IJCAI 2025 Guangzhou August 29-31

CENTRAL SOUTH UNIVERSITY 中南大學

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], Libo Qin [1,2,*]

[1] School of Computer Science and Engineering, Central South University, China

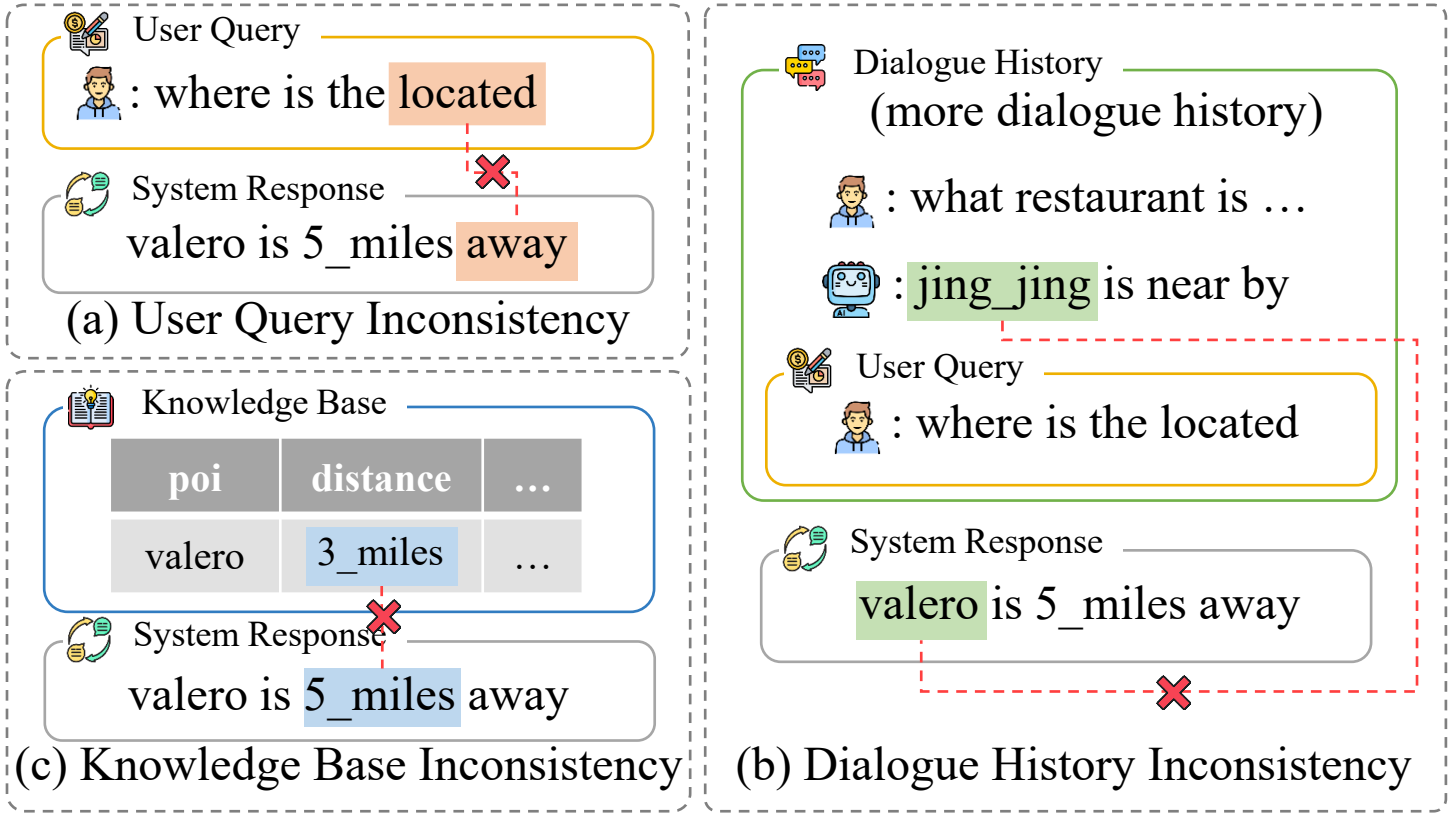
[2] Key Laboratory of Data Intelligence and Advanced Computing in Provincial Universities, Soochow University, China

