RAR: A Refined-Attitude Reasoning Framework on LLMs for Zero-Shot Stance Detection

Anonymous ACL submission

Abstract

In the context of the rapid evolution of social media, zero-shot stance detection emerges as a significant task within natural language processing, particularly in effectively identifying text attitudes towards given targets using Large Language Models (LLMs). However, existing methodologies inadequately leverage LLMs in terms of complex reasoning and contextual comprehension, often presenting cumbersome utilization methods. This paper introduces a novel framework known as RAR (Refined-Attitude Reasoning), which combines example reasoning with attitude refinement techniques to significantly enhance the performance of LLMs in stance detection tasks. Specifically, we first construct and optimize a repository of examples for effective in-context learning. Secondly, we refine the expression of attitudes through explicit reasoning processes to extract a broader range of attitude keywords. Lastly, we implement a multi-check mechanism to ensure the accuracy and rationality of generated results, necessitating only single API call during reasoning. Experiments demonstrate that our RAR framework surpasses other baseline methods in overall F1 score on the zero-shot stance detection dataset VAST, underscoring the importance of fine-grained attitude refinement and in-context learning in enhancing model performance.

1 Introduction

Stance detection is a natural language processing (NLP) task that automatically identifies attitudes towards specific targets within a given text (Du et al., 2017). For instance, "Profiling is the primary tool of crime prevention which is actually what the police are supposed to do - prevent crime." The stance of this sentence is against crime. Stance detection plays a crucial role in processing substantial textual information on the internet and social media, including issues related to public opinion (Lai et al., 2020), political elections (Müller et al., 2022), and social media ideologies (Kato et al., 2024).

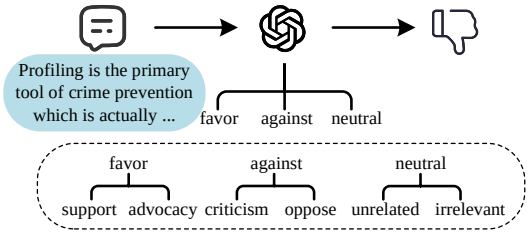

Figure 1: By refining the attitude words in the prompt, LLMs can better identify the stance of the text towards a given target.

Traditional stance detection models are designed for singular targets, with text in datasets focused on the same topic. To enhance the practicality of stance detection models, tasks such as cross-domain (Wei and Mao, 2019), few-shot, and zero-shot stance detection have emerged (Allaway and Mckeown, 2020). Few-shot and zero-shot stance detection eliminate the need to collect stance datasets for all topics, allowing the model to apply its training to unseen targets during inference. Fine-tuning models based on BERT (Liang et al., 2022a; Lin et al., 2021) and LLM-based models (Zhang et al., 2022a; Li et al., 2023) are commonly employed methodologies for zero-shot stance detection tasks. Notably, LLM-based models, endowed with rich background knowledge and robust logical comprehension capabilities, achieve impressive results in zero-shot stance detection without requiring training. Additionally, LLMs demonstrate substantial content support for the interpretability of stance reasoning. Their performance continues to improve through methods such as chain-of-thought (Zhang et al., 2023) and multi-agent (Lan et al., 2024; Wang et al., 2024) approaches.

Despite efforts to leverage chain-of-thought reasoning, reasoning paradigms, or domain experts

without fine-tuning the model, existing methods have failed to effectively stimulate deeper reasoning within LLMs using simple querying techniques, thereby impacting stance prediction accuracy. More complex approaches involving reasoning paradigms or domain experts rely on constructs such as paradigms or roles, leading to a lack of clarity during the reasoning process and often necessitating multiple calls to LLM API.

To address the aforementioned challenges, we propose the RAR framework, an advanced framework for zero-shot stance detection leveraging large models. First, we construct the case library to facilitate efficient in-context learning. Second, through an explicit example reasoning process, we refine the expression of stances and extract a broader range of attitude words. Finally, we implement a multi-check mechanism in generating new texts and determining target stances to ensure the accuracy and validity of results. The innovation of this method lies in the comprehensive utilization of sample features and semantic similarity, providing a more nuanced basis for stance detection while offering deeper insights into LLMs.

