Stance Detection with Collaborative Role-Infused LLM-Based Agents

Xiaochong Lan, Chen Gao, Depeng Jin, Yong Li
Department of Electronic Engineering, BNRist, Tsinghua University, China
lanxc22@mails.tsinghua.edu.cn, {chgao96, jindp, liyong07}@tsinghua.edu.cn

Abstract

Stance detection automatically detects the stance in a text towards a target, vital for content analysis in web and social media research. Despite their promising capabilities, LLMs encounter challenges when directly applied to stance detection. First, stance detection demands multi-aspect knowledge, from deciphering event-related terminologies to understanding the expression styles in social media platforms. Second, stance detection requires advanced reasoning to infer authors' implicit viewpoints, as stance are often subtly embedded rather than overtly stated in the text. To address these challenges, we design a three-stage framework COLA (short for **C**ollaborative r**O**le-infused **L**LM-based **A**gents) in which LLMs are designated distinct roles, creating a collaborative system where each role contributes uniquely. Initially, in the multidimensional text analysis stage, we configure the LLMs to act as a linguistic expert, a domain specialist, and a social media veteran to get a multifaceted analysis of texts, thus overcoming the first challenge. Next, in the reasoning-enhanced debating stage, for each potential stance, we designate a specific LLM-based agent to advocate for it, guiding the LLM to detect logical connections between text features and stance, tackling the second challenge. Finally, in the stance conclusion stage, a final decision maker agent consolidates prior insights to determine the stance. Our approach avoids extra annotated data and model training and is highly usable. We achieve state-of-the-art performance across multiple datasets. Ablation studies validate the effectiveness of each design role in handling stance detection. Further experiments have demonstrated the explainability and the versatility of our approach. Our approach excels in usability, accuracy, effectiveness, explainability and versatility, highlighting its value.

1 Introduction

Stance detection is commonly defined as automatically detecting the stance (as *Favor*, *Against*, or *Neutral*) of the text producer towards a target [30, 31, 6]. Stance detection plays a pivotal role in the analysis of large-scale text data on the web and social media platforms [23, 42]. Over the years, numerous methodologies have been proposed for stance detection [24, 3]. However, a persistent challenge lies in the need to train models specifically for the targets of interest. Even with advancements in cross-target stance detection[26] and zero-shot stance detection[4, 25], a suitable training on annotated corpora is often required. Acquiring large-scale labeled datasets is not trivial, which curtails the model's usability.

Recently, large language models (LLMs) have demonstrated remarkable capabilities across various applications [10, 34, 2]. The inherent semantic understanding of these large models presents an exciting opportunity for stance detection. Most LLMs can be easily interacted with by users through zero-shot prompting. This significantly enhances the usability of models. Thus, with their strength and usability, large language models could reshape how we approach stance detection.

Researchers have discerned the transformative potential LLMs bring to stance detection. Some works have proposed simple methods using LLMs for stance detection [51, 52]. Yet, while these works report satisfactory results on specific subset of certain datasets, our rigorous replications indicate these methods often underperform compared to the state-of-the-art non-LLM baselines. This can be attributed to two inherent challenges with stance detection, which can be listed as follows and are further illustrated in Figure 1.

- **First, stance detection demands multi-aspect knowledge.** Sentences may contain elements like domain-specific terms, cultural references, social media linguistic styles, and more. These are not immediately comprehensible to large language models and require specialized parsing to be truly understood.
- **Second, stance detection necessitates advanced reasoning.** Often, authors don't state their stances directly but inadvertently reveal them in various ways, such as through their attitudes towards related topics or events. Stance detection requires reasoning from various textual features to arrive at the correct stance.

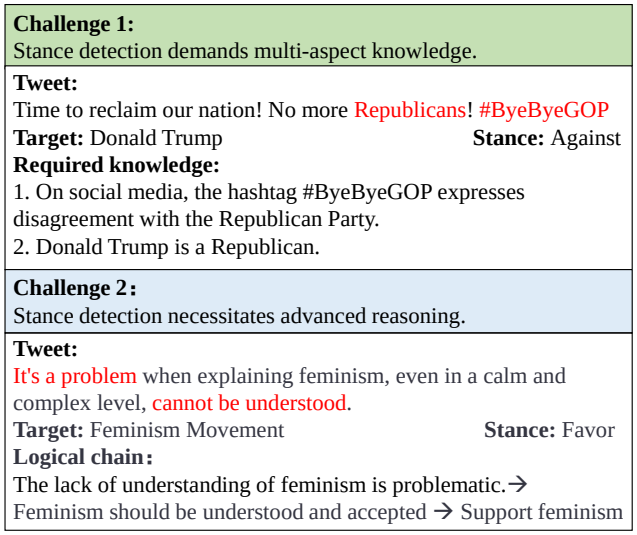

Figure 1: Illustration of the challenges of stance detection.

To address these challenges, we introduce our three-stage framework named COLA(short for **C**ollaborative r**O**le-infused **L**LM-based **A**gents). We design a stance detection system consisting of role-infused LLM-based agents, with each role bearing distinct responsibilities and significance. To counter the first challenge, we initiate a multidimensional text analysis stage. In this stage, LLMs are designated with three roles, named as linguistic expert, domain specialist, and social media veteran, to analyze text from various perspectives. While the linguistic expert delve into syntax, diction, and tenses, the domain specialist elucidate characters, events, and other textual elements. What's more, the social media veteran decode platform-specific terminologies and expression styles. Their combined insights help unearth stance indicators in the text. Addressing the second challenge, we propose a reasoning-enhanced debating stage. Here, we assign advocates for each potential stance category. Drawing evidence from the preceding phase, these advocates present arguments to bolster their respective stances, forcing the LLMs to discern the latent logic connecting textual features and actual stances. Lastly, a stance conclusion stage determines the text's stance, drawing insights both from the text itself and the debates.