[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

Background

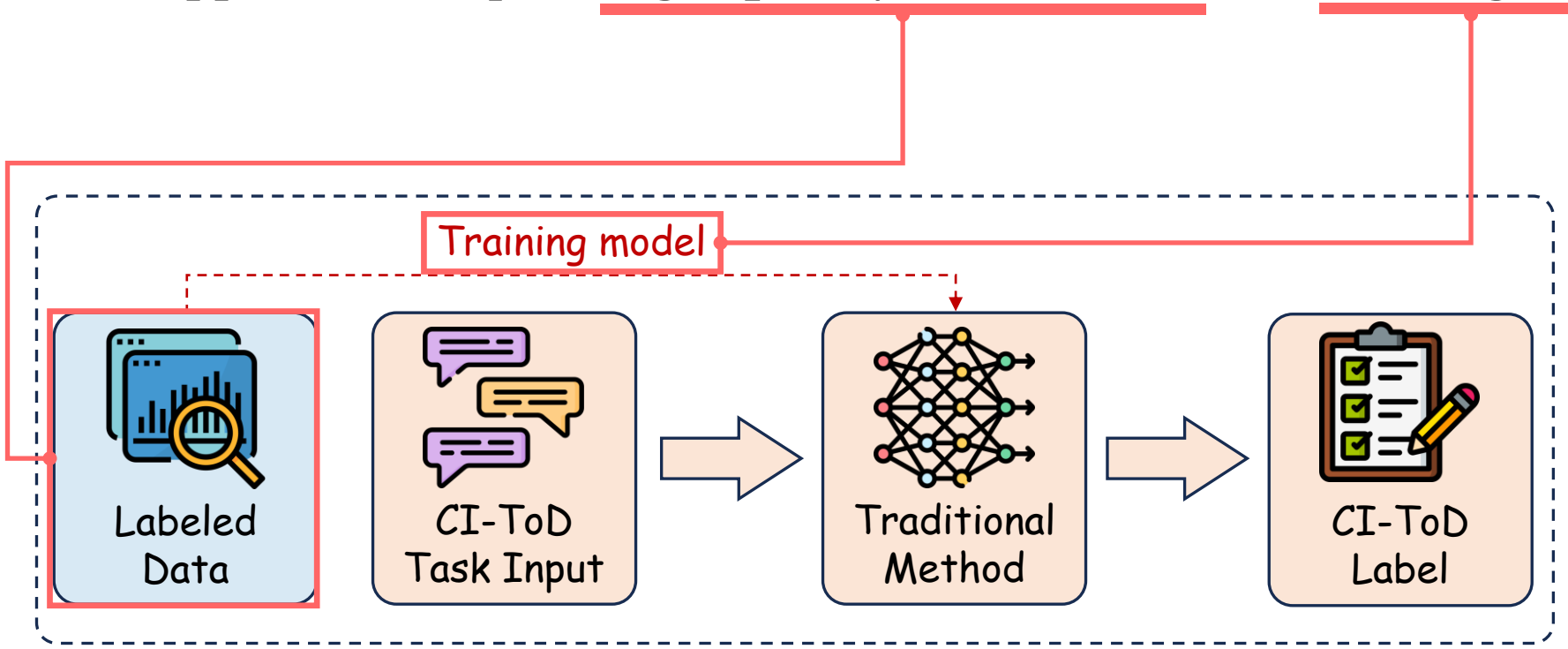


Consistency identification in task-oriented dialogue (CI-ToD)

- User query inconsistency identification (QI)
- Dialogue history inconsistency identification (HI)
- Knowledge base inconsistency identification (KBI)