Our contributions can be summarized as follows:

- We propose a novel stance detection framework, RAR, which enhances the contextual assistance for stance detection through explicit reasoning explanations, resulting in improved performance.
- We integrate fine-grained attitudes with in-context learning, thereby enhancing the efficacy of LLMs in zero-shot stance detection tasks.
- We conduct experiments on the widely used zero-shot stance detection dataset VAST, surpassing the overall performance achieved by other methods.

2 Related Work

2.1 LLM-based Stance Detection

With the rapid expansion of social media content, stance detection has emerged as an important research area in natural language processing (NLP), aiming to ascertain text stances (favor, against, or neutral) towards specific targets. Techniques based on deep learning have gradually supplanted traditional rule-based and machine learning methods. However, existing deep neural network models (Sun et al., 2023; Lin et al., 2021; Liang et al., 2022b) encounter challenges related to insufficient labeled data and the black-box nature of models that are difficult to explain.

With the emergence of large language models, their advanced logical reasoning capabilities and extensive background knowledge have been directly applied to stance detection tasks. Zhang et al. (2022a) investigated the advancement of stance detection technology following the launch of ChatGPT. ChatGPT provides more interpretable predictions compared to conventional models, offering a novel perspective for stance detection research. Building on this foundation, Zhang et al. (2023) further examined the efficacy of utilizing the Chain-of-Thought (CoT) methodology for stance detection, revealing that CoT exhibits greater accuracy and efficiency than traditional methods. Hu et al. (2024) proposed a "Ladder-of-Thought" (LoT) framework that extends CoT applications, aiming to infuse external knowledge into smaller language models, thereby augmenting the model's reasoning capabilities. Through two-phase progressive optimization, LoT effectively enhances predictive precision while preserving computational efficiency. Li et al. (2023) approached the issue of stance detection from a knowledge enhancement perspective, presenting the Knowledge-Augmented Stance Detection (KASD) framework, which integrates episodic knowledge and discourse knowledge while effectively extracting background knowledge using LLM to bolster stance detection. Ding et al. (2024) introduced the EDDA framework, employing an encode-decode structure for data augmentation and addressing the logic relationship deficiencies within target enhancement and text enhancement. This method significantly improves prediction accuracy and diversity by enabling the generation of semantically relevant expressions. As the application of language models deepens, Lan et al. (2024) designed the Collaborative Role Injection LLM Agent (COLA) framework, aimed at enhancing the debate phase through multi-dimensional text analysis and reasoning, addressing multi-faceted knowledge and implicit viewpoint reasoning challenges in stance detection.

Confronting the rapidly evolving social media environment, LLM-based stance detection technology is continuously evolving, offering new tools and methodologies for understanding public opinion.

2.2 In-Context Learning

In-Context Learning (ICL) is a learning methodology that employs a minimal number of labeled samples and task-related instructions. Since the inception of this concept with GPT-3, ICL has gradually emerged as a pivotal approach in the applications of LLMs (Xie et al., 2021). The core idea of ICL is to guide the model in generating corresponding outputs by designing prompt templates that integrate labeled samples with new testing data. Unlike traditional fine-tuning methods that entail gradient updates, ICL directly generates prompts by concatenating examples, allowing predictions for new samples. This strategy significantly outperforms zero-shot learning, proposing a fresh approach in few-shot scenarios.

In practical implementation, the selection of examples is crucial to the effectiveness of ICL. Current research primarily employs heuristic and LLM-based methods to optimize the selection of demonstrations. Researchers using heuristic methods have discovered that semantically similar examples can significantly enhance output quality (Liu et al., 2022), for example, by employing K-nearest neighbor algorithms to select samples akin to the test samples and improve the accuracy of generated outcomes. Moreover, diversity selection strategies have also been suggested to encompass more potential output scenarios, thereby enhancing model generalizability in novel contexts (Levy et al., 2023).