Our approach does not necessitate annotated data nor additional model training, hence ensuring high **usability**. Extensive experiments validate our method's superior performance over existing baselines, affirming its **accuracy**[1]. A representative result is that our zero-shot framework achieves a 21.7% absolute improvement compared to the best in-target labeled data dependent baseline on the F_{avg} metric on the CC target of the SEM16 dataset. Ablation studies elucidate the **effectiveness** of each

[1]In this article, unless explicitly stated otherwise, we use *accuracy* to express the overall strong performance of the model on classification tasks, rather than solely referring to the accuracy metric.

module in handling stance detection. Case studies and quantitative experiments substantiate our approach's **explainability**. The powerful performance of our proposed framework in a series of text classification tasks underscores its **versatility**. Our approach stands out for its usability, accuracy, effectiveness, explainability, and versatility, all of which highlight its value.

Our main contributions are summarized as follows:

- We are among the first to delve into harnessing LLMs to bolster stance detection.
- We introduce a approach based on collaborative role-infused LLM-empowered agents, which exhibits outstanding performance on stance detection and achieves high levels of usability and explainability.
- Our proposed three-stage framework—analyst, debater, and summarizer—offers significant potential for a range of text classification tasks, providing a powerful tool for text analysis on web and social media.

The subsequent sections are organized as follows. We first review related works. The wen describe our three-stage framework in detail. We then presents our experiments, providing robust empirical evidence that demonstrated the superiority of our method from multiple perspectives. Lastly, we conclude our work and highlight potential areas for future improvement.

2 Related Work

This section is structured as follows: First, we provide a detailed overview of advancements in stance detection. Next, we introduce recent progress in large language models. Lastly, we focus on reviewing a subset of works closely related to ours, specifically multi LLM-based agents systems.

Stance detection. Stance detection aims to discern the stance of the author towards a particular target from textual content. Typically, stances are categorized into favor, against, neutral. A plethora of algorithms for stance detection have been proposed by researchers, encompassing both feature-based methods [1, 9, 29] and deep learning techniques [21, 47, 28]. These methodologies have enabled in-depth analysis of content on the internet and social media platforms. For example, Jang et al. [23] develop a method to find controversies on social media by generating stance-aware summaries of tweets. Grcar et al. [20] examine the Twitter stance before the Brexit referendum, revealing the pro-Brexit camp's higher influence.

Conventionally, stance detection necessitates training on datasets annotated for the specific target. Such datasets are not trivially obtainable, thereby constraining the usability of many methods. Recognizing this limitation, researchers have ventured into cross-target stance detection, aiming to train classifiers that can adapt to unfamiliar but related targets after being trained on a known target [49, 46, 26]. Recently, there has been an emergence of zero-shot stance detection approaches that automatically detects the stance on unseen tasks [4, 25]. However, all these methods require training on annotated datasets. Unlike these methods, our approach uses pre-trained LLM, removing the need for additional annotated data. Through prompt engineering, we refine these models without extra training, offering a solution with high usability.

Large language models. Large language models (LLMs) represent one of the most significant advancements of artificial intelligence in recent years. With the release of ChatGPT[2] at the end of 2022, LLMs witnessed a meteoric rise in attention, predominantly driven by their outstanding performance. A myriad of LLMs, such as GPT-4 [33], Llama 2 [41], ChatGLM [50], and others, have been introduced at a rapid pace. In conventional NLP tasks, the zero-shot capabilities of these LLMs often rival or even surpass meticulously crafted, domain-specific models [45]. The emergence of robust capabilities, such as planning and reasoning within LLMs, has further enabled their adoption across diverse applications. Some endeavors integrate LLMs with existing tools [37, 38], others explore the potential of LLMs to create new tools [11], and there is a growing trend towards leveraging LLMs for dynamic decision-making, planning, and embodied intelligence [2, 39, 48].

Inherently, the vast knowledge and potent semantic understanding of LLMs offer immense potential in tackling stance detection tasks. Several research initiatives have indeed explored the application of LLMs in stance detection [51, 54, 52]. However, these existing methods often adopt a relatively

[2]chat.openai.com

straightforward approach, neglecting the intrinsic challenges specific to stance detection. As a result, our rigorous replication efforts have frequently found their performance to be subpar in comparison to training dependent baselines. In contrast, our method is specifically tailored to cater to the expert knowledge and intricate reasoning often required for stance detection, consequently achieving commendable results.