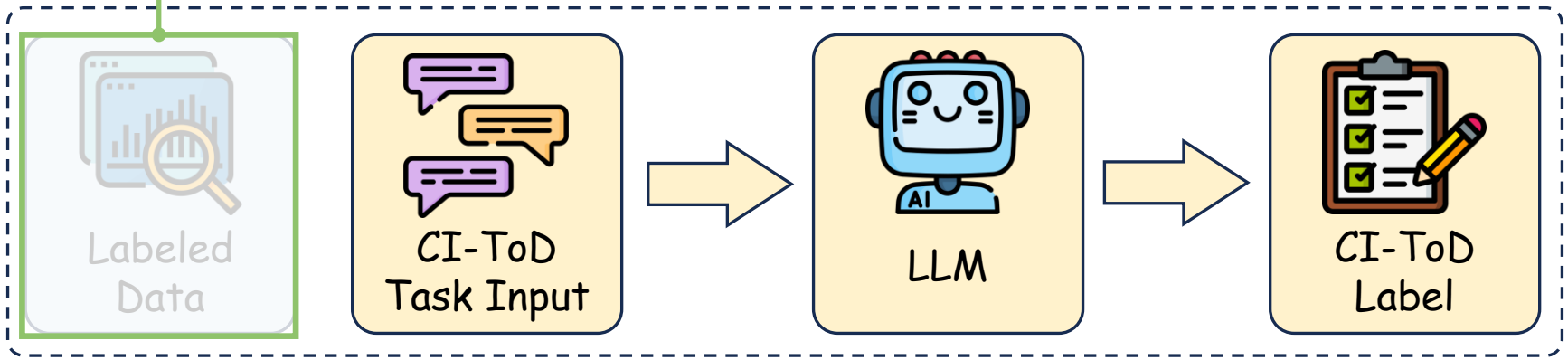
Motivation

Traditional approaches require **high-quality labeled data** and **training cost**.



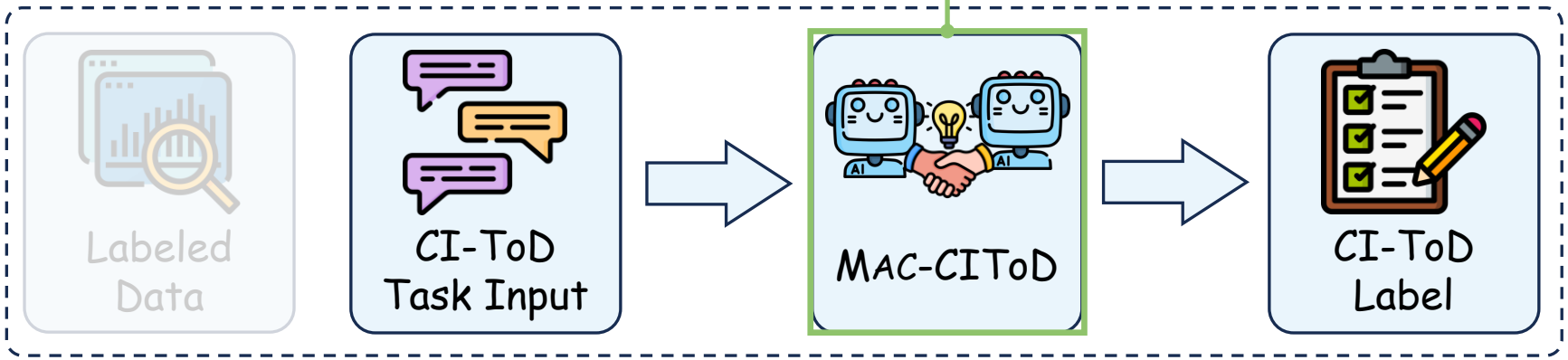
Motivation

Applying LLM to CI-ToD does **not require any training data**.

