TODRef: Agentic policy with self-reflection for task-oriented dialogue systems

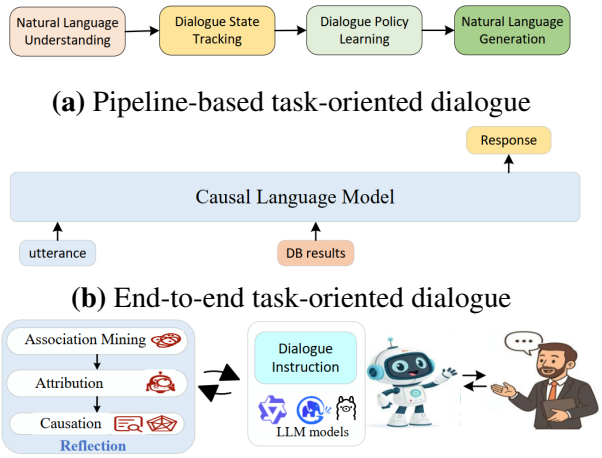
Anonymous ACL submission

Abstract

Task-oriented dialogue (TOD) systems consist of modular and end-to-end approaches. They rely heavily on extensive labeled data, struggle to adapt to evolving dialogues, and suffer cumulative errors without effective correction. These drawbacks result from insufficient intent perception and dynamic dialogue tracking, which are vital for stable task completion. To address these issues, this paper proposes a TOD method based on a reflection mechanism constructed by modeling intentions as a Markov transition process. Abandoning traditional fixed modules, the method takes a general instruction-following language model as the core. The reflection mechanism quantifies inter-intention transition probabilities via Markov process to trace intention evolution and identify deviation risks, enabling retrospective correction of historical intention misunderstandings. Without relying on annotated data, experiments on MultiWOZ and SGD benchmarks show significant improvements in intention alignment and task success rate, especially enhancing adaptability in complex multi-turn scenarios with frequent intention shifts.

1 Introduction

Task-oriented dialogue (TOD) systems act as crucial interfaces in human-computer interaction, streamlining the completion of various practical tasks. With the rising demand for intelligent services, TOD technology is being widely deployed in both general consumer scenarios and specialized vertical domains (Dong et al., 2025; Rim et al., 2025). The advancement of large language models (LLMs) has brought transformative progress to TOD research and application, endowed with massive pre-trained knowledge and superior contextual modeling capabilities (Li et al., 2025), have overcome many drawbacks of traditional systems, such as poor semantic understanding and weak multi-turn coherence. Using methods like instruction tuning and prompt learning, LLMs can quickly adapt to TOD tasks, greatly improving dialogue fluency and versatility (Chung et al., 2023; Raposo et al., 2023). However, despite these advancements, practical applications of TOD systems are still plagued by prominent limitations. These systems are not only challenged by the high reliance on large-scale annotated data, especially in low-resource domains, but also lack effective dynamic adaptation capabilities to the evolving nature of dialogues. More critically, the absence of robust error correction mechanisms often leads to cumulative errors, severely hindering the reliability of task completion.



(a) Pipeline-based task-oriented dialogue

(b) End-to-end task-oriented dialogue

(c) Self-reflection TOD agent

Figure 1: Different architectures of TOD systems, the self-reflection agent is our proposed method that can evolutionarily enhance performance.

In TOD, user intent determines the direction of dialogue interaction and the success of task completion. However, how intent tracking can be effectively leveraged to improve efficient task completion in TOD remains unclear and requires further investigation. Traditional solutions for TOD systems can be roughly divided into two categories: modular architectures (see figure 1(a)) and end-to-end frameworks (see figure 1(b)). Modular archi-

tectures typically consist of independent modules for intent recognition, dialogue state tracking, policy decision-making, and response generation, with each module performing specific tasks in a pipeline manner. This design ensures good interpretability but suffers from cumulative errors caused by module cascading and poor adaptability to cross-domain scenarios (Yi et al., 2025; Xu et al., 2025a). End-to-end frameworks, on the other hand, aim to directly map dialogue context to response outputs through a unified model (Sun et al., 2025; Medjad et al., 2025), simplifying the complex pipeline design, but they rely heavily on large-scale annotated data and lack transparency in decision-making.

In recent years, LLM-driven TOD systems have become the mainstream research direction. Researchers have developed a series of adaptation strategies to migrate general-purpose LLMs to TOD tasks. Instruction tuning (Wu et al., 2024; Vaillancourt and Thompson, 2024) guides LLMs to understand task objectives through natural language instructions, prompt learning (Wu et al., 2025; Ghazarian et al., 2025) elicits task-solving capabilities by constructing specific prompt templates, parameter-efficient fine-tuning (Feng et al., 2025; Stricker and Paroubek, 2024) achieves task adaptation by fine-tuning only a small number of model parameters, reducing computational costs.

