Improving Consistency Identification in Task-oriented Dialogue Through Multi-Agent Collaboration

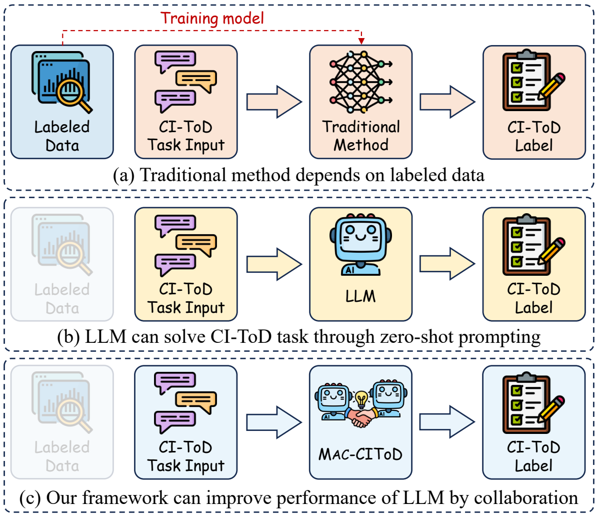
Peng Wang[1,2,5], **Shuo Li**[1], **Ruoxi Zhou**[1], **Qiguang Chen**[3], **Xiao Xu**[3],
Hao Fei[4], **Dagang Li**[5], **Wanxiang Che**[3] and **Libo Qin**[1,2,*]
[1]School of Computer Science and Engineering, Central South University, China
[2] Key Laboratory of Data Intelligence and Advanced Computing in Provincial Universities,
Soochow University, China
[3]Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology, China
[4]School of Computing, National University of Singapore, Singapore
[5]School of Computer Science and Engineering, Macau University of Science and Technology, China
wpengxss@gmail.com, lbqin@csu.edu.cn

Abstract

Consistency identification in task-oriented dialog (CI-ToD) typically consists of three sub-tasks: User Query Inconsistency (QI) identification, Dialogue History Inconsistency (HI) identification, and Knowledge Base Inconsistency (KBI) identification, which aim to determine inconsistent relationships between system response and user query, dialogue history, and knowledge base. Previous approaches focus on the exploration of deep learning models for CI-ToD. While these models achieve remarkable progress, they still rely on large amounts of labeled data, which is hard to achieve in real-world scenarios. Motivated by this, in the paper, we aim to explore large language models for CI-ToD, which do not require any training data. In addition, we further introduce a multi-agent collaboration framework (Mac-CIToD) to model the interaction across three sub-tasks in CI-ToD, including (1) *Full Connection paradigm*, (2) *Cycle Connection paradigm*, and (3) *Central Connection paradigm*, which effectively builds interaction across QI, HI, and KBI. Experiments on the standard benchmark reveal that our framework achieves superior performance. Additionally, we compare Mac-CIToD with the most advanced trained approaches and find that its zero-shot performance on most metrics even surpasses that of models after training on the CI-ToD dataset.

1 Introduction

Consistency identification in task-oriented dialogue (CI-ToD) is a critical task to detect various categories of dialogue hallucinations [Dziri *et al.*, 2021; Li *et al.*, 2022; Huang *et al.*, 2023; Zhang *et al.*, 2023; Liu *et al.*, 2024]. Such hallucinations in dialogue systems can lead to responses that are inconsistent with the user query, the knowledge base, or the dialogue history [Qin *et al.*, 2021], which results in users receiving incorrect or misleading feedback, ultimately diminishing the overall user experience. Given the importance of this issue, researchers are concentrating on detecting consistency and addressing hallucinations in system responses by aligning them with both the dialogue context and the knowledge base, which gains increasing attention.



(a) Traditional method depends on labeled data

(b) LLM can solve CI-ToD task through zero-shot prompting

(c) Our framework can improve performance of LLM by collaboration

Figure 1: Traditional method for CI-ToD requires large amounts of labeled data (a). Directly using the current large language model for CI-ToD tasks can solve the need for labeled data (b). Our framework can further improve the performance of LLM and do not rely on any labeled data (c).

In the literature, CI-ToD typically contains three sub-tasks: user query inconsistency identification (QI), dialogue history inconsistency identification (HI) and knowledge base inconsistency identification (KBI), to judge whether the dialogue response are inconsistent with the user query, the knowledge base, or the dialogue history [Qin *et al.*, 2021]. Since the three sub-tasks are highly tied, the previous approaches mainly focus on how to model the interaction across QI,

[*] Corresponding Author.

HI and KBI [Chen *et al.*, 2017; Balaraman *et al.*, 2021; Qin *et al.*, 2021]. Specifically, Qin *et al.* [2021] introduce a vanilla multi-task framework to implicitly model the interaction across QI, HI and KBI. CGIM [Qin *et al.*, 2022] introduces a cycle-interactive learning model that leverages triple-interaction mechanisms to explicitly consider the interaction between the three sub-tasks. Recently, Ding *et al.* [2024] develop a plug-and-play adapter to integrate external knowledge for CI-ToD, achieving promising performance.

While these methods have achieved notable success, traditional approaches still heavily rely on high-quality training data, which remains challenging to obtain in practical applications (see Figure 1 (a)). Recently, large language models (LLMs) have garnered considerable attention for their remarkable performance across numerous tasks [Zhao *et al.*, 2023; Qin *et al.*, 2024; Liang *et al.*, 2024; Vatsal and Dubey, 2024; Wang *et al.*, 2025; Qin *et al.*, 2025; Chen *et al.*, 2025]. Unfortunately, there is little study exploring LLM for CI-ToD, which is a potential approach to alleviate the data scarcity.