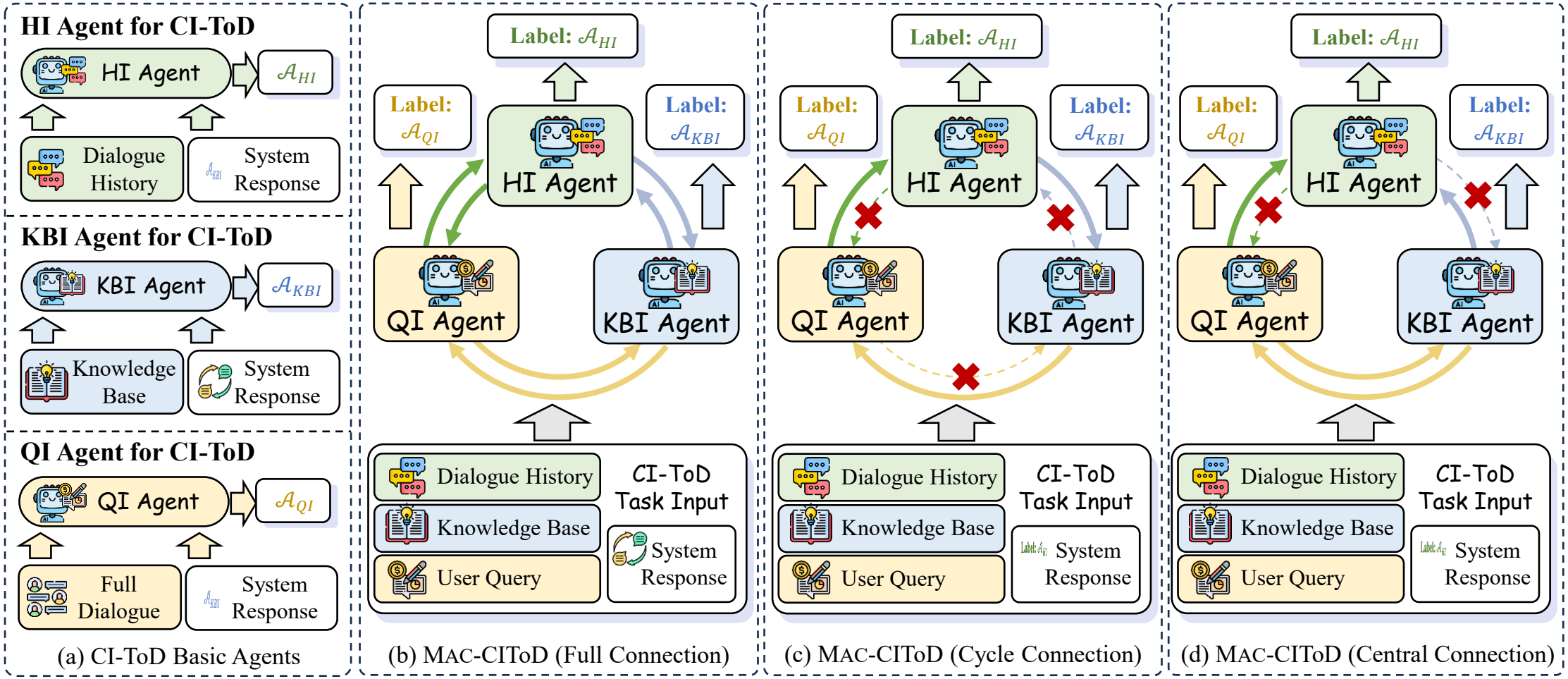


Motivation

The unique challenge for LLM in CI-ToD is how to effectively **model the interaction** across the related sub-tasks.

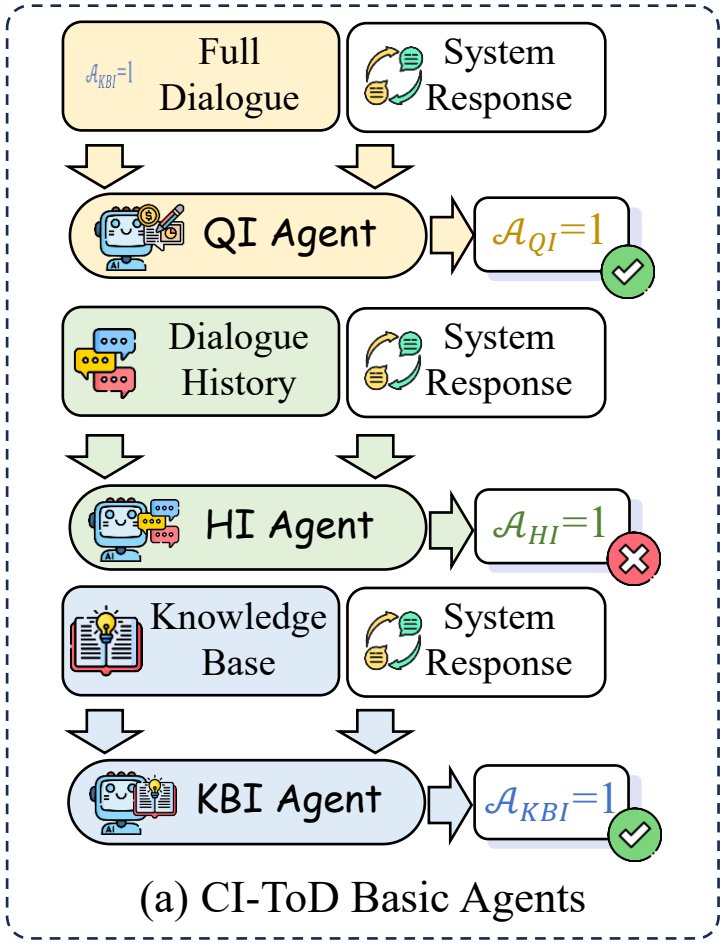


MAC-CIToD



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.