Despite great progress, current LLM-based TOD systems still have prominent defects: inadequate goal-oriented design, defective dialogue state management and missing reflection modules, which restrict their practical adaptability. Though equipped with powerful contextual comprehension, such models cannot retrospect and rectify historical dialogue errors (Xu et al., 2025b). Once intent recognition goes wrong, they fail to locate error sources, resulting in error accumulation across multi-turn dialogues. This problem severely lowers task completion rates and user trust, especially in complex interaction scenarios. In addition, mainstream LLM adaptation approaches are highly dependent on labeled data, limiting their deployment in low-resource domains. Accordingly, it is imperative to develop reflective self-evolving TOD systems to tackle the above limitations.

To address the aforementioned problems, this paper proposes a reflection mechanism by modeling intentions as a Markov transition process (Figure 1(c) shows the framework). It utilizes a general instruction-following model integrated with the reflection mechanism to compensate for the inherent flaws of LLMs in error correction. The core innovation lies in the designed reflection mechanism: by modeling the evolution of dialogue intentions as a Markov transition process (where the current intention state depends only on the previous adjacent intention state), the mechanism quantifies the transition probabilities between different intentions. This enables the system to effectively trace the evolutionary trajectory of intentions, identify potential risks of intention deviation in a timely manner, and achieve retrospective correction of historical intention misunderstandings. Notably, the method does not rely on additional annotated data—its reflection and intention modeling capabilities are activated through instruction prompts and pre-trained knowledge, significantly reducing the data dependency typically associated with traditional TOD systems and LLM adaptation methods. The main contributions of this paper are summarized as follows:

- This paper clarifies the dominant role of dialogue intention in TOD systems and proposes a Markov transition-based intention modeling approach, offering a new perspective for dynamic intention tracking.

- Integrates the reflection mechanism with the instruction-following language model, enabling effective error correction and reducing cumulative errors in multi-turn dialogues.

- Improves the adaptability of TOD systems in complex multi-turn scenarios with frequent intention shifts, achieving excellent performance on the MultiWOZ and SGD benchmark datasets, while also being free from reliance on annotated data, thus lowering the application threshold in low-resource settings.

2 Related Work

2.1 LLM-based TOD systems

TOD systems help users complete specific tasks through multi-turn interactions, with LLMs driving performance breakthroughs via pre-trained knowledge. Lightweight transfer adaptation methods (Liu et al., 2025) (instruction tuning (Nguyen et al., 2025), prompt learning (Thakkar and Pise, 2024), Parameter-Efficient Fine-Tuning (Nguyen et al., 2025; Zhang et al., 2026)) enable low-cost adaptation of general LLMs to TOD tasks. To address LLMs' limitations in core TOD tasks, researchers

have improved state tracking, multi-turn interaction and complex scenario adaptation by integrating dialogue state tracking modules, optimizing long-range attention and building multimodal fusion architectures. Zero-shot LLMs are crucial for low-resource TOD (Xu et al., 2024; Baidya et al., 2025), as they perform tasks via natural language instructions without task-specific annotated data, reducing costs and enabling rapid adaptation to new scenarios. Notably, reinforcement learning (RL) is core to dialogue policy learning in TOD systems, determining actions based on dialogue states. Recent work combines evolutionary algorithms with deep RL and elite injection to solve exploration-exploitation imbalance in high-dimensional spaces, while early studies used end-to-end deep RL to unify state tracking and policy learning for better task success (Young et al., 2013). Integrating RL and LLMs to enhance TOD generation is a promising direction.

2.2 Reflection Mechanism

The reflection mechanism empowers agents to simulate human cognition by reviewing decision trajectories, analyzing errors and formulating corrections, thereby realizing self-correction and iterative optimization (Radanliev, 2025; Yu, 2023). In deep learning, it operates as an error feedback and parameter adjustment closed loop dominated by backpropagation. Traditional TOD systems optimize state updates by reflecting on mismatches between slot-filling errors and user intentions. Existing LLM-based reflection methods (Bo et al., 2024; Wang et al., 2025) mainly adopt prompt design and lightweight tuning to achieve efficient reflection and optimization. This mechanism equips TOD systems with error recovery ability, mitigates cumulative errors in multi-turn dialogues and boosts system reliability.

Our designed self-reflection agent possesses three merits. It traces dialogue histories to detect and fix intent deviation, missing slots and factual mistakes. It leverages chain-of-thought reasoning (Wei et al., 2022; Xia et al., 2025) via extended retrospective paths to refine defective logic. It also supports dynamic policy adaptation, adjusting response strategies by reviewing historical interactions amid task shifts and new user requirements.