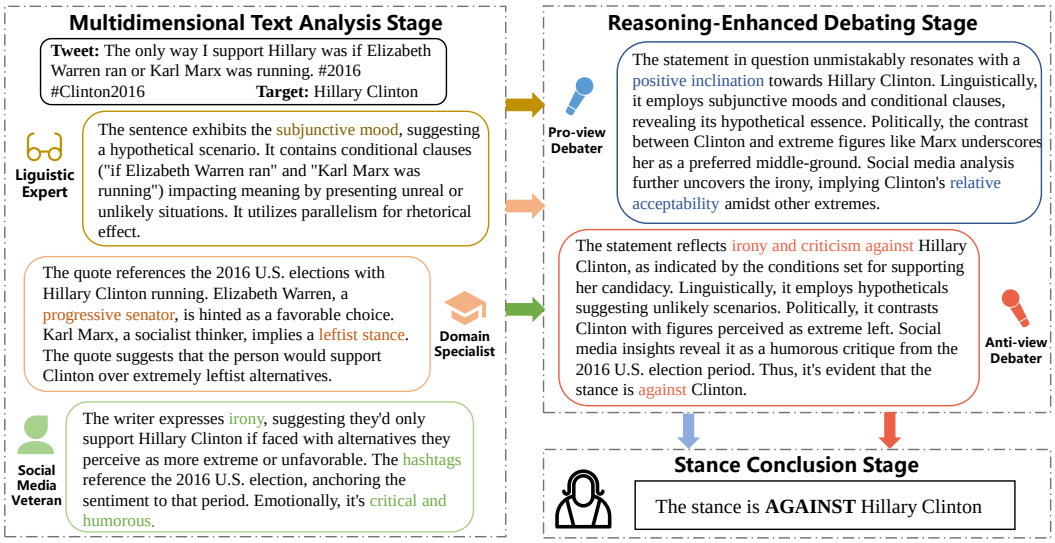

Figure 2: Architecture of our proposed COLA. In the multidimensional text analysis stage, the liguisic expert, the domain specialist and the social media veteran analyze the text from web or social media from various perspectives, providing a holistic understanding. In the reasoning-enhanced debating stage, for each possible stance, a debater defends it, seeking possible logical chains between text features and stance. Finally, in the stance conclusion stage, a final judge determines the stance based on the statements made by all debaters.

Multi LLM-based agents system. Systems comprised of multiple LLM-based agents have demonstrated complex and powerful capabilities not inherent to individual LLM. Leveraging the human-like capacities of LLM, systems formed from multiple LLM-based agents have been applied in both online and offline societal simulations, showcasing credibility at the individual level and emergent social behaviors. For instance, Part et al. [34] construct an AI town with 25 agents, witnessing phenomena such as mayoral elections and party organization. Gao et al. [19] conduct simulations of online social networks with thousands of LLM-based agents, observing group emotional responses and opinion shifts that mirrored real-world trends. What's more, some studies have employed collaborative efforts between LLMs with distinct roles to accomplish tasks. In METAGPT [22], LLM-based agents with different roles collaboratively develop computer software, while DERA [32] uses discussions among various agents to refine medical summary dialogues and care plan generation. Additionally, several efforts have utilized debates between large language model agents to enhance model performance. For example, ChatEval [12] improves text evaluation capabilities through multi-agent debates. Du et al. [18] amplify the factuality and reasoning capacities of large language models by facilitating debates among them.

To the best of our knowledge, our work is the pioneering effort in employing muilti LLM-based agents system for the task of stance detection.

3 Methods

In this section, we desribe our proposed COLA in detail. The architecture of COLA is shown in Figure 2.

3.1 Task Description

In stance detection, the objective is to decide the stance of a given opinionated document with respect to a specified target. Let us define a dataset $D = \{(x_i = (d_i, t_i), y_i)\}_{i=1}^{n}$ consisting of n instances. For each instance, x_i represents a tuple comprising a document d_i and a corresponding target t_i. The task is to detect the stance y_i, which can be one of the following categories: favor, against, or neutral.

3.2 Multidimensional Text Analysis Stage

3.2.1 Challenge:

Stance detection necessitates a profound grasp of multi-aspect knowledge. Sentences on social media that convey the author's stance may be influenced by various linguistic phenomena, such as grammatical structures, tenses, and moods. There is also often an abundance of domain-specific terminologies, including references to characters, political parties, and events, and their relationships with the target. Additionally, unique language features of social media, such as hashtags, come into play. Although large language models have assimilated vast knowledge from their training data, their direct application for stance detection often fails to adequately harness this knowledge, leading to suboptimal results, a fact corroborated by our subsequent experiments.

3.2.2 Approach:

To address this challenge and leverage the rich knowledge encoded within large language models, we designed a multidimensional text analysis stage. During this stage, we introduced three distinct LLM-based agents to parse the text from different perspectives, ensuring a comprehensive understanding of potential elements influencing the author's stance.These agents are the Linguistic Expert, Domain Specialist, and Social Media Veteran. We ask the LLM to behave in the way of the roles through prompting. Specifically, the inputs and outputs of the role-infused agents in this stage are as follows.

Input: A text with a stance.

Output: The individual analyses of the text by the linguistic expert, the domain specialist, and the social media veteran.

The detailed configurations of agents are as follows.

Liguistic Expert. This Agent is tasked with dissecting the text from a linguistic standpoint, exploring factors including but not limited to:

- *Grammatical structure.* The arrangement and relationship of words in a sentence, which determines how different elements combine to produce specific meanings.
- *Tense and inflection.* Tense identifies when an action occurs, influencing the stance's immediacy or distance. Inflection adjusts word forms, providing clues about the sentence's grammatical and relational context.
- *Rhetorical devices.* These are techniques used to enhance the expressiveness of language. By emphasizing, contrasting, or evoking emotions, they shape the tone and attitude of a statement.
- *Lexical choices.* The selection of particular words or phrases in writing, which can reveal deeper nuances, biases, or viewpoints about a topic.