Motivated by this, we make the first attempt to investigate LLM for CI-ToD, which does not require any training data (see Figure 1 (b)). However, unlike applying LLM for other tasks, the unique challenge for LLM in CI-ToD is how to effectively model the interaction across the related sub-tasks: QI, HI and KBI in CI-ToD. Inspired by this, we further introduce a multi-agent collaboration framework for CI-ToD (MAC-CIToD), which is shown in Figure 1 (c). Specifically, we explore three collaboration paradigms in MAC-CIToD: (1) *Full Connection paradigm*, where each agent is fully connected with all other agents, allowing for comprehensive information exchange among multiple agents; (2) *Cycle Connection paradigm*, where agents exchange information with neighboring agents to enhance response generation by introducing diverse perspectives, encouraging agents to incorporate knowledge from different angles; and (3) *Central Connection paradigm*, which prioritizes information exchange involving a central agent while allowing other agents to collaborate selectively, facilitating the transmission of effective information to the central agent and ensuring communication among the remaining agents.

We evaluate our framework in the standard benchmark, and experimental results show that MAC-CIToD outperforms all baseline LLM agent approaches. Additionally, it surpasses the previously trained state-of-the-art approach in most evaluation metrics, further demonstrating its superiority.

The contribution of this work can be summarized as:

- To the best of our knowledge, we are the first to investigate LLM for CI-ToD, which does not require any labeled data.
- We further introduce a novel multi-agent collaboration framework (MAC-CIToD) for CI-ToD task and systematically explore three collaboration paradigms, hoping to provide insights for further research.
- Experiments on the standard benchmark demonstrate that MAC-CIToD achieves superior performance. In addition, MAC-CIToD can even surpass previous approaches with training on most metrics.

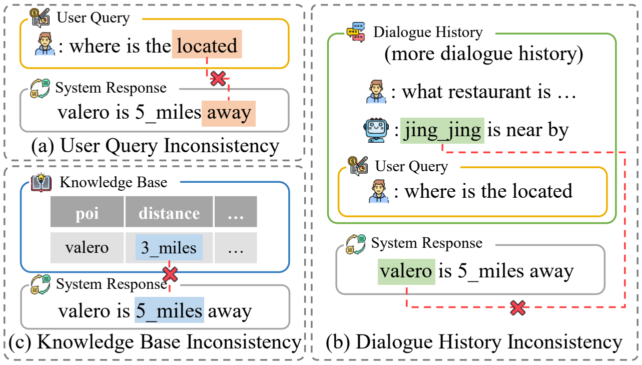


Figure 2: Consistency identification in task-oriented dialog (CI-ToD) including User Query Inconsistency (QI), Dialogue History Inconsistency (HI), and Knowledge Base Inconsistency (KBI). QI denotes that the system response is inconsistent with user query (a). HI denotes that the system response is inconsistent with previous dialogue history (b). KBI denotes that the system response is inconsistent with the knowledge base (c). Cross mark represents inconsistency.

To facilitate the further research, our code will be available at https://github.com/WPENGxs/MAC-CIToD.

2 Problem Definition

Consistency identification in task-oriented dialogue (CI-ToD) is considered as a multi-label classification task [Qin *et al.*, 2021]. Specifically, based on the input dialogue history $\mathcal{H} = \left\{ \left(d_1^u, d_1^s\right), \left(d_2^u, d_2^s\right), ..., \left(d_{n-1}^u, d_{n-1}^s\right) \right\}$, the corresponding knowledge base $\mathcal{KB}$, the user query d_n^u, and the system response d_n^s, the model needs to determine the inconsistency and outputs the corresponding inconsistency label set $\mathbb{L}$. It can be defined as:

$$\mathbb{L} = \text{Model}\left(\mathcal{H}, \mathcal{KB}, d_n^u, d_n^s\right), \tag{1}$$

where the label set $\mathbb{L} = \{\mathcal{L}_{QI}, \mathcal{L}_{HI}, \mathcal{L}_{KBI}\}$ represents the inconsistency labels in (i) User Query Inconsistency (QI) between the user query d_n^u and the system response d_n^s (see Figure 2 (a)); (ii) Dialogue History Inconsistency (HI) between the dialogue history $\mathcal{H}$ and the system response d_n^s, (see Figure 2 (b)); and (iii) Knowledge Base Inconsistency (KBI) between the corresponding knowledge base $\mathcal{KB}$ and the system response d_n^s, (see Figure 2 (c)).

3 MAC-CIToD

In this work, we introduce a multi-agent collaboration framework for CI-ToD (MAC-CIToD) to build interaction across QI, HI, and KBI. Specifically, MAC-CIToD includes two parts: (1) CI-ToD Basic Agents (§3.1) define basic agents for each type of inconsistency, and (2) Collaboration Paradigms (§3.2) describe the advanced multi-agent collaboration between basic agents.

3.1 CI-ToD Basic Agents

To model sub-tasks in CI-ToD, we construct three basic agents for each type of inconsistency, including QI Agent, HI Agent, and KBI Agent, which is shown in Figure 3 (a).