Conversely, LLM-based methods emphasize the automatic generation of examples through the model's inherent capabilities. Researchers enhance the selection of demonstrations by evaluating the output probabilities of each example concerning input data, thus reducing the complexity of manual selections (Rubin et al., 2022). Related works have explored the notion of viewing example selection as a sequential decision-making problem and optimizing demonstration combinations through reinforcement learning methods (Zhang et al., 2022b). Self-generated examples also show a degree of potential to replace external sample libraries (Kim et al., 2022).

3 Methods

In stance detection, the task entails determining the stance of a given document towards a specified target. The dataset is defined as $D = \{d_i, t_i, y_i\}$, consisting of n instances. Each instance includes a document d_i and a target t_i. The task is to detect the stance y_i, which encompasses pro, con, or neutral.

As illustrated in the figure, our approach comprises three phases, with the first two phases focusing on constructing the example library and refining attitudes. The third phase involves stance detection for new text-target pairs.

3.1 Case Library Construction and Retrieval

Case Library. Typically, training datasets are used for fine-tuning models. However, they can also serve as superior examples in-context learning (Shum et al., 2023). Utilizing the training set, we construct an case library $C = \{c_1, c_2, ..., c_n\}$, each containing full stance detection information,

$$c_i =< d_i, t_i, y_i >, \quad (1)$$

where d_i represents the i-th document, t_i denotes the i-th target, and y_i indicates the stance of d_i towards t_i.

Representation and Retrieval. Recent research suggests that semantically similar content significantly contributes to the efficacy of in-context learning (Liu et al., 2022). To retrieve semantically similar examples as demonstrations, we compute a mixed weighted representation for each text-target pair,

$$f(d, t) = \boldsymbol{w_d} \cdot Enc(d) + \boldsymbol{w_t} \cdot Enc(t), \quad (2)$$

where $f(\cdot)$ is the mixed weighted representation of the current document and target, $\boldsymbol{w_d}$ is the weight corresponding to the document, $\boldsymbol{w_t}$ is the weight corresponding to the target, and $Enc(\cdot)$ is a sentence encoder. Studies indicate that many stance detection models are insensitive to targets (Li and Scarton, 2024), thus we consider both documents and targets together (default ratio of 7:3). The stance reasoning content regarding similar documents and targets aids in conveying information during in-context learning.

In the reasoning phase, when presented with a new text-target pair d and t, we select the top n most similar cases as contextual demonstrations,

$$Demo = \underset{c_i \in C, n}{top\text{-}n}\ Sim(f(d, t), f(d_i, t_i)), \quad (3)$$

where $Demo \subset C$, encompasses n required examples, $top\text{-}n$ refers to the n examples yielding the highest computed values, and $Sim(\cdot)$ is a similarity calculation. It is noteworthy that this retrieval step does not depend on any LLM-generated content.