3 Self-reflection TOD agent

In this work, we propose a zero-shot TOD-based framework based on a reflective mechanism. Unlike traditional pipeline or end-to-end TOD systems, our model efficiently and accurately completes specific tasks solely through instructions, thereby advancing control-based generation in TOD systems. We extract user intentions from interaction trajectories and establish a dynamic intention-dependent Markov decision process, solving for the optimal policy through reinforcement learning. By utilizing this policy for task-oriented goal planning, we enhance the controllability of LLMs in TOD systems. Figure 2 illustrates the model structure proposed in this paper, which consists of three main components: intention extraction, intention relation mining, and decision making. The following sections provide a detailed introduction to each of these components.

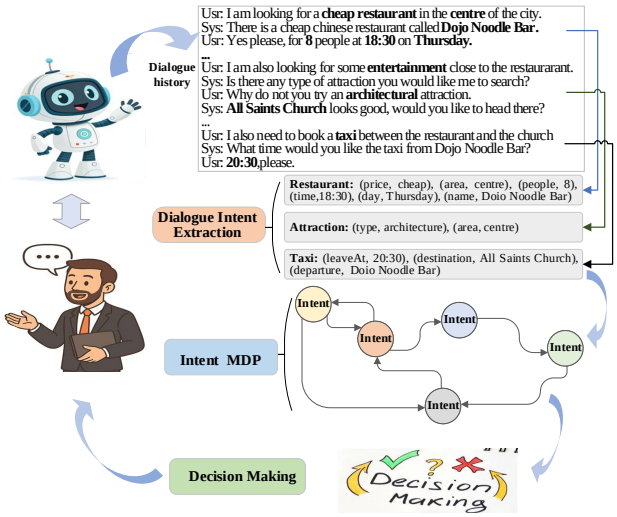


Figure 2: The framework of proposed self-reflection TOD agent.

3.1 Intent Recognition

Intent recognition serves as a fundamental prerequisite for ensuring that the TOD system accurately comprehends user requests and efficiently completes designated tasks. In this paper, an intent comprises two components: request slots and information slots. Request slots form the core of the intent, specifying the concrete task actions required by the user, such as 'inquire,' 'book,' 'modify,' or 'cancel.' Information slots represent essential attribute parameters for task execution, such as 'departure city' and 'travel date' for flight inquiries, or 'check-in time' for hotel

bookings. These two elements are interdependent, collectively constituting a complete semantic unit of intent. The intent can be represented as: $intent = [request : [slot, \ldots, slot], inform : slot : value, \ldots, slot : value]$.

Traditional multi-intent recognition and slot-filling techniques for TOD heavily rely on supervised deep learning models. Such approaches require extensive annotated data and demonstrate weak cross-domain generalization capabilities. The LLM prompt engineering methodology enables LLMs to perform multi-intent recognition and slot-filling tasks in zero-shot scenarios by designing precise prompting strategies. This significantly reduces the dependency on annotations while enhancing cross-domain adaptability. Therefore, in this paper, we employ prompt engineering to mine intents and slots. By embedding a small number of annotated domain examples within the prompts, we enable the LLM to learn the correspondence between intents and slot positions. For specific domains, domain-specific slot information is also incorporated as part of the prompt.

3.2 Intent Relationship Mining

Intent recognition involves parsing user utterances to extract core information. In TOD systems, a dialogue success is only when the agent fulfills all user requests. This paper establishes a Markov Decision Process (MDP) for intent transfer by mining user intent from historical dialogues. First, the state space S is defined, which must satisfy the Markov property. Consequently, states are characterized as multidimensional tuples that include: the set of currently recognized user intents, the slot-filling status for each intent, and the current dialogue turn.

The action space A is defined as the set of dialogue strategies that the system can execute in the current state. In alignment with the logic of dialogue interactions, the action space comprises four categories:

Query Slots: Initiate user questioning to obtain missing key slots for the current intent.

Intent Clarification: Prompt the user to clarify their actual requirements when user input exhibits ambiguous intent or unclear expression.

Task Execution: Trigger the business operation associated with the intent when all key slots corresponding to the current intent are fully populated.

Rejection/Transfer to Human Agent: Terminate the automated dialogue and guide the user to a human agent when the user's intent cannot be recognized, exceeds the system's service scope, or when key slot information cannot be obtained even after multiple follow-up questions.

The state transition probability P describes the likelihood of transitioning from the current state s_t to the next state s_{t+1} after executing action a_t. Its core calculation relies on statistical patterns derived from historical dialogue corpora. First, extract all triplet trajectories $s_t \rightarrow a_t \rightarrow s_{t+1}$ from historical dialogue data. For each (s_t, a_t) combination, count the number of transitions to different s_{t+1} states. Initial transition probabilities are calculated based on frequency proportions using the formula:

$$P(s_{t+1}|s_t, a_t) = \frac{\text{count}(s_t, a_t, s_{t+1}) + \lambda}{\sum_{s' \in S} \text{count}(s_t, a_t, s') + \lambda \cdot K}, \tag{1}$$

where λ is the Laplace smoothing coefficient, and K denotes the number of observable states in the historical dialogue corpus. For transition trajectories that are absent or infrequent in the historical data, techniques such as Laplace smoothing must be employed to prevent transition probabilities from becoming zero: using observable states K (instead of the potentially enormous $|S|$) avoids the difficulty of acquiring the full state space, while still ensuring the corrected probabilities satisfy the probability distribution constraint and maintaining the model's generalization capability.