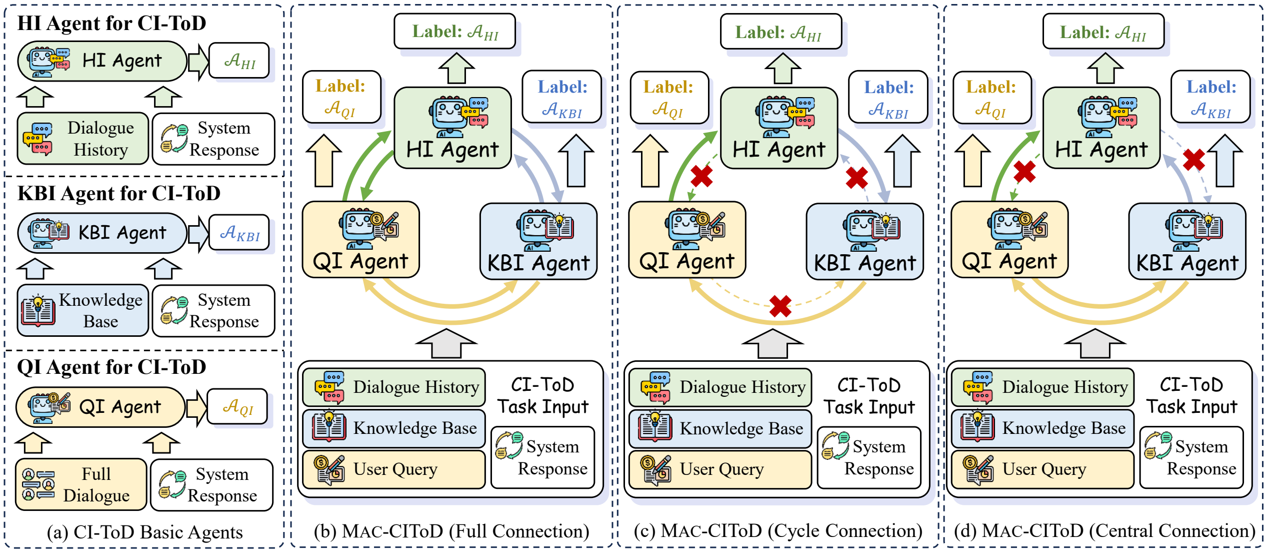


Figure 3: The main framework of MAC-CIToD. Figure (a) presents the architecture of CI-ToD basic agents. Figure (b, c, d) presents the different multi-agent collaboration paradigms.

QI Agent. Based on the full dialogue $\mathcal{D}$, QI Agent determines whether the dialogue contains User Query Inconsistency. It can be defined as:

$$\mathcal{A}_{QI} = \underset{QI}{\operatorname{argmax}}\, \mathrm{P}\left(\mathcal{A}^{p}_{QI}|\mathcal{P}_{QI},\mathcal{D}\right), \tag{2}$$

where $\mathcal{P}_{QI}$ represents the prompt about QI. Full dialogue $\mathcal{D} = \{(d^u_1, d^s_1), (d^u_2, d^s_2), ..., (d^u_n, d^s_n)\}$. $\mathcal{A}_{QI}$ represents QI answer of the LLM, where $\mathcal{L}_{QI}$ can be parsed from it. $\mathcal{A}^{p}_{QI}$ is the potential answering path.

HI Agent. Based on the dialogue history $\mathcal{H}$ and the system response d^s_n, HI Agent determines whether the dialogue contains Dialogue History Inconsistency. It can be defined as:

$$\mathcal{A}_{HI} = \underset{HI}{\operatorname{argmax}}\, \mathrm{P}\left(\mathcal{A}^{p}_{HI}|\mathcal{P}_{HI},\mathcal{H},d^s_n\right), \tag{3}$$

where $\mathcal{P}_{HI}$ represents the prompt about HI. $\mathcal{A}_{HI}$ represents HI answer of the LLM, where $\mathcal{L}_{HI}$ can be parsed from it. $\mathcal{A}^{p}_{HI}$ is the potential answering path.

KBI Agent. Based on the knowledge base $\mathcal{KB}$ and the system response d^s_n, KBI Agent determines whether the dialogue contains Knowledge Base Inconsistency. It can be defined as:

$$\mathcal{A}_{KBI} = \underset{KBI}{\operatorname{argmax}}\, \mathrm{P}\left(\mathcal{A}^{p}_{KBI}|\mathcal{P}_{KBI},\mathcal{KB},d^s_n\right), \tag{4}$$

where $\mathcal{P}_{KBI}$ represents the prompt about KBI. $\mathcal{A}_{KBI}$ represents KBI answer of the LLM, where $\mathcal{L}_{KBI}$ can be parsed from it. $\mathcal{A}^{p}_{KBI}$ is the potential answering path.

Prompt Detail. To enhance performance in this task and unlock the ability of LLM, we design the specialized prompt for QI Agent, HI Agent, and KBI Agent, which is shown as:

[Basic Prompt P_{basic}] : You are an expert at ...
[Task Definition P_{task}] : User Query Inconsistency (QI) / Dialogue History Inconsistency (HI) / Knowledge Base Inconsistency (KBI) denotes that ...
[Output Constraints P_{cons}] : ...give your reasons and output the json format of ...

These prompt can be defined as:

- Basic Prompt P_{basic} allows LLM to perform role-playing through expert prompts.
- Task Definition P_{task} provides definitions of each type of inconsistency in CI-ToD task for QI Agent, HI Agent, and KBI Agent.
- Output Constraints P_{cons} restrict the output format of LLM so that label $\mathcal{L}$ can be accurately extracted.

The overall structure of our prompt $\mathcal{P}$ is as follows:

$$\mathcal{P} = (P_{basic}, P_{task}, P_{cons}). \tag{5}$$

3.2 Collaboration Paradigms

After obtaining the answers $\mathcal{A}^{init}_{QI}, \mathcal{A}^{init}_{HI}, \mathcal{A}^{init}_{KBI}$ from the initial turn of CI-ToD Basic Agents, we set up different collaboration paradigms in the following turns to model their contribution on the final answer of CI-ToD. Below is a detailed explanation of three collaboration paradigms.