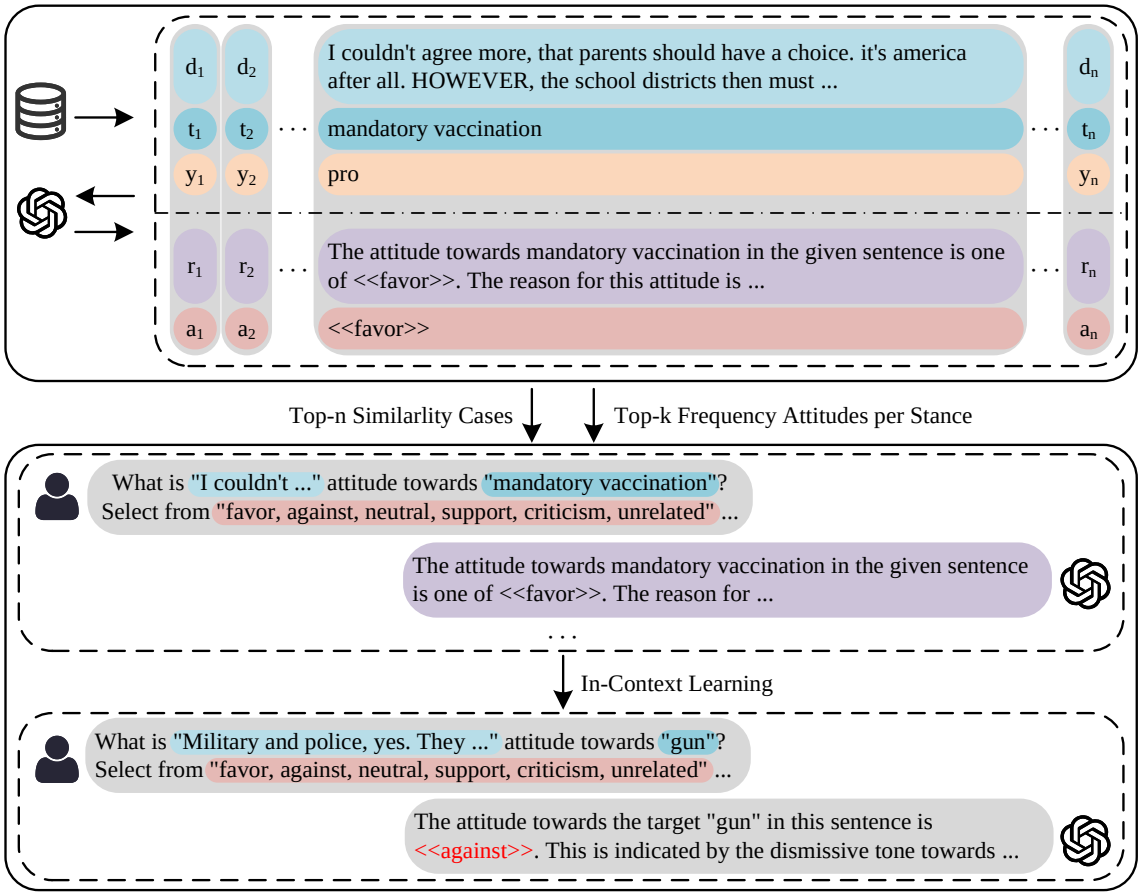

Figure 2: The figure illustrates the methodology for stance detection. Firstly, a case library $C = \{c_1, c_2, \ldots, c_n\}$ is constructed using the training dataset. Mixed weighted representations are computed to retrieve. Secondly, each case c_i undergoes stance reasoning, resulting in an extended case $\tilde{c}_i =< d_i, t_i, y_i, r_i, a_i >$. Lastly, new documents d and targets t are combined with the refined attitude words to generate stance detection results.

3.2 Case Reasoning and Attitude Refinement

Case Reasoning. To convey the principles of stance detection during in-context learning, we explicitly utilize the reasoning process of case content as guidance. Thus, we perform stance reasoning on $c_i \in C$. The prompt template is as follows:

> The attitude of the sentence: "%s" to the target "%s" is "%s". What is the reason for this attitude? Think about the question from a more flexible perspective and give a reasonable answer that fits the attitude given. Please mark the attitude with special symbols as «attitude».

Therefore, we extend c_i as follows:

$$\tilde{c}_i =< d_i, t_i, y_i, r_i, a_i >, \quad (4)$$

where r_i is the reasoning rationale for the ith example, and a_i is the indicated attitude word in special symbols. Consequently, the original case library is expanded to $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_n\}$.

Attitude Refinement. By reasoning on the contents of the example library, we obtain a series of keywords expressing attitudes in the text. Although we used standard terms like favor, against, and neutral in the prompt, the words we obtained for expressing attitudes extend beyond these categories, as illustrated below. We assert that apart from the three basic attitude representation terms, the additional vocabulary reflects LLMs' more nuanced judgments of text attitudes and can be incorporated into the prompt as optional enhancements for reasoning.We conduct a count of each attitude keyword under every category.

$$Count(a_j^y) = \sum_{\tilde{c}_i \in \tilde{C}} (\delta(y_i = y) \cdot \delta(a_i = a_j)), \quad (5)$$

where $\delta(\cdot)$ taking the value 1 when the condition is true and 0 otherwise.