3.3 Decision Making

Upon completion of the MDP construction, the core objective is to learn the optimal policy $\pi^*(a|s)$. This paper introduces confidence upper bounds to optimise the policy, addressing the uncertainty inherent in intent recognition. This approach ensures that rewards depend not only on the success or failure of the dialogue but also reflect the certainty of intent recognition and slot filling, thereby enhancing the rationality and robustness of policy learning. The definition of confidence relies on the outcomes of intent recognition and slot filling, establishing the probability Conf_I for LLM intent recognition confidence and Conf_F for slot filling confidence. Thus, a composite confidence $\text{Conf} = \alpha \cdot \text{Conf}_I + (1 - \alpha) \cdot \text{Conf}_F$ is constructed, where $\alpha \in [0, 1]$ is a weighting coefficient adjustable according to the relative importance of intent recognition and slot filling in the task context. This composite confidence is used to modify the reward.

The reward function modified with confi-

dence is formulated as $R_{\text{conf}}(s_t, a_t, s_{t+1}) = R(s_t, a_t, s_{t+1}) \cdot \text{Conf}$. This paper employs the PPO algorithm for policy optimisation. The agent learns policies based on structure-based intent transition data $\tau = \{s_0, a_0, r_{\text{conf},0}, s_1, a_1, r_{\text{conf},1}, ..., s_T, a_T, r_{\text{conf},T}\}$ to learn the policy, where r_{conf} denotes the reward adjusted by confidence, and the composite confidence at step t is denoted as Conf_t. Secondly, state representation enhancement is employed by concatenating the composite confidence Conf_t with the original state s_t as input to the Actor-Critic network (i.e., input $s'_t = [s_t, \text{Conf}_t]$). This enables the network to simultaneously perceive dialogue context and decision certainty, thereby enhancing the precision of value assessment and action selection. The computation uses generalized advantage estimation. It combines the critic's state-value estimate $V(s'_t)$ with the corrected reward $r_{\text{conf},t}$ to compute the action advantage A_t.

To implement the policy optimization based on the confidence-modified reward and enhanced state representation, we adopt the PPO algorithm, and the specific policy update process is detailed as follows. During the policy update phase, we construct the PPO clipped objective function to ensure stable policy iteration:

$$L(\theta) = \mathbf{E}_t \left[\min \left(\frac{\pi_\theta(a_t|s'_t)}{\pi_{\theta_{old}}(a_t|s'_t)} A_t, clip\left(\frac{\pi_\theta(a_t|s'_t)}{\pi_{\theta_{old}}(a_t|s'_t)}, 1-\epsilon, 1+\epsilon \right) A_t \right) \right], \tag{2}$$

where θ_{old} is the pre-update policy parameter, and ϵ is the clipping coefficient to constrain the old-new policy probability ratio for stable updates. We also optimize the Critic network's value loss $L_V(\theta_v) = \mathbf{E}_t \left[(V(s'_t) - \hat{V}_t)^2 \right]$, where $\hat{V}_t$ is the target value from the corrected reward. Alternately optimizing the Actor's policy and Critic's value loss until convergence yields the dialogue-specific optimal policy π^*, which serves as prompts to guide LLM response generation. Algorithm 1 presents the model's pseudocode.

Algorithm 1 Self-reflection TOD Agent

Require: Input query q, initial policy model π, database DB used for querying, maximum dialogue turn T, dialogue intent history H

Output: Response y

```
Initialize response y ← ϕ, dialogue turn t ← 0
while t < T do
    Initialize curren rollout y' ← ϕ
    Perform intent analysis on user utterances
    Query the database based on slots and slot values to obtain query result res
    Record the intent transition in H
    Perform policy analysis based on H to derive next action prompt act
    Sample next output y_{t+1} based on dialogue history and action prompt y_{t+1} ~ LLM(·|x, y', res, act)
    y' = y_t + y_{t+1}
    if res == None then
        break
    end if
end while
return Response y_{t+1}.
```