Domain Specialist. This agent focuses on domain-relevant knowledge, exploring facets such as:

- *Characters.* Key individuals or entities in a text.
- *Events.* Significant occurrences within a text. How they're portrayed can hint at the author's stance on certain issues or topics.
- *Organizations.* Established groups mentioned. Their depiction can showcase the author's feelings towards certain societal structures or institutions.
- *Parties.* Political groups with distinct ideologies. A text's treatment of these can provide insights into the author's political leanings or criticisms.

- *Religions.* Specific faiths or spiritual beliefs. How they are referenced might shed light on the author's personal beliefs or societal observations.

Social Media Veteran. This agent delves into the nuances of social media expression, focusing on aspects like:

- *Hashtags.* Specific labels used on social media platforms, assisting in categorizing posts or emphasizing specific themes, making content easily discoverable.
- *Internet slang and colloquialisms.* These refer to informal terms and expressions often used in online communities. Their usage can introduce nuances, cultural contexts, or specific attitudes, making them significant indicators of the underlying stance in a statement.
- *Emotional tone.* This captures the sentiment inherent in a piece of writing, revealing the author's feelings, whether positive, negative, or neutral, about a particular subject.

3.3 Reasoning-Enhanced Debating Stage

3.3.1 Challenge:

The task of stance detection requires sophisticated reasoning. Authors often do not explicitly state their positions in a text. Instead, their stance may be implied through their sentiment towards certain entities or by mechanisms like comparison and contrast. Identifying these implicit stances requires detailed reasoning. Although large-scale language models possess some reasoning capabilities, their performance can be suboptimal in intricate reasoning tasks without proper guidance, which can affect the quality of stance detection results.

3.3.2 Approach:

Drawing inspiration from recent works that leverage discussions or debates among large models to enhance their performance [18, 12, 27], especially in reasoning tasks, we introduce a reasoning-enhanced debating stage. In this stage, for every potential stance, an agent is designated. This agent seeks evidence from expert analyses of the text and advocates for its designated stance. Specifically, the inputs and outputs of agents in this stage are as follows.

Input: A text with a stance. The analyses of the text by the linguistic expert, the domain specialist, and the social media veteran.

Output: The debate from each agent for the stance they support, including the evidence it chooses and its logical chain.

In our framework, we only engage in a single round of debate, reserving multi-round debates for future exploration. Directing agents to search for evidence and defend their aligned stances compels the large language model to establish logical connections between discerned textual features (as well as their multifaceted interpretations) and the actual underlying stance of the text. By having multiple agents debate in favor of different stances, the system encourages the large model's divergent thinking. This generates a plethora of potential text stance interpretations, ensuring that the probable correct interpretation has a higher likelihood of being produced by the system. These outputs subsequently feed into the stance conclusion stage, which renders a final, judicious judgment.

3.4 Stance Conclusion Stage

To infer a conclusive stance from diverse agent debates, we introduce the stance conclusion stage. In this stage, a judger agent determines the final stance of a text based on both the text itself and the arguments presented by debater agents. The process is delineated as:

Input: A text with an embedded stance. Arguments from each agent, including evidence and their logical reasoning.

Output: The identified stance of the text.

The judger agent evaluates the text's inherent qualities, the evidence provided by debaters, and their logical frameworks to reach an informed decision.

Dataset	Target	Pro	Con	Neutral	Unrelated
SEM16	DT	148	299	260	-
	HC	163	565	256	-
	FM	268	511	170	-
	LA	167	544	222	-
	A	124	464	145	-
	CC	335	26	203	-
WT-WT	CA	2469	518	5520	3115
	CE	773	253	947	554
	AC	970	1969	3098	5007
	AH	1038	1106	2804	2949
VAST	-	6952	7297	4296	-

Table 1: Statistics of our utilized datasets.

After going through the three stages mentioned above, we have effectively extracted the underlying stance towards the given target from the text.

4 Experimental Setup

In this section, we describe the specific setup of our experiments.

4.1 Datasets

We conduct experiments on three distinct datasets:

SEM16 [30]. This dataset features six specific targets from diverse domains, namely *Donald Trump* (DT), *Hillary Clinton* (HC), *Feminist Movement* (FM), *Legalization of Abortion* (LA), *Atheism* (A), and *Climate Change is Real Concern* (CC). Each instance is classified into one of three stance categories: *Favor*, *Against*, or *None*.

WT-WT [15]. Specializing in discourse about mergers and acquisitions between companies, this dataset comprises four targets: *CVS_AET* (CA), *CI_ESRX* (CE), *ANTM_CI* (AC), *and AET_HUM* (AH). Stance labels include *Support*, *Refute*, *Comment* (Neutral), or *Unrelated*.

VAST [4]. This dataset is characterized by its large number of varying targets. An instance in VAST includes a sentence, a target, and a stance, which may be *Pro*, *Con*, or *Neutral*.

The statistics of our utilized datasets are shown in Table 1. Due to the zero-shot nature of our method, we do not split the dataset into training, development, and testing sets, but instead conduct experiments on the entire dataset. For zero-shot stance detection approaches, we evaluate their performance across all three datasets. However, for in-target stance detection methods, we assess their performance on SEM16 and WT-WT, because the targets within the VAST dataset are mainly few-shot or zero-shot. The datasets contain no personally identifiable information, but may contain offensive content because the text has a clear stance on topics such as religion, politics, climate, etc. We strictly adhere to the requirements of the respective licenses when using all datasets mentioned in the paper.