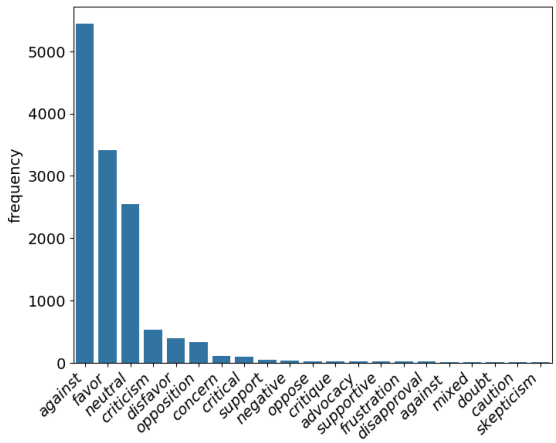

Figure 3: The distribution of attitude words in the reasoning process obtained from the stance of the text-target pair in the case library. When LLMs give reasons, they give a deeper analysis of attitudes.

We select the top-k keywords from each category as optimized attitude words.

$$Attitudes^y = \underset{\tilde{c}_i \in \tilde{C}, k}{\text{top-}k\ Count}(a_i^y), \tag{6}$$

where $Attitudes^y$ refers to chosen attitude words for prompt. Here, we implement certain modifications by removing derived and semantically ambiguous words, such as disfavor being derived from favor, and concerning or mixed attitudes being challenging for stance determination.

3.3 Generation and Verification

Generation. By this stage, we have amassed all necessary content for in-context learning, comprising the n documents d and targets t from $Demo$, and the refined attitude words selected from $Attitudes$, with $3k$ terms possessing k in each category. We utilize the following prompt template:

> What is the attitude of the sentence: "%s" to the target "%s"? select from "%s". Please mark the answer with special symbols as «attitude».

Subsequently, the new document d and new target t are contated to prompt.

$$r, a = LLM(Prompt|Attitudes|Demo|d|t), \tag{7}$$

$$\hat{y} = \begin{cases} y & a \in Attitude^y \end{cases}, \tag{8}$$

where r is the reasoning by LLM, a is the attitude word from r. $\hat{y}$ can be extracted from a.

Verification. Throughout our experimental processes, we observed that LLMs occasionally return contents lacking special markers or erroneously apply markers to the word attitude, or introduce spelling errors. Although we found that the probability of error is less than 1%, We established two types of verification methodologies.

Weak Verification: We perform a format check on the returned content, necessitating the inclusion of << and >>, while forbidding the content from being solely attitude. Strong Verification: Building upon weak verification, we require the content within << and >> to conform to a specific lexical range.

If verification fails, the inquiry is resent. Through our multitude of experiments, it was noted that LLM responses always carry a minimal probability of error. Implementing verification mechanisms mitigates this risk. For case reasoning, we employed weak verification to capture a broader range of LLM-expressed attitude vocabulary. For generation reasoning, strong verification is leveraged to ensure a reasonable reply is obtained.

4 Experiments

4.1 Dataset

To evaluate our approach, we conducted experiments on the well-utilized zero-shot stance detection dataset VAST (Allaway and Mckeown, 2020). VAST encompasses a multitude of distinct targets, with each data sample comprising a document, a target, and the stance of the document towards the target, classified as "favor," "against," or "neutral."

4.2 Experiment Implementation

This experiment utilizes GPT-3.5 turbo (OpenAI, 2023) as the LLM. We employ AoE (Li and Li, 2024) as the text encoder and cosine to assess the similarity between sentences. For the VAST dataset, we use the average F1 scores for each of the three label categories alongside the macro F1 score.

4.3 Comparison with Baseline Methods

To evaluate the effectiveness of our proposed RAR framework, we compare it against a range of baseline methods, primarily encompassing fine-tuning methods and LLM-based methods.