Full Connection. To enable all agents to make decisions based on comprehensive information, the full connection paradigm assists the agent in the second turn by providing the labels from CI-ToD basic agents in the initial turn, as illustrated in Figure 3 (b). This approach allows the large language model to comprehensively reference all information,

Method	QI F1	HI F1	KBI F1	Overall Acc.
Traditional method†				
BERT-multi-task [Devlin *et al.*, 2019]	0.691	0.555	0.740	0.500
XLNet-multi-task [Yang, 2019]	0.725	0.487	0.736	0.509
Longformer-multi-task [Beltagy *et al.*, 2020]	0.717	0.500	0.710	0.497
BART-multi-task [Lewis *et al.*, 2020]	0.744	0.510	0.761	0.513
CGIM [Qin *et al.*, 2022]	0.764	0.567	0.772	0.563
PPA [Ding *et al.*, 2024]	0.772	0.624	0.781	0.592
Llama-3.1-8B-Instruct [Dubey *et al.*, 2024]				
`Reflexion` [Shinn *et al.*, 2024]	0.376	0.175	0.312	0.094
`Debate` [Liang *et al.*, 2023]	0.372	0.152	0.403	0.075
`S`^{3} `Agent` [Wang *et al.*, 2024b]	0.693	0.350	0.591	0.213
MAC-CIToD (Full Connection)	0.706 (+0.013)	0.480 (+0.130)	0.619 (+0.028)	0.242 (+0.029)
MAC-CIToD (Cycle Connection)	0.727 (+0.034)	0.483 (+0.133)	**0.677 (+0.086)**	**0.301 (+0.088)**
MAC-CIToD (Central Connection)	**0.753 (+0.060)**	**0.500 (+0.150)**	0.586 (-0.005)	0.283 (+0.070)
gpt-3.5-turbo [OpenAI, 2022]				
`Reflexion` [Shinn *et al.*, 2024]	0.491	0.285	0.530	0.330
`Debate` [Liang *et al.*, 2023]	0.579	0.351	**0.626**	0.194
`S`^{3} `Agent` [Wang *et al.*, 2024b]	0.328	0.165	0.332	0.191
MAC-CIToD (Full Connection)	**0.800 (+0.221)**	0.545 (+0.194)	0.513 (-0.113)	**0.418 (+0.088)**
MAC-CIToD (Cycle Connection)	0.748 (+0.169)	0.528 (+0.177)	0.573 (-0.053)	0.415 (+0.085)
MAC-CIToD (Central Connection)	0.756 (+0.177)	**0.597 (+0.246)**	0.537 (-0.089)	0.406 (+0.076)
GLM-4-9B-chat [GLM *et al.*, 2024]				
`Reflexion` [Shinn *et al.*, 2024]	0.734	0.357	0.588	0.342
`Debate` [Liang *et al.*, 2023]	0.633	0.360	0.668	0.230
`S`^{3} `Agent` [Wang *et al.*, 2024b]	0.583	0.056	0.312	0.336
MAC-CIToD (Full Connection)	**0.804 (+0.070)**	0.366 (+0.006)	**0.697 (+0.029)**	0.427 (+0.085)
MAC-CIToD (Cycle Connection)	0.782 (+0.048)	0.467 (+0.107)	0.660 (-0.008)	**0.437 (+0.095)**
MAC-CIToD (Central Connection)	0.742 (+0.008)	**0.488 (+0.128)**	0.680 (+0.012)	0.408 (+0.066)
Gemma-2-9B-It [Team *et al.*, 2024]				
`Reflexion` [Shinn *et al.*, 2024]	0.410	0.304	0.384	0.201
`Debate` [Liang *et al.*, 2023]	0.481	0.207	0.448	0.198
`S`^{3} `Agent` [Wang *et al.*, 2024b]	0.815	0.538	0.660	**0.522**
MAC-CIToD (Full Connection)	0.884 (+0.069)	**0.624 (+0.086)**	0.687 (+0.027)	0.474 (-0.048)
MAC-CIToD (Cycle Connection)	**0.902 (+0.087)**	0.621 (+0.083)	**0.688 (+0.028)**	0.474 (-0.048)
MAC-CIToD (Central Connection)	0.896 (+0.081)	0.468 (-0.070)	0.671 (+0.011)	0.333 (-0.189)
gpt-4o [Achiam *et al.*, 2023]				
`Reflexion` [Shinn *et al.*, 2024]	0.702	0.482	0.724	0.506
`Debate` [Liang *et al.*, 2023]	0.798	0.520	0.766	0.484
`S`^{3} `Agent` [Wang *et al.*, 2024b]	0.700	0.254	0.670	0.455
MAC-CIToD (Full Connection)	0.886 (+0.088)	0.550 (+0.030)	0.835 (+0.069)	0.512 (+0.006)
MAC-CIToD (Cycle Connection)	**0.910 (+0.112)**	0.582 (+0.062)	**0.840 (+0.074)**	0.556 (+0.050)
MAC-CIToD (Central Connection)	0.904 (+0.106)	**0.629 (+0.109)**	0.831 (+0.065)	**0.584 (+0.078)**

Table 1: Main results. **Bold number** presents the best results achieved by these methods on the current model. † represents that these models or frameworks in traditional methods have been trained on the CI-ToD dataset. The performance of MAC-CIToD gains/drops relative to the baseline best result are highlighted with blue/red in the Table.

thereby promoting a more accurate determination. Taking HI as an example, it can be formally defined as:

$$\mathcal{A}_{HI} = \operatorname*{argmax}_{HI} \mathrm{P}\left(\mathcal{A}^{p}_{HI} | \mathcal{P}_{HI}, \mathbb{L}^{init}, \mathcal{H}, d^{s}_{n}\right), \tag{6}$$

where $\mathbb{L}^{init}$ represents all inconsistent labels. For CI-ToD task, $\mathbb{L}^{init} = \left(\mathcal{L}^{init}_{QI}, \mathcal{L}^{init}_{HI}, \mathcal{L}^{init}_{KBI}\right)$.