4.2 Experimental Implementation

4.2.1 Implementation of COLA

In our study, we employ the GPT-3.5 Turbo model, provided by OpenAI, as our backbone. We opt for GPT-3.5 Turbo primarily due to its superior performance, cost-effectiveness, and the ease of interaction offered via the OpenAI API. These attributes not only facilitate efficient research but also ensure the usability of our methodology for future applications. By utilizing the system instruction feature available through the OpenAI API, we instruct the model to act as various agent roles, feeding text inputs via prompts and collecting textual outputs from the model. To maximize reproducibility,

we set the temperature parameter to 0. The reported results are the average of 5 repeated runs to ensure statistical reliability.

4.2.2 Evaluation Metric

For SEM16 dataset, following Allaway et al. [5], we calculate F_{avg}, which represents the average of F1 scores for *Favor* and *Against*. For the WT-WT dataset, we follow the guidelines set by Conforti et al. [15] and calculate the Macro-F1 score for each target. For the VAST dataset, we adopt the method from Allaway et al. [4] and compute the F1 score for *Pro*, *Con* and the Macro-F1 score to assess model performance.

4.3 Comparison Methods

We compare COLA with state-of-the-art (SOTA) methods in stance detection. We conduct comparisons with methods for two tasks: zero-shot stance detection and in-target stance detection.

We compare our method with various zero-shot stance detection methods. This includes adversarial learning method: TOAD [5], contrastive learning methods: PT-HCL [25], Bert-based techniques: TGA-Net [25] and Bert-GCN [28]. We also include two baselines based on large language models: GPT-3.5 Turbo and GPT-3.5 Turbo+Chain-of-thought(COT), both of which can be considered zero-shots, implemented in strict accordance with Zhang et al. [51] and Zhang et al. [52], respectively.

To further verify the performance of our model, we compare our model to in-target stance detection methods. Such methods undergo extensive training on datasets for a given target and are then evaluated on the test set of the same target. In contrast, our method remains strictly zero-shot, with **no fine-tuning** applied to our backbone model. We compare our approach with various in-target stance detection baselines, including RNN-based methods: BiCond [8], and ATT-LSTM [44]; Attention-based method: CrossNet [49]; Bert-based method: BERT [17]; and Graph-based methods: ASGCN [53] and TPDG [26].

For non-LLM approaches, we retrieve results from existing literature for a comprehensive comparison [15, 4, 5, 28, 26, 25].

5 Experimental Results

In this section, we aim to answer the following research questions (RQs) with the help of experimental results:

RQ1: How is the performance of COLA compare with state-of-the-art stance detection models? **(Accuracy)**

RQ2: Is every component in our model effective and contributory to performance enhancement? **(Effectiveness)**

RQ3: Can our model explain the rationale and logic behind its stance determinations? **(Explainability)**

RQ4: Is our framework adaptable for other text classification tasks related to web and social media content analysis? **(Versatility)**

5.1 Overall Performance (RQ1)

In Table 2, we present the zero-shot stance detection performance of COLA across three datasets in comparison to baseline methods. Furthermore, Table 3 showcases the results of both our zero-shot COLA and the in-target labeled data dependent baselines on the SEM16 and WT-WT datasets for the in-target stance detection task. Overall results have demonstrated the strong performance of our approach. Specifically, the key findings are enumerated below.

- Our method outperforms the state-of-the-art zero-shot stance detection approaches across the majority of metrics. On most metrics across three datasets, our model demonstrates statistically significant improvements over the best baseline. For the CC and LA targets in the SEM16 dataset, our approach achieves substantial gains over the best baseline, with

Model	SEM16(%)						WT-WT(%)				VAST(%)		
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH	Pro	Con	All
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	55.3	57.7	58.6	61.7	42.6	36.7	41.0
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	65.7	63.5	69.9	68.7	55.4	58.5	66.6
BERT-GCN	42.3	50.0	44.3	44.2	53.6	35.5	67.8	64.1	70.7	69.2	58.3	60.6	68.6
PT-HCL	50.1	54.5	54.6	50.9	56.5	38.9	73.1	69.2	76.7	76.3	61.7	63.5	71.6
GPT-3.5	69.5	74.0	59.1	52.0	8.1	24.7	65.5	61.1	64.3	66.4	66.2	67.5	65.0
GPT-3.5+COT	69.0	75.5	60.8	55.3	10.3	25.2	66.2	63.3	65.5	66.7	68.5	66.4	66.4
COLA(ours)	**71.2**	**75.9**	**69.1***	**71.0***	**62.3***	**64.0***	**80.8***	**76.2***	**83.0***	**78.9**	**73.4***	**77.2***	**73.4***

Table 2: Comparison of COLA and baselines in zero-shot stance detection task. Best scores are in bold. * denotes COLA improves the best baseline at $p < 0.05$ with paired t-test.