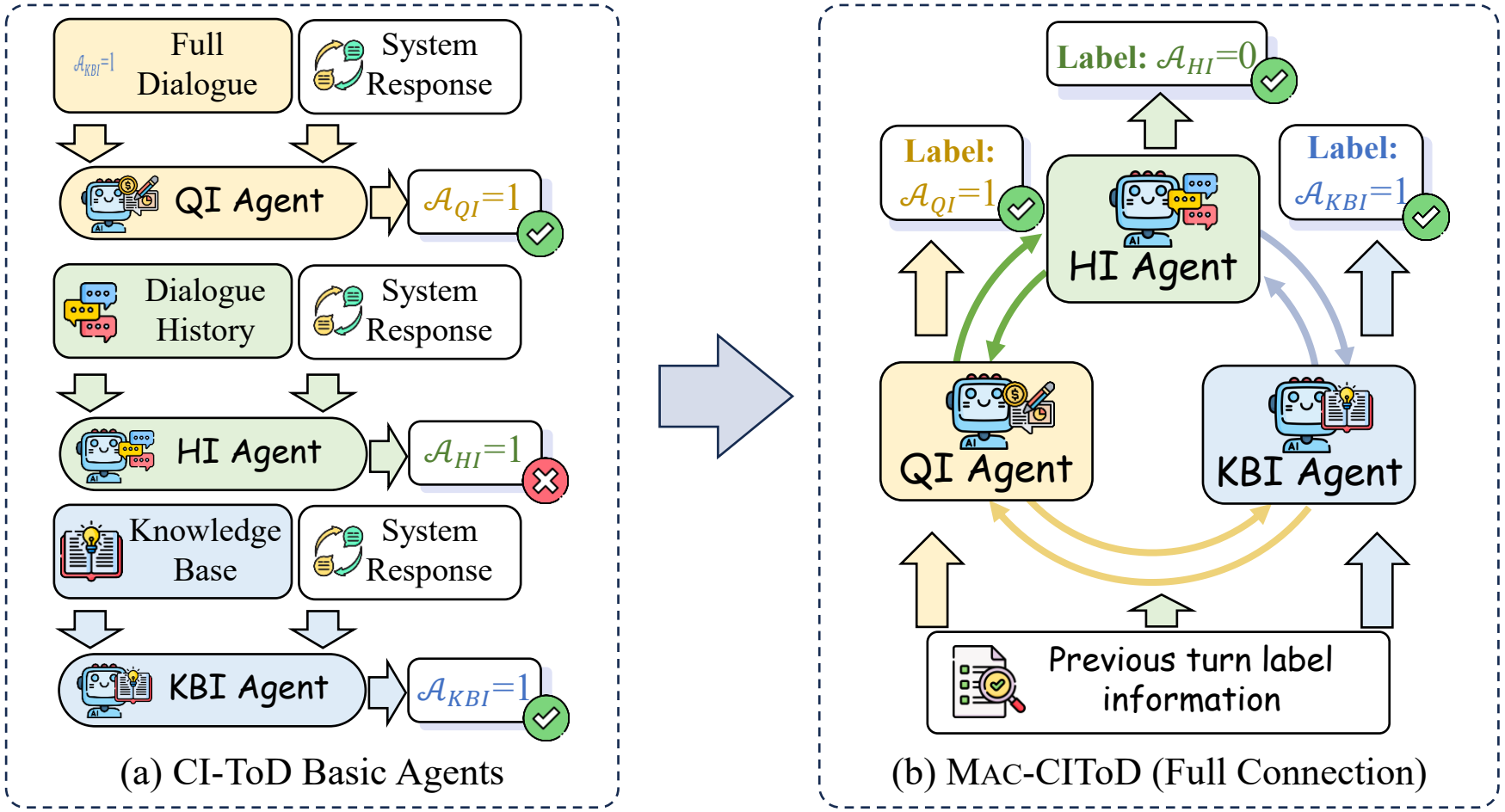
MAC-CIToD



(a) CI-ToD Basic Agents

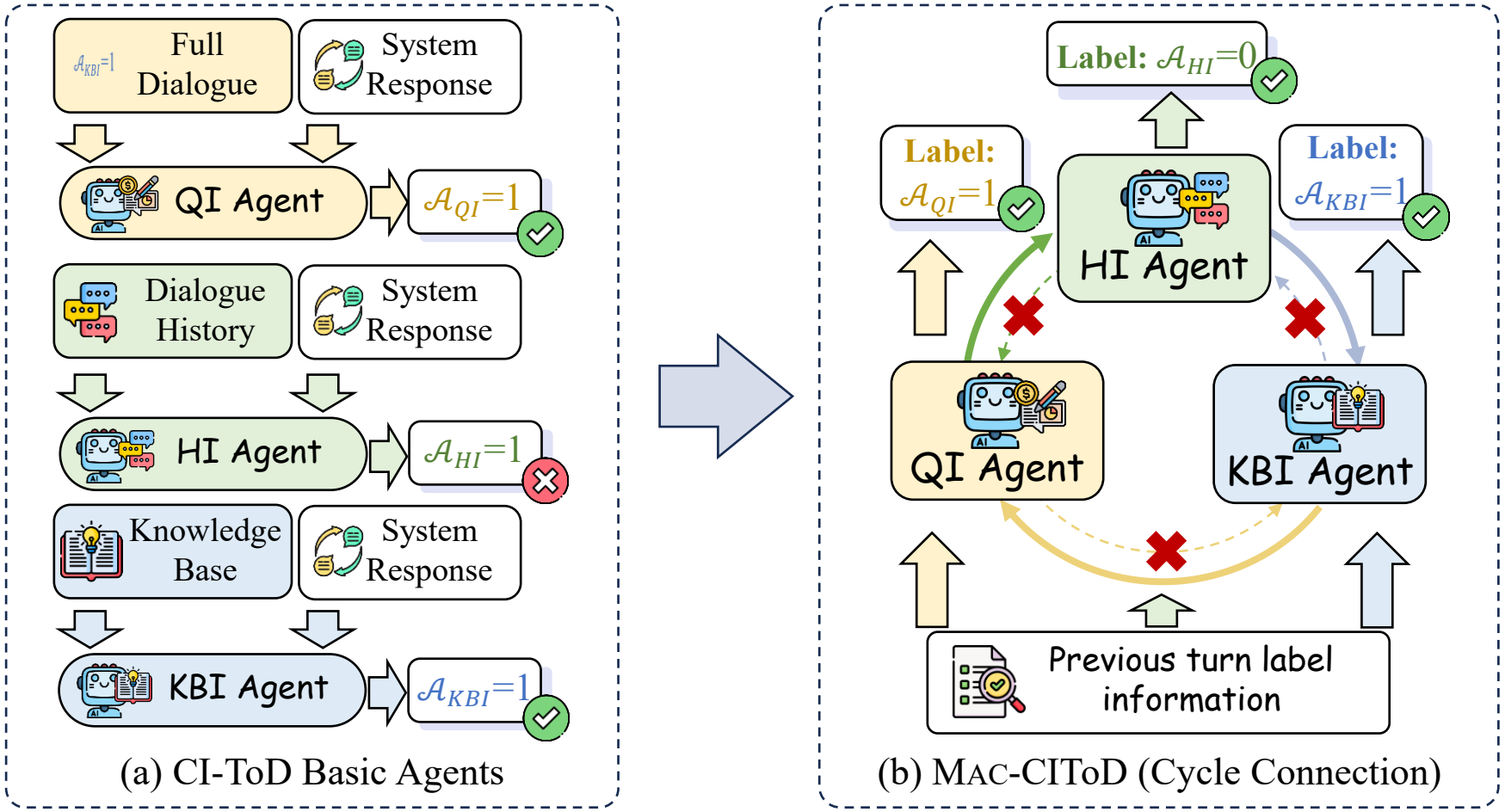
Firstly, we introduce three basic agents for each type of inconsistency: QI Agent, HI Agent, and KBI Agent.