Fine-Tuning Methods. **BERT** (Kenton and Toutanova, 2019) fine-tunes the stance detection task by simply adding an output layer. **BertGCN**

	Methods	VAST			
		Pro	Con	Neutral	F1
Fine-Tuning	Bert	54.6*	58.4*	85.3*	66.1*
	Bert-GCN	58.3*	60.6*	86.9*	68.6*
	JointCL	64.9*	63.2*	**88.9***	72.3*
LLM-Based	DQA	66.2†	67.5†	61.3†	65.0†
	StSQA	68.5†	66.4†	64.3†	66.4†
	KASD-GPT	-	-	-	67.03$^{\sharp}$
	GPT-EDDA	65.3‡	64.2‡	76‡	68.5‡
	COLA	**73.4**†	**77.2**†	69.6†	73.4†
	RAR(ours)	70.32	73.16	81.2	**74.89**

Table 1: Overall results of our RAR and baseline models in the average F1 scores for each of the three label categories and the Macro-F1(%) score(t-test with p-value < 0.05). Best scores are in bold. The results with † are retrieved from (Lan et al., 2024), ‡ from (Ding et al., 2024), ⋆ from (Liang et al., 2022b), ♯ from (Li et al., 2023).

(Lin et al., 2021) integrates large-scale pretraining with graph learning, handling stance classification by constructing heterogeneous graphs. **JointCL** (Liang et al., 2022b) is a joint contrastive learning framework that learns target-specific representations by constructing prototype graphs, facilitating better generalization of stance features.

LLM-Based Methods. **DQA** (Zhang et al., 2022a) directly queries LLMs using prompts to classify the stance of the text. **StSQA** (Zhang et al., 2023) employs a zero-parameter chain-of-thought (CoT) method by adding the phrase "Let's think step by step." **KASD** (Li et al., 2023) framework introduces external knowledge, distinguishing between episodic knowledge and discourse knowledge. **EDDA** (Ding et al., 2024) framework utilizes an encoder-decoder structure for data augmentation to enhance the logical relationship between the target and the text. **COLA** (Lan et al., 2024) framework designs a three-stage collaborative role injection system for LLM agents specifically for stance detection.

5 Results and Analysis

5.1 Overall Performance

In this study, we evaluated the effectiveness of our proposed RAR framework against various fine-tuning methods and prompt-tuning strategies. The experimental outcomes reveal that the RAR framework demonstrates superior performance in stance detection tasks, particularly in terms of overall F1 score metrics.

From Table 1, it is evident that different models exhibit varying performances across the categories of "pro," "con," and "neutral." Specifically:

- **Fine-tuning Methods**: Models like BERT and its derivatives (BertGCN and JointCL) exhibit higher metrics in the "neutral" category, reflecting their effectiveness in addressing sentence similarity in neutral contexts. However, their classification accuracy in "pro" and "con" scenarios remains generally lacking, with BERT achieving an F1 score of just 66.1, while JointCL only attains an enhancement of 72.3. Comparatively, BERT-GCN employing graph convolution networks shows improvements across categories but retains overall performance below our RAR method.

- **Prompt-tuning Methods**: StSQA enhances overall classification performance via chain-of-thought strategies. Nevertheless, their F1 scores still fail to surpass 70, indicating limitations in prompt-tuning methods when addressing complex stances.

- **Superiority of RAR Framework**: Within our RAR framework, the F1 score reaches 74.89, particularly achieving an accuracy of 73.16 in the "con" category and 70.32 in the "pro" category. This indicates that the integration of fine-grained attitude extraction and in-context learning distinctly enhances model performance. In comparison to the COLA framework, although COLA excels in certain categories (e.g., "con"), it still falls short of the overall F1 score established by RAR.

The analysis of results reveals several key points:

- **Influence of Model Architecture**: The RAR architecture achieves profound comprehen-

sion of targets by integrating in-context learning and example reasoning, thereby improving the accuracy of stance discernment. This flexibility allows our model to better navigate complex sentiments and shifts in stance.

- **Contribution of Fine-Grained Attitudes**: By refining attitude vocabulary, the model can make determinations in more nuanced contexts, overcoming traditional models' limitations when confronting polysemous words—an essential aspect of stance judgments.

5.2 Analysis of Example Quantity n and Refined Attitude Quantity k

In this study, we conducted a systematic analysis of the impact of example quantity n and the quantity of refined attitude words k. The experimental results indicate significant influences of these parameters on the performance of our models.