Category	Model	SEM16(%)						WT-WT(%)			
		DT	HC	FM	LA	A	CC	CA	CE	AC	AH
In-target Labeled Data Dependent Methods	BiCond	59.0	56.1	52.9	61.2	55.3	35.6	71.1	72.3	72.6	72.0
	BERT	57.9	61.3	59.0	63.1	60.7	38.8	73.6	73.2	76.6	75.5
	CrossNet	60.2	60.2	55.7	61.3	56.4	40.1	71.7	71.2	73.8	72.5
	ATT-LSTM	55.3	59.8	55.3	62.6	55.9	39.2	72.0	71.4	74.3	73.5
	ASGCN	58.7	61.0	58.7	63.2	59.5	40.6	72.2	72.9	75.1	74.3
	TPDG	63.0	73.4	67.3	**74.7**	**64.7**	42.3	79.3	**77.6**	81.5	**80.2**
Zero-shot Method	COLA(ours)	**71.2***	**75.9**	**69.1**	71.0	62.3	**64.0***	**80.8**	76.2	**83.0**	78.9

Table 3: Comparison of zero-shot COLA and baselines fully trained on labeled data for the in-target stance detection task. Best scores are in bold. * denotes COLA improves the best baseline at $p < 0.05$ with paired t-test.

absolute increases in F_{avg} of 15.7% and 25.1% respectively. In the WT-WT dataset, our method realizes significant improvements over the best baseline for all targets except for AH. In the VAST dataset, which comprises tens of thousands of instances, our model secures a notable absolute boost of 1.8% in the overall Macro-F1 Score. This attests to the robust zero-shot stance detection capabilities of our approach.

- The zero-shot stance detection performance of our method is closely aligned with that of the state-of-the-art in-target stance detection techniques, even when they are fully trained on corresponding targets. On the SEM16 dataset, our approach significantly outperforms the best baseline, TPDG, on the DT and CC targets, while maintaining comparable performance on other targets. In the WT-WT dataset, our method consistently matches the performance of TPDG across all targets. Remarkably, even though these comparison methods have been extensively trained on their respective targets, our approach still sustains comparable or superior performance, underscoring our method's strong performance.

- Direct application of large language models may yield poor performance, especially on abstract concept targets. In the SEM16 dataset, for the targets A (*Atheism*) and CC (*Climate Change is a Real Concern*), GPT-3.5 achieves only 8.1% and 24.7% in F_{avg} respectively. Even with the enhanced GPT-3.5+COT, the scores are merely 10.3% and 25.2%. Across almost all datasets and metrics, the performance of simply deploying large language models significantly lags behind our proposed method. This underscores the limitations of directly using large language models for stance detection tasks, especially in handling stances towards abstract concept targets, highlighting the necessity and validity of our design.

Model	SEM16(%)					
	DT	HC	FM	LA	A	CC
COLA	71.2	75.9	69.1	71.0	62.3	64.0
w/o Liguisitic Expert	69.1	74.2	67.8	67.2	46.0	62.1
w/o Domain Specialist	70.4	75.0	66.5	60.1	42.4	58.2
w/o Social Media Veteran	67.8	75.5	68.2	64.4	54.6	60.0
w/o Multidimensional Text Analysis Stage	67.4	72.8	65.2	52.2	23.3	55.9
w/o Reasoning-Enhanced Debating Stage	64.7	73.3	64.0	53.8	26.6	49.1

Table 4: Experimental results of ablation study.

5.2 Ablation Study (RQ2)

To investigate the impact of each module in our design, we conduct ablation studies to assess the performance of our framework when each module is removed. The results are shown in Table 4, which demonstrate that every module in our framework contributes to performance enhancement. In the following, we provide a detailed description of the results.

5.2.1 Study on multidimensional text analysis stage.

During the multidimensional text analysis stage, three expert agents from different domains concurrently analyze the text. We individually removed each of these experts to assess the performance of our approach. We also evaluated the performance when all expert analyses are excluded. The results show that the removal of any expert agent results in a certain degree of performance degradation. Moreover, eliminating the entire multidimensional text analysis stage leads to a significant performance drop. The most pronounced performance decline was observed for the A (*Atheism*) target. Removing the Linguistic Expert, Domain Specialist, and Social Media Veteran leads to an F_{avg} decrease to 46.0%, 42.4%, and 54.6%, respectively. What's more, without the multidimensional text analysis stage, the F_{avg} drops to a mere 23.3%. This could be attributed to the complexity of the *Atheism* topic across various domains such as religion and society. These findings underscore the effectiveness of our multidimensional text analysis stage and the design of each agent therein.

5.2.2 Study on reasoning-enhanced debating stage.

In the reasoning-enhanced debating phase, we introduce debates among agents with differing perspectives to augment the reasoning capabilities of our LLM-based system. We remove this stage and let the judger agent directly deduce the text's stance from the expert agents' text analysis, aiming to verify the effectiveness of the debating design. Upon the removal of the debating stage, our method experiences a noticeable performance degradation. The most significant drops are observed for the abstract concept targets A (*Atheism*), CC (*Climate Change is Real Concern*), and LA (Legalization of Abortion), with the absolute F_{avg} declining by 35.7%, 14.9%, and 10.6%, respectively. This indicates that the reasoning-enhanced debating stage offers substantial benefits, especially when dealing with relatively abstract targets. The results validate the effectiveness of the reasoning-enhanced debating stage design.