Cycle Connection. To enable agents to leverage information from other perspectives, the cycle connection allows all neighboring agents to transmit the corresponding inconsistent labels in one direction of the cycle, as shown in Figure 3 (c). This ensures that all agents receive the same amount of information and incorporate knowledge from multiple angles. Taking HI as an example, it can be defined as:

$$\mathcal{A}_{HI} = \operatorname*{argmax}_{HI} \mathrm{P}\left(\mathcal{A}^{p}_{HI} | \mathcal{P}_{HI}, \mathcal{L}^{init}_{near}, \mathcal{H}, d^{s}_{n}\right), \tag{7}$$

where $\mathcal{L}^{init}_{near}$ represents the inconsistent label from nearby agent. For CI-ToD task and set the starting direction to $QI \rightarrow HI$, $\mathcal{L}^{init}_{near} = \mathcal{L}^{init}_{KBI}$ in the QI agent, $\mathcal{L}^{init}_{near} = \mathcal{L}^{init}_{QI}$ in the HI agent, $\mathcal{L}^{init}_{near} = \mathcal{L}^{init}_{HI}$ in the KBI agent.

Central Connection. To enable certain agents to access more information for enhanced performance, while allowing other agents to collaborate selectively. The central connection paradigm selects a central agent to receive all labels from the other agents. In contrast, the remaining agents exchange labels with each other, as shown in Figure 3 (d). This ensures that more information flows into the central agent while other agents get the necessary reference information. Taking HI as the central agent as an example, it can be defined as:

$$\mathcal{A}_{HI} = \overset{central}{\operatorname*{argmax}_{HI}} \mathrm{P}\left(\mathcal{A}^{p}_{HI} | \mathcal{P}_{HI}, \mathbb{L}^{init}_{other}, \mathcal{H}, d^{s}_{n}\right), \tag{8}$$

$$\mathcal{A}_{QI} = \underset{QI}{\overset{other}{\operatorname{argmax}}} \mathrm{P}\left(\mathcal{A}^{p}_{QI} | \mathcal{P}_{QI}, \mathcal{L}^{init}_{KBI}, \mathcal{D}\right), \qquad (9)$$

$$\mathcal{A}_{KBI} = \underset{KBI}{\overset{other}{\operatorname{argmax}}} \mathrm{P}\left(\mathcal{A}^{p}_{KBI} | \mathcal{P}_{KBI}, \mathcal{L}^{init}_{QI}, \mathcal{KB}, d^{s}_{n}\right), \qquad (10)$$

where central agent accepts more inconsistent information, $\mathbb{L}^{init}_{other} = \left(\mathcal{L}^{init}_{QI}, \mathcal{L}^{init}_{KBI}\right)$.

4 Experiments

4.1 Experimental Settings

Following previous work [Qin *et al.*, 2021; Qin *et al.*, 2022; Ding *et al.*, 2024], we use the standard CI-ToD benchmark for experiments. For the evaluation, we employ four metrics based on prior work: QI F1, HI F1, KBI F1, and overall Acc.

We conduct experiments on five backbones: Llama-3.1-8B-instruct [Dubey *et al.*, 2024], GLM-4-9B-Chat [GLM *et al.*, 2024], Gemma-2-9B-It [Team *et al.*, 2024], gpt-3.5-turbo [OpenAI, 2022], and gpt-4o [Achiam *et al.*, 2023]. All open source models are obtained from HuggingFace Library [Wolf *et al.*, 2020]. For the GPT series models, the temperature is 0.3, the top p is 1, and the output max token length is 512. For the open source model, the temperature is 0.7, the top p is 0.8, and the output max token length is 512.

4.2 Baselines

For traditional methods, we select six models and frameworks, including BERT [Devlin *et al.*, 2019], XLNet [Yang, 2019], Longformer [Beltagy *et al.*, 2020], BART [Lewis *et al.*, 2020], CGIM [Qin *et al.*, 2022], and PPA [Ding *et al.*, 2024]. In the traditional method, the performance of PPA is derived from Ding *et al.* [2024], and other performance is derived from Qin *et al.* [2022].

For the LLM-based approaches, we select the following representative approaches, including:

- `Reflexion` [Shinn *et al.*, 2024]: Reflexion enhances the performance of the LLM on a certain task through LLM's self-reflection and iterative optimization.
- `Debate` [Liang *et al.*, 2023]: Debate aims to enhance LLM's understanding of the same problem through discussion between two roles, and finally let the LLM referee decide the party most likely to be correct, and finally get an answer.
- `S`^{3} `Agent` [Wang *et al.*, 2024b]: S^3 Agent uses the multi-agent system to solve the problem from multiple perspectives, and the decision agent makes the final decision based on these perspectives. This avoids bias and errors caused by a single perspective and improves the performance of the LLM on specific tasks.