MAC-CIToD



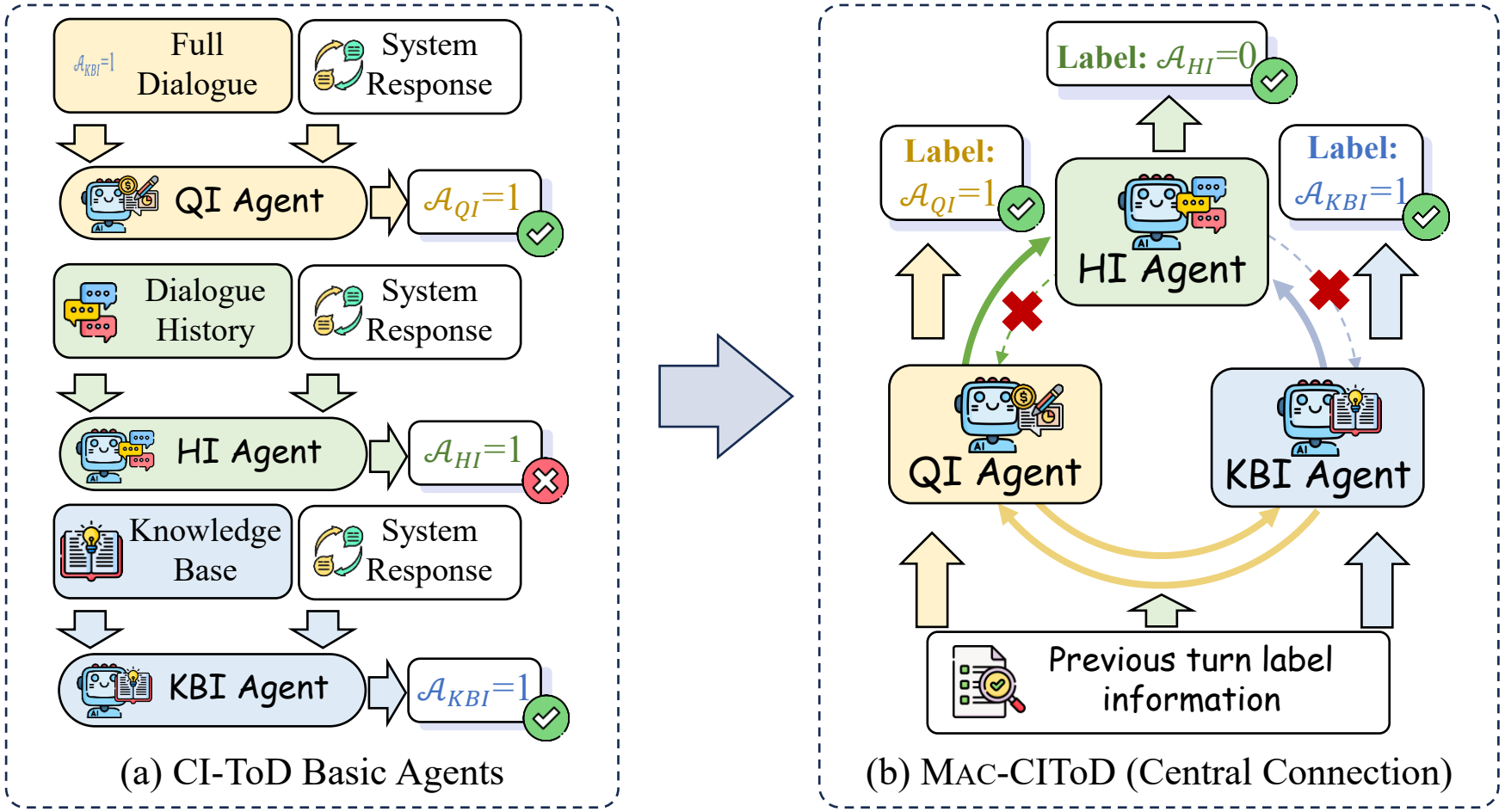
MAC-CIToD (Full Connection) assists the agent by providing all labels from CI-ToD basic agents, allowing them to reference all information comprehensively.

MAC-CIToD



MAC-CIToD (Cycle Connection) allows all neighboring agents to transmit labels in one direction of the cycle, which incorporates knowledge from multiple angles.

MAC-CIToD



MAC-CIToD (Central Connection) selects a central agent to receive all labels from the other agents; the remaining agents exchange labels with each other.

Main Results

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)**

These advanced agent methods still have a gap from the performance of traditional methods.

Main Results

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)**

Our framework attains the best performance after collaboration. Notably, MAC-CIToD on gpt-4o comprehensively surpasses the results of the best traditional method, PPA.

Main Results

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)**

Our framework can still achieve competitive performance on smaller LLMs. The performance on Llama and GLM is close to the traditional method, BERT.

MAC-CIToD Analysis

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.

MAC-CIToD Remains Robust For All Connection Paradigms.

MAC-CIToD Analysis

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.