Case 1 **Tweet:** The ruling by @Scotus is a major setback for @EPA & the environment. #dirtycoal **Target:** Climate Change is Real Concern
Detected Stance: Favor **Explaination:** The use of the hashtag #dirtycoal conveys a negative stance on coal, associated with environmental harm and climate change. Mentioning the Environmental Protection Agency (@EPA) and describing a ruling as an environmental setback highlights a concern for environmental protection, including climate change mitigation. The emotional tone of frustration and negative environmental hashtags further reinforce this pro-environmental perspective.
Case 2 **Tweet:** @GovtsTheProblem This is what I see: Make way 4 ur queen peasants! Don't touch or talk 2 her U filth! #NoHillary2016 #Benghazi **Target:** Hillary Clinton
Detected Stance: Against **Explaination:** The use of derogatory language, the mention of the Benghazi incident in a negative context, and the hashtag #NoHillary-2016 all indicate thje stance against Clinton and a desire for her not to be elected.

Figure 3: Cases of explainations generated by our approach.

Method	SEM16(%)					
	DT	HC	FM	LA	A	CC
GPT-3.5	69.0	75.5	60.8	55.3	10.3	25.2
COLA	**71.2**	75.9	69.1	**71.0**	**62.3**	**64.0**
GPT-3.5+COLA's Explainations	69.4	**77.7**	**70.7**	66.7	61.9	54.5

Table 5: Performance of GPT-3.5 Turbo, COLA and GPT-3.5 Turbo with explainations generated by COLA. Best scores are in bold.

5.3 Study on Explainablity (RQ3)

An explainable artificial intelligence (XAI) is one that offers clear insights or justifications to make its decisions comprehensible [7]. By elucidating its decision-making processes, an XAI augments transparency and reinforces model trustability [16]. Large language models inherently possess the capability to explain their outputs. By prompting them about the rationale behind their decisions, we can obtain explanations for their determinations directly. To delve deeper into the explainablility of our approach, we conduct both case studies and quantitative experiments to verify its ability to generate clear and reasonable explanations.

During the stance conclusion stage, we mandate the judger agent to provide outputs in a JSON format, consisting of two components: the stance and a concise explanation not exceeding 100 tokens. We conduct our experiment on the SEM16 dataset. After closely examining the generated outputs, we find that our model can provide clear explanations for its decisions. In Figure 3, we show two cases to illustrate.

In the first case, the tweet *"The ruling by @Scotus is a major setback for @EPA & the environment. #dirtycoal"* agrees that climate change is a real concern. Our model detects this stance. In its generated explanation, the model discerns the mention of the EPA and the usage of the #dirtycoal tag, indicating an environmental concern. Moreover, the model perceives an emotional tone of frustration, further reflecting a pro-environmental perspective.

In the second case, the tweet *"@GovtsTheProblem This is what I see: Make way 4 ur queen peasants! Don't touch or talk 2 her U filth! #NoHillary2016 #Benghazi"* portrays an opposing stance toward Hillary. Our model rationally explains its judgment from a linguistic perspective (utilization of derogatory language), a domain-specialist perspective (mentioning the Benghazi incident in a negative context), and a social media lens (the hashtag #NoHillary2016). These cases validate the model's proficiency in generating clear and reasonable explanations.

To further validate our model's ability to produce clear and logical explanations, we conduct quantitative experiments. For the SEM16 dataset, we collect explanations (from the second part of the JSON output) related to each instance's stance generated by COLA. These explanations, along with the original text, are fed into the GPT-3.5 Turbo model. We inform the model that these explanations could be used as references for its decisions. As a result, we obtain a new set of judgments from the model. It's evident that the performance of GPT-3.5 Turbo significantly improves by incorporating explanations generated by COLA in addition to the original text, as presented in Table 5. There is a noticeable increase for the A(*Atheism*) and CC(*Climate Change is Real Concern*) targets, with F_{avg} improving by 51.6 and 29.3 points, respectively. For the HC(*Hillary Clinton*) and FM(*Feminist Movement*) targets, the results even exceed that of COLA. This further confirms our model's strong ability in generating clear and logical explanations.

Category	Model	Restaurant14(%)		Laptop14(%)		Restaurant15(%)	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Labeled Data Dependent Methods	DGEDT	**86.3**	80.0	79.8	75.6	84.0	71.0
	dotGCN	86.2	**80.5**	81.0	**78.1**	85.2	72.7
Zero-shot Methods	GPT-3.5-Turbo	74.3	69.6	69.9	61.0	80.4	67.7
	Ours	84.1	77.7	**81.6**	77.0	**85.4**	**74.9**

Table 6: Performance of our framework and baselines on aspect-based sentiment analysis. Best scores are in bold.

Model	Accuracy(%)	F1-Score(%)
Hybrid RCNN	74.8	59.6
GPT-3.5 Turbo	67.6	56.0
Ours	**76.5**	**63.9**

Table 7: Performance of our framework and baselines on persuasion prediction. Best scores are in bold.

5.4 Study on Versatility (RQ4)

Our proposed COLA can be summarized as an Analyst-Debater-Summarizer framework. In this section, we conduct experiments to validate that the Analyst-Debater-Summarizer framework can be applied to other text classification tasks for text analysis on web and social media, not just as an ad-hoc approach for stance detection. We perform experiments on two additional text classification tasks: aspect-based sentiment analysis and persuasion prediction. We select aspect-based sentiment analysis because it demands precise understanding of sentiments tied to specific elements in text, reflecting the detailed analysis capability of our framework. Meanwhile, persuasion prediction is chosen due to its emphasis on detecting underlying intent, highlighting COLA's ability to adeptly handle intricate conversational dynamics commonly seen in web and social media exchanges.