4 Experiments

The experiments in this paper were implemented on two widely used TOD datasets. MultiWOZ (Budzianowski et al., 2018) and SGD (Rastogi et al., 2020). MultiWOZ is a fully crowdsourced, manually annotated multi-domain, multi-turn dialogue dataset. It covers seven relevant domains: hotels, restaurants, taxis, attractions, trains, hospitals, and police stations. Conversations frequently involve switching between and combining multiple domains (e.g., first booking a hotel, then finding a nearby restaurant, and finally hailing a taxi). Each dialogue averages approximately 14 turns and encompasses rich conversational actions such as requesting information, providing information, affirming, negating, and domain switching. SGD is the *Schema-Guided Dialogue* dataset released by Google. Unlike MultiWOZ, most of its dialogues are automatically generated by advanced dialogue simulators based on predefined schemas, followed by human polishing and quality control. Each domain within SGD is defined by an independent schema file. MultiWOZ simulates a complex yet closed world, testing models' nuanced operations within this environment. SGD simulates an ever-expanding open world, testing models' meta-capabilities to rapidly learn and integrate new services APIs. The experiment was conducted on the mainstream GPT-3.5 (gpt-3.5-turbo-0613). We constructed prompts by extracting user objectives and building interaction items, utilising the GPT-3.5 model to establish a user simulator. This simulator interacted with our proposed TODRef to obtain the final output. All models employed a greedy decoding strategy. This paper evaluates the proposed

model on the TOD datasets using three key metrics: Inform, Success, and Book.

Inform: This metric verifies whether the dialogue agent correctly matches entities to the user's intent. A general pre-trained language model extracts entity names (dialogue intents) from the dialogue text, and complete entity attributes are retrieved from the database based on these names. The metric is considered met by comparing the constraints within the user's goal against the attributes of the system's recommended entities, assessing compliance by their degree of alignment.

Success: This metric assesses whether the system provides all required entity attributes to the user. It is calculated by extracting the attributes already provided by the system from the dialogue using a pre-trained language model. The Success metric is deemed met only when all three conditions are satisfied: The matched entity is correct (i.e., the Inform metric criteria are satisfied). The system has provided all entity attributes required by the user. All attribute values are accurate and error-free.

Book: A supplementary metric for MultiWOZ (where users often book after locating venues), assessing booking success rate. It uses a pre-trained LLM to extract the system's booking reference number, retrieve corresponding database records, and match them with user booking constraints; the matching score is the metric result. We also define a Combine score as the average of all three metrics.

4.1 Baselines

The models compared in this paper include: **SimpleTOD** (Hosseini-Asl et al., 2020): A unified generative model based on GPT-2, which unifies user understanding, action decision and response generation into a single sequence prediction task.
UBAR (Yang et al., 2021): It is fine-tuned on complete dialogue trajectories including utterances, belief states, database results and system responses. It adopts generated content as dialogue context to ensure output coherence. **GALAXY** (He et al., 2022): A pre-trained TOD model learning dialogue policies via semi-supervised learning. It leverages dialogue act prediction, consistency regularization and sample gating to fully exploit unlabeled dialogue data. **Mars** (Sun et al., 2022): An end-to-end TOD model adopting dual contrastive learning to model contextual semantic relations, so as to optimize belief-action states and improve multi-turn dialogue performance. **TOATOD** (Bang et al., 2023): It applies task-specific adapters to optimize NLU, DST and NLG subtasks while freezing pre-trained backbones, and further boosts performance via metric-aware reinforcement learning. **AutoTOD** (Xu et al., 2024): A zero-shot TOD agent built on instruction-following LLMs. It adopts concise instruction templates to realize turn-level dialogue decision-making and breaks the constraints of traditional modular frameworks.

4.2 Main Results

In this experiment, seven TOD models are multi-dimensionally evaluated on MultiWOZ and SGD benchmarks, with results listed in Table 1. SimpleTOD obtains 32.5 Inform and 20.6 Combine scores on MultiWOZ without Book task support, showing the poorest baseline performance. UBAR fine-tunes GPT-2 on complete dialogue sequences to boost consistency, reaching 40.8 Inform and 24.7 Combine yet restricted by GPT-2's inherent capacity. GALAXY adopts semi-supervised learning to mine dialogue strategies from unlabeled data and hits 44.4 Inform and 26.5 Combine, while its modular design is confined by conventional pre-training schemes. Mars improves belief and action states via dual contrastive learning, attaining a higher Success score of 35.4 than GALAXY, yet its overemphasis on context-state correlation leads to a slightly lower Combine score of 26.0. TOATOD leverages task adapters for subtask-specific learning and reinforcement learning optimization, achieving 45.3 Inform and 27.3 Combine, ranking first among conventional TOD architectures. Built on instruction-tuned LLMs with zero-shot decision mechanisms, AutoTOD achieves superior results including 62.5 Inform, 51.4 Book task score and 55.5 Combine. Benefiting from powerful instruction following, it outperforms traditional models remarkably and possesses great optimization potential.