4.3 Main Results

The main results are shown in Table 1. We otain the following observations:

(1) These advanced agent methods still have a gap from the performance of traditional methods. Simply using stronger LLM could not fully solve the CI-ToD task. Even the most advanced model (gpt-4o) can only exceed the performance of the early traditional method BART-multi-task, and it still could not surpass CGIM, with a gap of 5.7% in overall Acc.

Method	Overall Acc	
Llama-3.1-8B-Instruct		
CI-ToD Basic Agents	0.145	
MAC-CIToD (worst connection method)	0.242	↑0.097
MAC-CIToD (best connection method)	0.301	↑0.156
gpt-3.5-turbo		
CI-ToD Basic Agents	0.406	
MAC-CIToD (worst connection method)	0.406	↑0.000
MAC-CIToD (best connection method)	0.418	↑0.012
gpt-4o		
CI-ToD Basic Agents	0.484	
MAC-CIToD (worst connection method)	0.512	↑0.028
MAC-CIToD (best connection method)	0.584	↑0.100

Table 2: The results of CI-ToD Basic Agents and MAC-CIToD. We selected the best and worst performance of MAC-CIToD for comparison based on overall Acc.

(2) Our framework attains best performance after collaboration. MAC-CIToD demonstrates superior performance compared with LLM agent baselines, and outperforms trained traditional methods. The zero-shot performance of MAC-CIToD on gpt-4o comprehensively surpasses the results of the best traditional method, PPA. Specifically, the improvements in MAC-CIToD (Central Connection) are 13.2% for QI F1, 0.5% for HI F1, and 5.0% for KBI F1.

(3) Even on smaller LLMs, our method can still achieve competitive performance compared to traditional methods. On smaller LLMs, we can see that the F1 metrics are equal to or even exceed traditional methods, for example, the performance on Llama-3.1-8B-instruct and GLM-4-9B-Chat is close to the performance of the traditional method BERT. This shows that our multi-agent collaboration paradigms can help smaller LLMs better complete complex CI-ToD tasks.

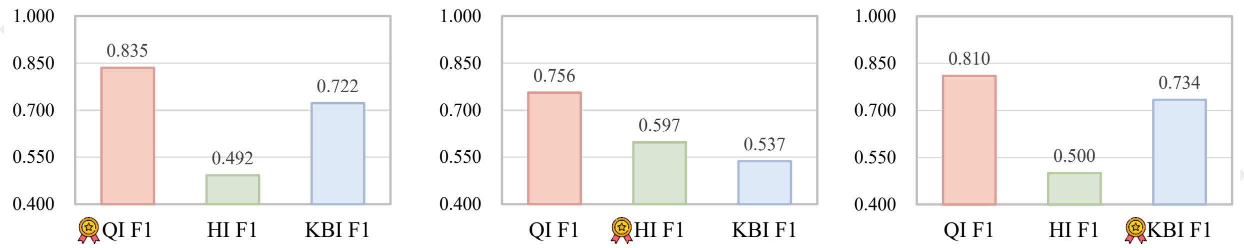
4.4 MAC-CIToD Remains Robust For All Connection Paradigms

To explore the robustness of MAC-CIToD connection paradigms, we show the detailed performance of the CI-ToD Basic Agents and MAC-CIToD, which is shown in Table 2.

It can be seen that the performance of CI-ToD Basic Agents is improved after multi-agent collaboration, with an average improvement of 6.5%, which shows that LLM can make full use of the label information to further improve performance. Additionally, our results from the worst and best connection paradigms show that, in the CI-ToD task, the performance of LLMs improves across all connection paradigms. This demonstrates the robustness of the different connection strategies employed by MAC-CIToD.

Input information	gpt-4o		GLM-4-9B-chat		gpt-3.5-turbo		Llama-3.1-8B-Instruct	
HI Agent	0.537		0.432		0.545		0.356	
+ QI Information	0.582	↑0.045	0.467	↑0.035	0.528	↓0.017	0.483	↑0.127
+ QI, KBI Information	**0.629**	↑0.092	**0.488**	↑0.056	**0.597**	↑0.052	**0.500**	↑0.144
+ QI, KBI, HI Information	0.550	↑0.013	0.366	↓0.066	0.545	↑0.000	0.480	↑0.124

Table 3: The performance of different input information. HI Agent presents the performance of HI F1 in CI-ToD Basic Agents. "+ Information" the performance of HI F1 in MAC-CIToD when inputting different information. **Bold number** presents the best results achieved by these input information on the current model.



(a) QI Agent as the central agent (b) HI Agent as the central agent (c) KBI Agent as the central agent

Figure 4: The results of different central agents in MAC-CIToD (Central Connection). This figure illustrates the F1 performance of MAC-CIToD (Central Connection) while QI Agent (a), HI Agent (b), and KBI Agent (c) as the central agent. The reward sign on the left side of F1 indicates that the performance of this central agent is the best when compared to other agents serving as the central agent.

4.5 Information From Different Sub-tasks Can Effectively Boost The Performance of The Target Sub-task

In order to explore the impact of different sub-task information on the further reasoning of LLM, we analyze the impact of inputting different information on the HI Agent. The final results are shown in Table 3.

From Table 3, we can draw three findings: (1) For most LLMs, inputting the information from CI-ToD Basic Agents can help MAC-CIToD to improve performance. For HI Agent, it improved whether the amount of input information is more or less. (2) The appropriate amount of information input is conducive to the performance improvement. When the HI Agent receives three different inputs of agents, the QI and KBI input can fully stimulate the ability of the HI Agent in most models, so that it can finally get the highest HI F1. (3) The input of identical information leads to a disruptive effect on MAC-CIToD. For the HI agent, optimal performance is consistently achieved when only QI and KBI are input. When HI information is additionally included, performance degrades compared to the scenario with just QI and KBI. This suggests that redundant input information can mislead the framework, resulting in less accurate label.

