social media stance detection expert. You are given a multi-turn multi-party conversation and a target. Please combine the history utterances ,speaker information and the reply relationship of the utterances to comprehensively judge the stance of the current utterance towards the given target. The stance should be marked as one of the three labels: favor, against, or neutral. Neutral stance includes two cases: 1) The utterance explicitly or implicitly talks about the target, but neither favor nor against; 2) The utterance is completely unrelated to the target.

The input format of the target is a text string.
The input format of each utterance in the conversation is:
user id: utterance content [utterance id, reply = the id of the replied utterance].

The output format is json format: the prediction of each utterance contains two keys: "utterance id" and "stance label".

The input target is:___
The input conversation is:____

Figure 4: An example of the prompt.

User 3 is merely providing an objective analysis in response to User 2's question and is not expressing their own subjective stance. In the future, the model could be enhanced by adding auxiliary tasks for analyzing user dialogue intentions, which would help the model better understand the user's intentions and prevent it from determining the user's stance based solely on verbal cues. As shown in Figure 8 , the SITPCL model failed to understand that User 3 was using sarcasm to oppose User 2's viewpoint, and instead interpreted it as supporting User 2's viewpoint, thus classifying it as an opposing stance. Future work should focus on enhancing the model's ability to recognize sarcasm, metaphors, and other implicit expressions.

Target	Parents should check their children's homework.
Target type	Claim
Conversation	User 1: How shameless! What the teacher should do, you're asking parents to do. User 2: How shameless! You gave birth to the child, but now you want the teacher to raise them.
Gold label	Favor
SITPCL Prediction	Against
Error type	Failure to Understand the target

Figure 5: Error of failure to understand the target.

Target	Disgust towards "super male" gene babies.
Target type	Claim
Conversation	User 1: It's the couple's freedom to have a "super male" baby, and outsiders can't interfere. But if something goes wrong after the baby is born, who's going to take responsibility? The family? The school? Society? Four months of consultation, diagnosed at week 26, meaning the problem was noticed at week 10. Is this really worth it? User 2: There are plenty of genetically normal children who are poorly educated and end up committing crimes. It's unreasonable, unscientific, and unfair to label children with "abnormal" genes as inherently bad. User 3: Well said. The term "normal" is also a human-made definition. User 2: Exactly, humans are just herd animals.
Gold label	Against
SITPCL Prediction	Favor
Error type	Failure to Consider the Context of User Interaction

Figure 6: Error of failure to consider the context of user interaction.

Target	Disgust towards "super male" gene babies
Target type	Claim
Conversation	User 1: It's the couple's freedom to have a "super male" baby, and outsiders can't interfere. But if something goes wrong after the baby is born, who's going to take responsibility? The family? The school? Society? Four months of consultation, diagnosed at week 26, meaning the problem was noticed at week 10. Is this really worth it? User 2: I don't understand why they would have the baby. Isn't the purpose of prenatal testing to ensure a healthy child? User 3: Some people just love their children, regardless of whether they're healthy or not. Plus, wasn't there a study showing that a "super male" baby releases signals in the womb, and this makes the mother's love for them grow infinitely?
Gold label	Neutral
SITPCL Prediction	Against
Error type	Over-Reasoning

Figure 7: Error of over-reasoning.

Target	Ask for the older child's opinion before having a second child.
Target type	Claim
Conversation	User 1: When you want to have a second child, you should definitely ask the older child's opinion. User 2: Reproductive rights belong to the parents, no need for the child's opinion. Besides, a child's thoughts aren't mature enough to be taken seriously. User 3: Then just keep having children and ignore the first one. Wait until the first one is in their twenties, and then have the first one take care of them.
Gold label	Favor
SITPCL Prediction	Against
Error type	Failure to Recognize Sarcasm

Figure 8: Error of failure to recognize sarcasm.