The TODRef model achieves comprehensive superiority over all existing models across all metrics. On MultiWOZ, it attains the highest scores in Inform (65.6), Success (54.3), Book (52.8), and Combine (57.6), and its performance on the SGD dataset also far surpasses that of other models. This outcome benefits from TODRef's integration of the strengths of existing models, coupled with further enhancements to LLM instruction adaptability and fine-grained task processing capabilities. TODRef not only retains the flexible zero-shot decision-making characteristics of AutoTOD but

Model	MultiWOZ				SGD		
	Inform	Success	Book	Combine	Inform	Success	Combine
SimpleTOD	32.5	29.4	–	20.6	13.1	10.4	11.8
UBAR	40.8	33.3	–	24.7	18.0	15.1	16.5
GALAXY	44.4	35.1	–	26.5	20.1	17.5	18.6
Mars	42.7	35.4	–	26.0	19.6	16.5	18.1
TOATOD	45.3	36.7	–	27.3	21.7	17.9	19.8
AutoTOD	62.5	52.7	51.4	55.5	35.0	20.0	27.5
TODRef	**65.6**	**54.3**	**52.8**	**57.6**	**49.2**	**24.8**	**37.0**

Table 1: Task completion results of various models on MultiWOZ and SGD. − indicates the model lacks reservation capability.

also compensates for its deficiencies in task completion accuracy, ultimately achieving both superior performance and strong adaptability in TOD tasks.

This paper evaluates the linguistic diversity of MultiWOZ, with metrics and results presented in Table 2. Restricted by small model scale and single-sequence modeling, SimpleTOD scores the lowest in all diversity indicators with inadequate linguistic richness. UBAR slightly boosts diversity via GPT-2 dialogue-level fine-tuning and dynamic output adjustment, still limited by inherent model defects. GALAXY improves lexical diversity using semi-supervised learning and gated sample screening. Mars enriches dialogue outputs by differentiating context and semantic states via contrastive learning. TOATOD optimizes subtask generation with task adapters and reinforcement learning to elevate diversity performance (SE=7.65, MSTTR=68.23). Equipped with powerful universal semantic comprehension, AutoTOD attains prominent diversity gains (SE=8.11, MTLD=65.09). Built on AutoTOD, the proposed TODRef further optimizes instruction adaptation and fine-grained task processing, reaching the best overall diversity performance and outperforming AutoTOD in linguistic richness and lexical diversity.

Model	SE	CE	MSTTR	MTLD	HDD
SimpleTOD	7.13	2.00	59.32	31.28	76.91
UBAR	7.20	2.05	61.21	34.33	77.21
GALAXY	7.45	2.16	62.55	34.19	80.23
Mars	7.61	2.14	64.98	40.15	83.43
TOATOD	7.65	2.20	68.23	42.84	80.25
AutoTOD	8.11	2.62	76.61	65.09	86.07
TodRef	**8.12**	**2.63**	**76.66**	**65.30**	**87.45**

Table 2: Language diversity evaluation on MultiWOZ. SE, CE, MSTTR, MTLD and HDD stand for Shannon Entropy, Conditional bigram Entropy, Mean Segmental Type-Token Ratio, Measure of Textual Lexical Diversity, and Hypergeometric Distribution Function separately.

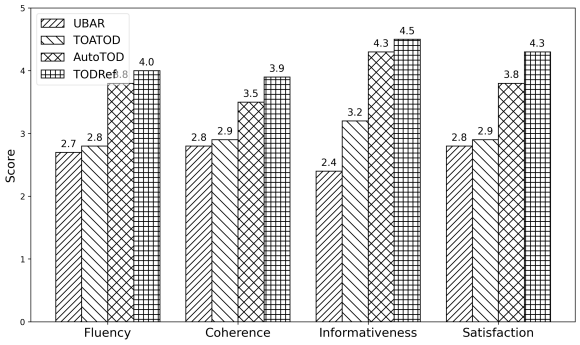
Table 3 presents the language diversity evaluation results for the SGD dataset. SimpleTOD's diversity metrics rank lowest among the comparison methods UBAR, GALAX, Mars, and AutoTOD progressively improve these metrics, with SE rising from 7.06 to 7.67 and MSTTR increasing from 71.33 to 73.18. However, traditional architectures demonstrate limited adaptability to complex dialogue scenarios, yielding far smaller metric gains than large language model-based approaches. Large language model-driven architectures achieved a leap in diversity metrics. AutoTOD, leveraging its generalised semantic understanding, saw SE directly jump to 8.44, with all metrics substantially surpassing traditional models. The proposed TODRef further optimises instruction adaptation and fine-grained generation logic for complex scenarios based on AutoTOD, achieving optimal performance across all diversity metrics on the SGD dataset. This approach not only preserves the generalisation advantages of large models in handling complex dialogues but also achieves further enhancements in linguistic richness and lexical diversity, demonstrating its superiority in language generation within highly complex TOD scenarios.