4.6 The Performance of The Central Agent is Consistently The Best in MAC-CIToD Central Connection

In order to test the impact of the different central agent in the MAC-CIToD central connection, we use gpt-3.5-turbo to test the performance of using QI, HI, and KBI Agent as the central agent. The final results are shown in the Figure 4.

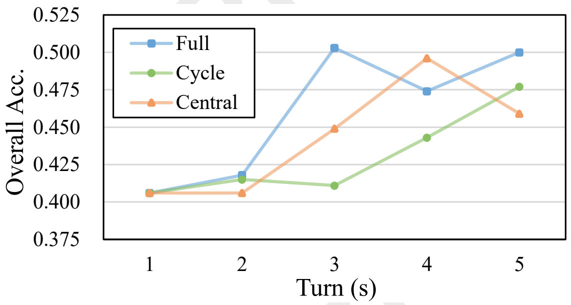
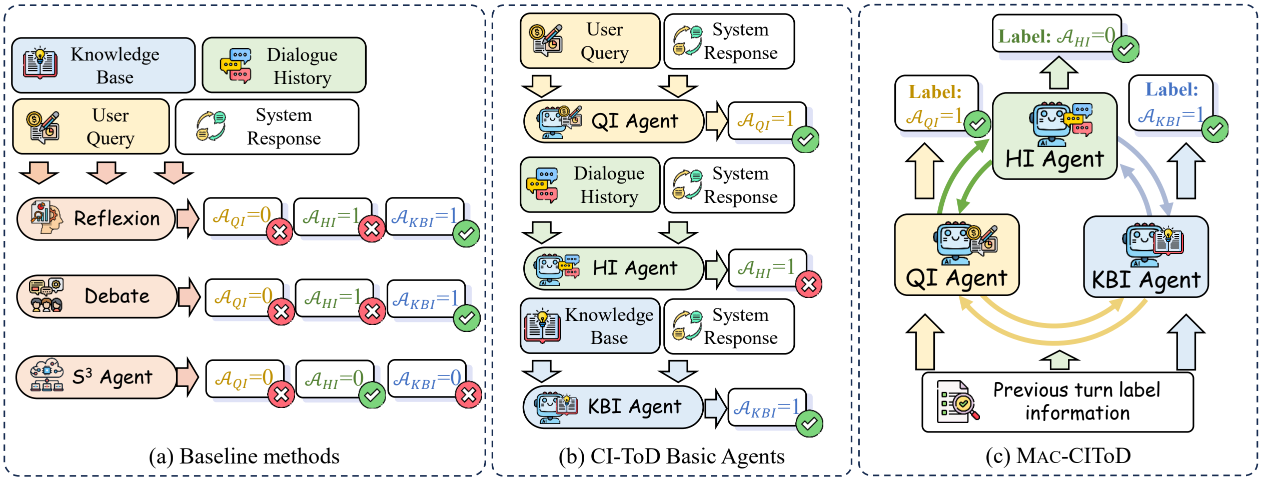


Figure 5: The performance of MAC-CIToD in n-th turn. turn 1 presents CI-ToD Basic Agents.

It can be seen that the performance of the central agent is always the best. When the QI Agent is the central agent, it outperforms the HI Agent and KBI Agent by 7.9% and 2.5% on QI F1; The HI Agent is the central agent, it outperforms the QI Agent and KBI Agent by 10.5% and 9.7% on HI F1; The KBI Agent is the central agent, it outperforms the QI Agent and HI Agent by 1.2% and 19.7% on KBI F1. This indicates that the performance of the central agent is consistently superior, highlighting that the flow of information can be effectively leveraged when the central agent coordinates the interaction.

4.7 The Multi-turn Information Exchange in MAC-CIToD Can Further Improve The Performance

We also explore the impact of multi-turn on performance. Specifically, we use gpt-3.5-turbo to test the performance of



(a) Baseline methods

(b) CI-ToD Basic Agents

(c) MAC-CIToD

Figure 6: Case study of (a) baseline methods, (b) CI-ToD Basic Agents, and (c) MAC-CIToD. When inputting knowledge base and dialogue, baseline methods all output incorrect answers. The HI answer of CI-ToD Basic Agents is also incorrect firstly, but MAC-CIToD can modify it correctly across information communication and utilization.

2-5 turns of three connection paradigms. The performance of overall Acc. is shown in the Figure 5.

It can be seen that for the three connection paradigms, overall Acc. show an upward trend as the turn increases, cycle connection can continue to rise, and full connection and central connection achieve the best performance in the third and fourth turns respectively. This shows that for multi-agent collaboration, more turns are needed so that the agents can fully utilize the information and get the correct answer. At the same time, we observe that all turns are higher than the initial turn in overall Acc., which shows that multi-turn collaboration can enhance the performance of the agent.

4.8 Case Study

In order to understand our method more intuitively, we show a case study to specifically analyze the difference between the baseline methods and our method.

When inputting a dialogue and knowledge base, the correct labels are QI = 1, HI = 0, and KBI = 1. As shown in Figure 6 (a), all baseline methods fail to predict labels correctly. In Figure 6 (b), MAC-CIToD also makes an error in predicting HI in the initial turn. However, after collaboration, which is shown in Figure 6 (c), the HI agent can correct the mistake by incorporating information from the previous turn. Ultimately, MAC-CIToD outputs three correct labels.