Impact of the Quantity of Attitude Words k. Fixing the example quantity to $n = 3$, we assessed the performance of the model across various attitude word counts k in relation to "pro," "con," and "neutral" categories. The experiments demonstrate in table 2.

	Attitudes	VAST			
		Pro	**Con**	**Neutral**	**F1**
n=3	**k=1**	68	69.74	72.3	70.01
	k=2	69.3	71.94	77.86	73.03
	k=3	**70.32**	**73.16**	**81.2**	**74.89**
	k=4	68.7	72.25	78.71	73.22

Table 2: Results of our RAR when refined attitude quantity k increases in the average F1 scores for each of the three label categories and the Macro-F1(%) score. Best scores are in bold.

At $k = 1$, the F1 metric stands at 70.01, indicating that a limited selection of attitude words restricts the model's ability to effectively differentiate subtle contexts. With the increase of k, a rising trend in F1 scores is observed. At $k = 2$, the F1 score elevates to 73.03, showcasing that an optimal quantity of attitude words significantly enhances classification performance. At $k = 3$, the F1 score reaches its peak at 74.89, substantiating that this number of attitude words provides the best contextual refinement capability for the model, enabling it to address complex sentiment attitudes better. At $k = 4$, the F1 score slightly declines to 73.22, indicating that an excessive number of attitude words may lead to information redundancy or interference, thereby highlighting the necessity of selecting an appropriate k value.

Impact of Example Quantity n. Fixing $k = 3$, we evaluated the influence of various n values, leading to the results in table 3.

		VAST			
		Pro	**Con**	**Neutral**	**F1**
k=3	**n=0**	52.85	60.65	44.87	52.79
	n=1	67.37	70.88	77.06	71.77
	n=2	69.73	72.73	79.20	73.89
	n=3	**70.32**	73.16	**81.20**	**74.89**
	n=4	69.55	73.27	79.82	74.21
	n=5	69.88	**73.61**	81.04	74.84

Table 3: Results of our RAR when example quantity n increases in the average F1 scores for each of the three label categories and the Macro-F1(%) score. Best scores are in bold.

At $n = 0$, the F1 metric is merely 52.79, significantly lower than the DQA method. This reveals that the absence of effective examples severely impairs model performance, emphasizing the critical role of examples in guiding model comprehension. As n increments to 1 and 2, the F1 scores progressively improve to 71.77 and 73.89, indicating that an increase in the number of examples enhances contextual information, thereby improving reasoning accuracy. At $n = 3$, model performance peaks at 74.89, further validating the effectiveness of example quantity on in-context learning. However, as n increases to 4 and 5, performance improvements are marginal, with F1 scores of 74.21 and 74.84 respectively. This denotes that beyond a certain number of examples, further increasing sample sizes contributes negligibly to performance enhancement, with a potential risk of information overload.

Comprehensive analysis indicates that both example quantity n and the selected refinement attitude word quantity k significantly influence the performance of the RAR framework in stance detection tasks. The optimal configuration is established at $n = 3$ and $k = 3$, where the model can leverage sample contextual information and attitude refinement to enhance overall classification accuracy. Conversely, excessive or insufficient examples and attitude words may yield diminishing model capacity for capturing nuanced context.

5.3 Case Study

> Text: Military and police, yes. They're sworn and trained to uphold the country and the laws of the country. The rest is an ego trip. As for hunters, Ogden Nash said it best, '——————————the hunter with pluck and luck is trying to outwit a duck'.
> Target: gun
> DQA: The attitude of the sentence towards the target "gun" is neutral.
> RAR: The attitude towards the target "gun" in this sentence is «against». This is indicated by the dismissive tone towards the idea of individuals using guns for activities outside of military, police, or hunting purposes. The statement suggests that the use of guns by individuals for reasons other than these specific contexts is portrayed as being driven by egotistical motivations rather than a legitimate or necessary purpose.