Model	SE	CE	MSTTR	MTLD	HDD
SimpleTOD	6.88	1.95	69.22	51.44	76.19
UBAR	7.06	1.98	71.33	52.12	79.17
GALAXY	7.12	2.01	72.14	52.98	79.86
Mars	7.45	2.17	72.76	54.39	80.26
TOATOD	7.67	2.21	73.18	55.64	81.09
AutoTOD	8.44	2.45	73.87	58.91	86.87
TodRef	**8.56**	**2.61**	**74.38**	**69.60**	**87.56**

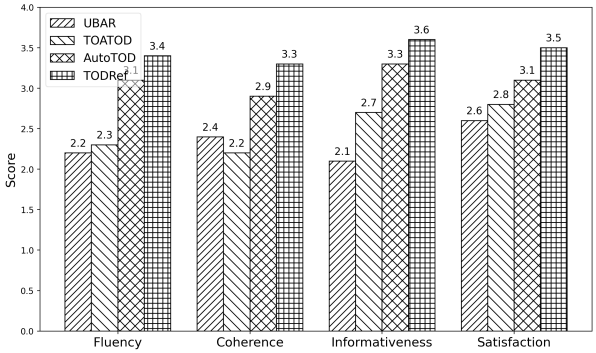
Table 3: Language diversity evaluation results on SGD.

4.3 Human Evaluation

To objectively evaluate dialogue quality, we recruited 20 evaluators (10 students and 10 professionals), each receiving a fountain pen as compensation.Model-generated dialogues on MultiWOZ were randomly shuffled, and evaluators independently scored each sample across fluency, coherence, informativeness, and satisfaction using a 5-point scale.Figure 3(a) shows the results. Traditional models such as UBAR and TOATOD achieved scores of 2.4–2.9, limited by naturalness, coherence, and information efficiency.In contrast, AutoTOD, benefiting from LLM instruction-following abilities, scored above 3.5 on all dimen-



(a) Results on the MultiWOZ dataset



(b) Results on the SGD dataset

Figure 3: Human evaluation results.

sions, demonstrating clear improvements in dialogue quality. Notably, TODRef, the model proposed in this paper, surpassed all others in every evaluation dimension, achieving a fluency score of 4.0 and an informativeness score of 4.5, with higher coherence and satisfaction scores than AutoTOD. This demonstrates that TODRef balances language naturalness, logical coherence, information effectiveness, and user experience more effectively, establishing it as the model with the best overall performance in this evaluation.

Figure 3(b) displays results from the SGD dataset. Traditional models like UBAR and TOATOD scored between 2.1 and 2.8, reflecting their struggles with the flexible dialogue flow of SGD and amplified shortcomings in fluency and coherence. In contrast, AutoTOD leveraged LLM generalization to achieve scores of 2.9–3.3, demonstrating basic adaptability to complex scenarios.Notably, TODRef outperformed all models, achieving a fluency score of 3.4, an informativeness score of 3.6, and a satisfaction score of 3.5. It outperforms traditional models and exceeds AutoTOD. With optimized instruction adaptation and fine-grained dialogue processing, TODRef adapts well to flexible SGD dialogue flows and achieves better linguistic naturalness, logical consistency and user experience.

4.4 Dialogue Example

Figure 4 presents a dialogue example for the proposed TODRef model. The user's goal remains unknown to TODRef. Based on the user's query 'I'm looking for entertainment venues in the town centre', it analyses the user's latent intent, queries the database according to that intent, and obtains the query results. Subsequently, it reflects on the previous interaction history and formulates a new query. The TODRef model employs reflection to generate responses that better align with the user's objective, rather than directly requesting information from the user. This approach renders the dialogue more human-like.

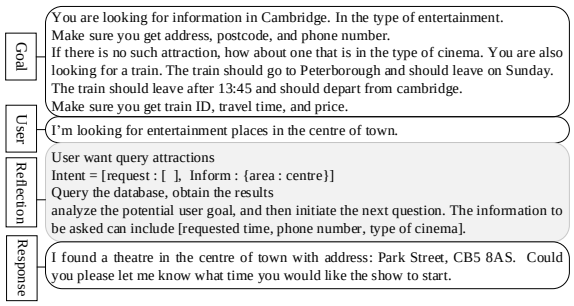


Figure 4: Example of Goal, User Input, Reflection, and System Response.

5 Conclusion and Limitations

This paper introduces TODRef, a task-oriented dialogue autonomous agent based on reflective mechanisms and API invocation. Unlike traditional pipeline and end-to-end approaches, it employs reinforcement learning to construct sequential decision-making for TOD systems, establishing goal-directed policies. Through instruction fine-tuning combined with policy guidance, it directs large language models to generate final task-oriented dialogue outputs. The system enables continuous reflection, fostering evolutionary decision-making. Experimental results validate the advanced performance of the proposed model.