5 Related Work

Consistency identification in task-oriented dialog aims to check and ensure system response is consistent with knowledge base and dialogue. Due to its importance in building task-oriented dialogue systems, some works have focused on building relevant datasets to improve the dialogue consistency of dialogue models. Welleck *et al.* [2019] create a Dialogue NLI dataset that can be used as training data to improve the consistency of a dialogue model. Song *et al.* [2020] create a large-scale dataset for open-domain dialogue and construct the profile consistency identification model to improve dialogue consistency. Nie *et al.* [2021] introduce the DialoguE COntradiction DEtection task (DECODE), it is a new conversational dialogue contradictory dataset between human-human and human-bot. Qin *et al.* [2021] propose a new benchmark for CI-ToD and evaluate some common traditional methods on this task.

To solve the CI-ToD task, some researchers focus on building the framework to reduce the occurrence of dialogue inconsistency. Qin *et al.* [2022] propose a cycle interactive learning model named CGIM, which enhances the performance by the triple-interaction and achieve best results in strong pre-trained models. Wang *et al.* [2024a] propose an efficient Multi-round Interactive Dialogue Tuning (Midi-Tuning) framework for improving dialogue consistency. Ding *et al.* [2024] propose a plug-and-play adapter, which allows the integration of knowledge base into model reasoning and enhancement of the performance in the CI-ToD task.

In contrast to these previous works, our approach is the first to explore large language models for the CI-ToD task. Furthermore, we introduce a multi-agent collaboration framework, MAC-CIToD, which fully leverages the different relationships between the three sub-tasks in CI-ToD.

6 Conclusion

In this paper, we initiate the exploration of LLMs for the CI-ToD task to alleviate the data scarcity problem. In addition, we further introduce a multi-agent collaboration framework, MAC-CIToD, including (1) *Full Connection paradigm*, (2) *Cycle Connection paradigm*, and (3) *Central Connection paradigm* to effectively model the interaction across three sub-tasks: QI, HI, and KBI in CI-ToD. Experiments on the standard benchmark show that MAC-CIToD achieves superior performance. Additionally, MAC-CIToD can outperform previous traditional methods on most metrics without any training. We hope that this research provides new insights and inspirations in this domain.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grant 62306342, 62236004, 62441603, and 62476073. This work was sponsored by the Excellent Young Scientists Fund in Hunan-Province (2024JJ4070), the Science and Technology Innovation Program of Hunan Province under Grant 2024RC3024, and CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202406). This work was also supported by the Key Laboratory of Data Intelligence and Advanced Computing in Provincial Universities, Soochow University. We are grateful for resources from the High Performance Computing Center of Central South University.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Balaraman *et al.*, 2021] Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proc. of SIGDIAL*, pages 239–251, 2021.

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[Chen *et al.*, 2017] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35, November 2017.

[Chen *et al.*, 2025] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186, 2019.

[Ding *et al.*, 2024] Zeyuan Ding, Zhihao Yang, and Hongfei Lin. A plug-and-play adapter for consistency identification in task-oriented dialogue systems. *Information Processing & Management*, 61(3):103637, 2024.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Dziri *et al.*, 2021] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proc. of EMNLP*, pages 2197–2214, 2021.

[GLM *et al.*, 2024] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[Huang *et al.*, 2023] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, pages 7871–7880, 2020.

[Li *et al.*, 2022] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*, 2022.

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

[Liang *et al.*, 2024] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, page 405–409, 2024.

[Liu *et al.*, 2024] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[Nie *et al.*, 2021] Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proc. of ACL*, pages 1699–1713, 2021.

[OpenAI, 2022] OpenAI. Introducing chatgpt. https://openai.com/index/chatgpt/, 2022. Accessed: 2025-05-16.

[Qin *et al.*, 2021] Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proc. of EMNLP*, pages 2357–2367, 2021.

[Qin *et al.*, 2022] Libo Qin, Qiguang Chen, Tianbao Xie, Qian Liu, Shijue Huang, Wanxiang Che, and Zhou Yu. CGIM: A cycle guided interactive learning model for consistency identification in task-oriented dialogue. In *Proc. of COLING*, pages 461–470, 2022.

[Qin *et al.*, 2024] Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*, 2024.

[Qin *et al.*, 2025] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. A survey of multilingual large language models. *Patterns*, 6(1):101118, 2025.

[Shinn *et al.*, 2024] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Song *et al.*, 2020] Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. Profile consistency identification for open-domain dialogue agents. In *Proc. of EMNLP*, pages 6651–6662, 2020.

[Team *et al.*, 2024] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[Vatsal and Dubey, 2024] Shubham Vatsal and Harsh Dubey. A survey of prompt engineering methods in large language models for different nlp tasks. *arXiv preprint arXiv:2407.12994*, 2024.

[Wang *et al.*, 2024a] Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *Proc. of ACL*, pages 3993–4010, 2024.

[Wang *et al.*, 2024b] Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. S3 agent: Unlocking the power of vllm for zero-shot multi-modal sarcasm detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, August 2024. Just Accepted.

[Wang *et al.*, 2025] Peng Wang, Wenpeng Lu, Chunlin Lu, Ruoxi Zhou, Min Li, and Libo Qin. Large language model for medical images: A survey of taxonomy, systematic review, and future trends. *Big Data Mining and Analytics*, 8(2):496–517, 2025.

[Welleck *et al.*, 2019] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proc. of ACL*, pages 3731–3741, 2019.

[Wolf *et al.*, 2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*, pages 38–45, 2020.

[Yang, 2019] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[Zhang *et al.*, 2023] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.