Aspect-based sentiment analysis is to determine the sentiment polarity (*Positive*, *Negative*, or *Neutral*) expressed towards each aspect mentioned in the text [36]. In this task, we modify the debater component in our original framework to engage in sentiment debates instead of stance debates, while keeping other design unchanged. We evaluate our approach's performance on the Restaurant14 and Laptop14 datasets from SemEval14 [36], as well as the Restaurant15 dataset from SemEval15 [35]. We follow Chen et al. [14] and use Accuracy and Macro-F1 score as evaluation metrics. We compare our approach with state-of-the-art models that require training, namely DGEDT [40] and dotGCN [13].

The experimental results are presented in Table 6. It can be observed that our zero-shot method performs comparably to the best baseline models that rely on labeled data. On the Restaurant15 dataset, our approach even outperforms the top baseline. Another crucial finding is that our approach consistently outperforms directly applying GPT-3.5 Turbo while maintaining ease of use.

Following Ziems et al. [54], we define persuasion prediction as determining whether one party in a conversation is persuaded after the conversation ends. In this task, we replace the three experts in our original framework with two experts: a domain expert and a psychologist. They provide detailed analysis of various concepts and nouns in the conversation topic and analyze the psychological changes of the individuals involved. The debaters are modified to argue for whether a participant in the conversation has been persuaded. We use the dataset provided by Wang et al. [43] and follow their evaluation metrics, using Accuracy and Macro-F1.

We compare our approach with Hybrid RCNN [43] and GPT-3.5 Turbo, and the results are presented in Table 7. The experimental results show that our approach achieves better performance compared to the baseline and a significant improvement over GPT-3.5 Turbo.

The Analyst-Debater-Summarizer framework has proven to be highly successful in both aspect-based sentiment analysis and persuasion classification tasks. On a series of tasks, our zero-shot framework performs on par with state-of-the-art baselines that rely on training data and significantly outperforms direct application of GPT-3.5 Turbo. These experiments demonstrate the versatility of our approach.

5.5 Discussions

In the aforementioned experiment, we extensively evaluate the performance of our approach across various dimensions. From the perspective of our method's design rationale, the ablation study confirms that every component in our approach contributes to a performance boost, indicating that the design is free of redundancy and can be considered efficacious. In comparison with existing methods, experimental evidence shows that our approach outperforms all other zero-shot methods on stance detection. Furthermore, its performance is on par with in-target stance detection methods that rely on in-target labeled data, exhibiting impressive accuracy. In addition, for two other text classification tasks related to web and social media content analysis, our method achieves results comparable

to state-of-the-art baselines, underscoring its versatility. From a practical application standpoint, our method does not require additional training for the model. Instead, it can be implemented by interacting with existing large language models through APIs or other means, showcasing its strong usability. The experiments also prove that our framework can provide clear and rational explanations for its decisions, ensuring a high degree of explainability. Such generated explanations can bolster users' trust in our approach and are conducive to further analysis. Given these advantages, our method promises a broad range of applications.

6 Conclusion and Future Work

In this work, we harness the formidable capabilities of LLMs for advanced stance detection. We propose COLA, where multiple LLM-based agents collaborate to reach an conclusion. This method encompasses three stages: the multidimensional text analysis stage, the reasoning-enhanced debating stage, and the stance conclusion stage. Experimental results demonstrate that our approach achieves high accuracy, effectiveness, explainability, and versatility, showcasing its significant applicability.

Our method is not without limitation. Due to the absence of real-time training data for large language models, the performance in analyzing real-time topics might be slightly compromised. For future work, we intend to incorporate a real-time updating knowledge base into the text analysis stage to enhance our framework's capability to analyze texts that include current events. Furthermore, there remains vast potential for exploring its implementation in addressing extensive text analysis tasks on web and social media.

References

[1] Aseel Addawood, Jodi Schneider, and Masooda Bashir. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on Social Media & Society*, pages 1–10, 2017.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

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

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

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

[6] Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144, 2023.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[8] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

[9] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, 2017.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

[12] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[13] Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064, 2022.

[14] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461, 2017.

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

[16] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[19] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

[20] Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. Stance and influence of twitter users regarding the brexit referendum. *Computational social networks*, 4:1–25, 2017.

[21] Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. Detecting stance in czech news commentaries. *ITAT*, 176:180, 2017.

[22] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[23] Myungha Jang and James Allan. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1221–1224, 2018.

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

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

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

[27] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[28] Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 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, 2021.

[29] Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. Stance prediction for russian: data and analysis. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018 6*, pages 176–186. Springer, 2020.

[30] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

[31] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

[32] Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. Dera: enhancing large language model completions with dialog-enabled resolving agents. *arXiv preprint arXiv:2303.17071*, 2023.

[33] OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[34] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[35] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics, 2016.

[36] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[37] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

[38] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[39] N Shinn, F Cassano, B Labash, A Gopinath, K Narasimhan, and S Yao. Reflexion: Language agents with verbal reinforcement learning (arxiv: 2303.11366). arxiv, 2023.

[40] Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6578–6588, 2020.

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

[42] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. A multi-task model for sentiment aided stance detection of climate change tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 854–865, 2023.

[43] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019.

[44] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

[45] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[46] Penghui Wei and Wenji Mao. 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, 2019.

[47] Penghui Wei, Wenji Mao, and Daniel Zeng. A target-guided neural memory model for stance detection in twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[48] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. *arXiv preprint arXiv:2305.10626*, 2023.

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

[50] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

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

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

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

[54] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.