TODRef has several limitations. First, constructing MDP via reinforcement learning and intent mining introduces extra computational overhead. Second, the evaluation should cover more comprehensive analyses of reflection mechanisms and intent transfer mining. Finally, testing on more LLMs would help readers better assess the effectiveness of our method.

References

Avinash Baidya, Kamalika Das, and Xiang Gao. 2025. The behavior gap: Evaluating zero-shot llm agents in complex task-oriented dialogs. *arXiv preprint arXiv:2506.12266*.

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. *arXiv preprint arXiv:2305.02468*.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. *Advances in Neural Information Processing Systems*, 37:138595–138631.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. Instructtods: Large language models for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2310.08885*.

Wenjie Dong, Sirong Chen, and Yan Yang. 2025. Protod: Proactive task-oriented dialogue system based on large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9147–9164.

Shutong Feng, Hsien-chin Lin, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2025. Emotionally intelligent task-oriented dialogue systems: Architecture, representation, and optimisation. *arXiv preprint arXiv:2507.01594*.

Sarik Ghazarian, Abhinav Gullapalli, Swair Shah, Anurag Beniwal, Nanyun Peng, Narayanan Sadagopan, and Zhou Yu. 2025. Tod-procbench: Benchmarking complex instruction-following in task-oriented dialogues. *arXiv preprint arXiv:2511.15976*.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, and 1 others. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10749–10757.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, and 1 others. 2025. Fundamental capabilities and applications of large language models: A survey. *ACM Computing Surveys*.

Chuanhong Liu, Caili Guo, Yang Yang, Mingzhe Chen, and Tony QS Quek. 2025. Lightweight task-oriented semantic communication empowered by large-scale ai models. *IEEE Transactions on Vehicular Technology*.

Maya Medjad, Hugo Imbert, Bruno Yun, Raphaël Szymocha, and Frédéric Armetta. 2025. Leveraging graph structures and large language models for end-to-end synthetic task-oriented dialogues. *arXiv preprint arXiv:2501.11977*.

Vinh Quang Nguyen, Nguyen Quang Chieu, Hoang Viet Pham, and Khac-Hoai Nam Bui. 2025. Spec-tod: A specialized instruction-tuned llm framework for efficient task-oriented dialogue systems. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 133–145.

Petar Radanliev. 2025. Artificial intelligence: reflecting on the past and looking towards the next paradigm shift. *Journal of Experimental & Theoretical Artificial Intelligence*, 37(7):1045–1062.

Gonçalo Raposo, Luisa Coheur, and Bruno Martins. 2023. Prompting, retrieval, training: An exploration of different approaches for task-oriented dialogue generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–412.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.

Daniel Rim, Minsoo Cho, Changwoo Chun, and Jaegul Choo. 2025. To chat or task: a multi-turn dialogue generation framework for task-oriented dialogue systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 576–592.

Armand Stricker and Patrick Paroubek. 2024. A few-shot approach to task-oriented dialogue enhanced with chitchat. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 590–602.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022. Mars: Semantic-aware contrastive learning for end-to-end task-oriented dialog. *arXiv preprint arXiv:2210.08917*.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2025. Comet: Dialog context fusion mechanism for end-to-end task-oriented dialog with multi-task learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10541–10553.

Manisha Thakkar and Nitin N Pise. 2024. A unified framework for efficient task-oriented dialogue system using prompts. In *International Conference on Communication and Computational Technologies*, pages 157–177. Springer.

Emily Vaillancourt and Christopher Thompson. 2024. Instruction tuning on large language models to improve reasoning performance. *Authorea Preprints*.

Ziyan Wang, Yingpeng Du, Zhu Sun, Haoyan Chua, Kaidong Feng, Wenya Wang, and Jie Zhang. 2025. Re2llm: reflective reinforcement large language model for session-based recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12827–12835.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Di Wu, Zhizhi Yu, and Yuheng Li. 2025. Cat-cpr-tod: A framework for context-aware training with clue-guided prompt reasoning in task-oriented dialogue. In *International Conference on Intelligent Computing*, pages 174–185. Springer.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2024. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2025. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10795–10809.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and He-Yan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763.

Kai Xu, Zhenyu Wang, Yuxuan Long, and Rui Zhang. 2025a. Personalized dialogue policy learning framework based on implicit user profiles. *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, page Early Access.

Yang Xu, Xuanming Zhang, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, and Sharon Li. 2025b. Simulating and understanding deceptive behaviors in long-horizon interactions. *arXiv preprint arXiv:2510.03999*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14230–14238.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 58(6):1–38.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Hao Yu. 2023. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712.

Shuyu Zhang, Yujie Liu, Xinru Wang, Cheng Zhang, Yanmin Zhu, and Bin Li. 2026. Darwintod: Llm driven lifelong self evolution for task oriented dialog systems. *arXiv preprint arXiv:2601.07248*.