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

MAC-CIToD Analysis

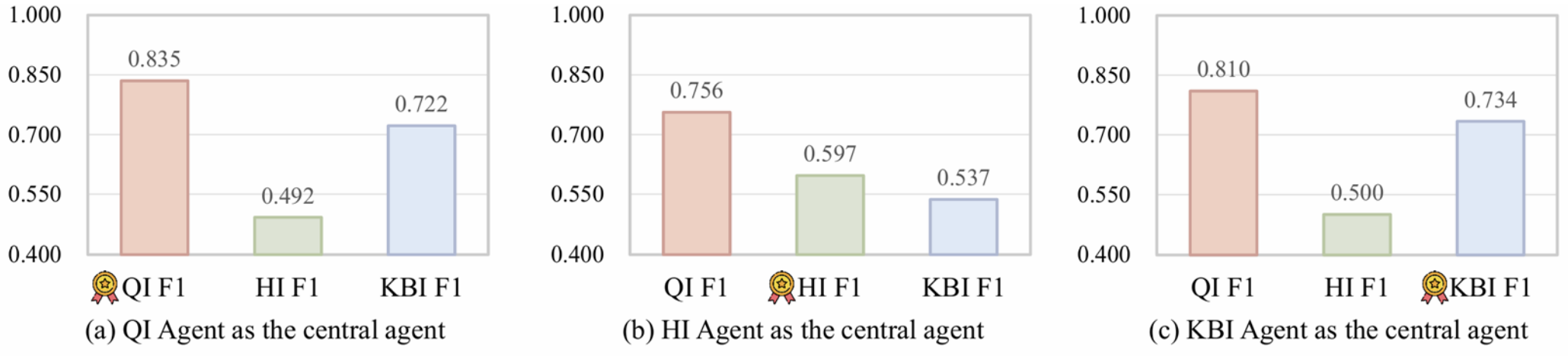


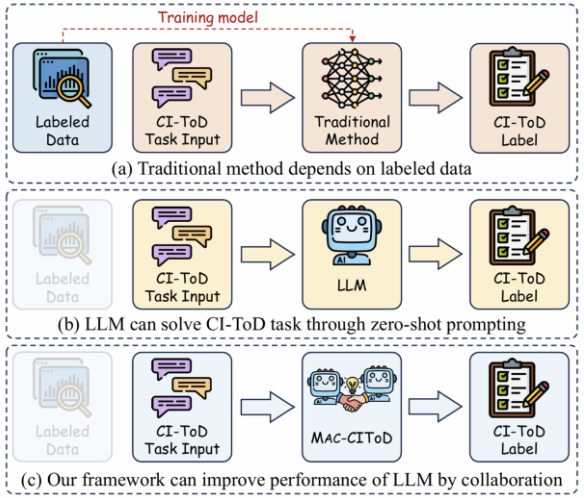
Figure 4: The results of different central agents in MAC-CIToD (Central Connection). This figure illustrates the F1 performance of MAC-CIToD (Central Connection) while QI Agent (a), HI Agent (b), and KBI Agent (c) as the central agent. The reward sign on the left side of F1 indicates that the performance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

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

Conclusion

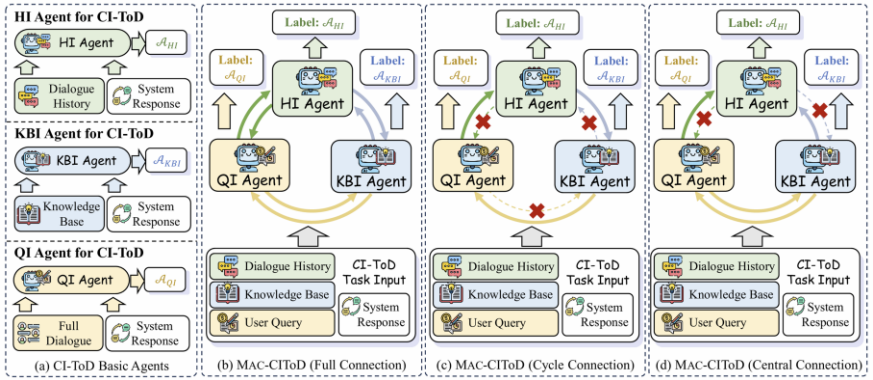
Motivation

Traditional methods depend heavily on high-quality labeded data.



Approach

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 and systematically explore three collaboration paradigms.



Result

MAC-CIToD achieves advanced performance, surpassing methods that require extensive training.

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)**

Thanks