The responses from RAR elucidate the framework's efficacy in capturing negative attitudes towards firearms. While DQA arrives at a neutral stance, RAR distinctly reflects contempt towards the notion of individuals utilizing guns outside military, police, or hunting contexts.

Regarding RAR's reasoning characteristics, the discussions are as follows:

- **Contextual Understanding** The RAR framework emphasizes contextual comprehension, which is critical for identifying intricate sentiments and positions. In this case, RAR can identify the distinctions the speaker makes between gun usage contexts (i.e., military and hunting as opposed to personal use), allowing for a more accurate stance classification of «against». This ability enables RAR to excel in addressing sentences with puns or multiple interpretations.

- **Fine-Grained Attitude Recognition** By synthesizing background information, RAR can extract more nuanced attitude markers. In contrast to DQA, which offers a neutral response, the RAR framework accentuates the speaker's disdain, signifying opposition to specific gun uses. This deep exploration of attitude intricacies renders RAR's outputs more compelling and practical, facilitating clearer distinction between positions across various contexts.

6 Conclusion and Future Work

6.1 Conclusion

The RAR framework presented in this paper substantially enhances the performance of LLMs in zero-shot stance detection tasks through the integration of case reasoning and attitude refinement. Our research findings indicate that detailed in-context learning effectively guides models toward profound understanding of the relationships between texts and targets, allowing for more accurate stance detections. Compared to traditional fine-tuning methods, the RAR framework exhibits remarkable advantages for addressing zero-shot and few-shot contexts. Additionally, through the establishment of multiple verification mechanisms, we address reliability issues associated with model outputs. This study carries significant implications for improving social media content analysis and public opinion comprehension.

6.2 Future Work

Despite the commendable performance exhibited by our RAR framework in stance detection, several potential avenues for improvement remain. First, exploring the applicability and efficacy of the RAR framework across diverse stance detection text structures (e.g., longer texts or multi-turn dialogues) could enhance its generalizability. Moreover, incorporating a broader range of background knowledge and contextual information will further elevate the model's reasoning capabilities. Lastly, considering the performance of the RAR model in other NLP classification tasks, such as sentiment analysis, may yield fruitful insights.

Limitations

In this study, while the introduction of fine-grained attitude labels has augmented the model's expressive capabilities, some limitations persist. We utilize a ranking mechanism based on model responses to select fine-grained attitudes, aiming to improve stance judgment accuracy. However, in practical applications, the selection and definition of these labels remain subjective. The definition and choice of fine-grained attitude tags may be context-dependent. Within varied contexts, a single attitude word might elicit different interpretations, indicating that the model might struggle to effectively differentiate these potential ambiguities during inference.

References

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.

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.

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

Kairui Hu, Ming Yan, Wen Haw Chong, Yong Keong Yap, Cuntai Guan, Joey Tianyi Zhou, and Ivor W Tsang. 2024. Ladder-of-thought: Using knowledge as steps to elevate stance detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ken Kato, Annabelle Purnomo, Christopher Cochrane, and Raeid Saqur. 2024. L (u) pin: Llm-based political ideology nowcasting. *arXiv preprint arXiv:2405.07320*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

Xianming Li and Jing Li. 2024. Aoe: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839.

Yue Li and Carolina Scarton. 2024. Can we identify stance without target arguments? a study for rumour stance classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2844–2851.

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

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gnn and bert. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Arthur Müller, Jasmin Riedl, and Wiebke Drews. 2022. Real-time stance detection and issue analysis of the 2021 german federal election campaign on twitter. In *International Conference on Electronic Government*, pages 125–146.

OpenAI. 2023. Gpt-3.5 turbo. https://platform.openai.com/docs/models/gpt-3.5-turbo. Accessed: 2024-10-14.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings*

of the Association for Computational Linguistics: EMNLP 2023, pages 12113–12139.

Yepeng Sun, Jicang Lu, Ling Wang, Shunhang Li, and Ningbo Huang. 2023. Topic-aware contrastive learning and k-nearest neighbor mechanism for stance detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2362–2371.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. Deem: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541.

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*, pages 1173–1176.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022a. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

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. *arXiv preprint arXiv:2304.03087*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148.