From Turns to Dialogues: Rethinking Task-Oriented Dialogue Evaluation by Combining Turn-Level Precision with Dialogue-Level Comparisons

Anonymous ACL submission

Abstract

Task-oriented dialogue (TOD) systems are experiencing a revolution driven by Large Language Models (LLMs), yet the evaluation methodologies for these systems remain insufficient for their growing sophistication. While traditional automatic metrics effectively assessed earlier modular systems, they focus solely on the dialogue level and cannot detect critical intermediate errors that can arise during user-agent interactions. In this paper, we introduce TD-EVAL (**T**urn and **D**ialogue-level **Eval**uation), a two-step framework that unifies fine-grained turn-level analysis with holistic dialogue-level comparisons. At turn-level, we assess each response along three TOD-specific dimensions: conversation cohesion, backend knowledge consistency, and policy compliance. Meanwhile, we design Conversational Task Completion Agent Arena that uses pairwise comparisons to provide a measure of dialogue-level quality. Through experiments on MultiWOZ and τ-Bench, we demonstrate that TD-EVAL effectively identifies the conversational errors that conventional metrics miss. Furthermore, TD-EVAL achieves strong alignment with human judgments while remaining simple to integrate into LLM-based agent pipelines. These findings demonstrate that TD-EVAL offers a new paradigm for the evaluation of dialogue systems, providing a comprehensive assessment across both turn and system levels to capture the full spectrum of TOD systems.

1 Introduction

Task-oriented dialogue (TOD) systems are conversational agents that help users complete specific tasks such as booking hotels, ordering food, or scheduling appointments. Advances in artificial intelligence and large language models (LLM) have significantly transformed these systems, enhancing their capabilities and flexibility (Hudeček and Dusek, 2023; Xu et al., 2024). However, evaluating their true conversational capabilities remains a challenge (Nekvinda and Dušek, 2021). Although human evaluation serves as the gold standard for evaluating dialogue systems, conducting tests with real users is time-consuming, costly, and difficult to scale across multiple systems and iterations. This limitation creates a significant gap in measuring and ensuring accountability in TOD systems research.

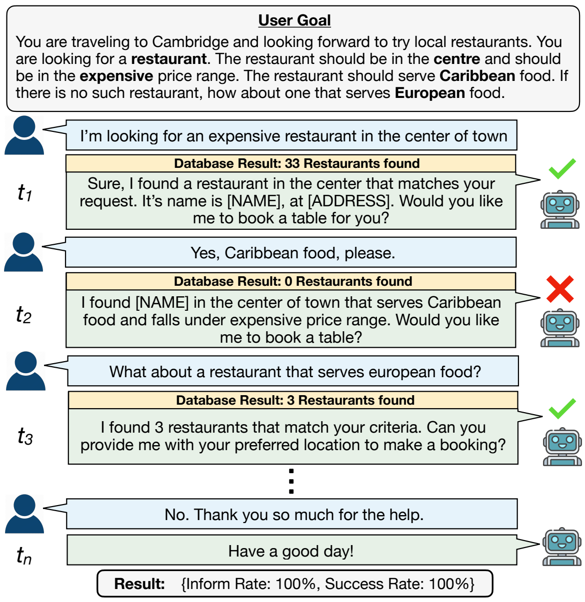

Figure 1: Example of a conversation between a user and an LLM-based agent on MultiWOZ 2.4. In the second turn (t_2), the agent recommends a restaurant even though the backend results are empty. However, in t_3, the restaurant entities are correctly detected, overwriting the previous error. Commonly used evaluation metrics in TOD, Inform and Success, fail to detect this first turn-level error and assign a perfect, 100% score.

To automate TOD evaluation, previous work has converged to various offline metrics. For example, the BLEU score (Zang et al., 2020; Papineni et al., 2002) is used to estimate the quality of the generated response (Si et al., 2024) while metrics like *Inform* and *Success* rates are utilized to estimate task completion (Budzianowski et al., 2018). Traditionally, the overall performance of the system is evaluated at a dialogue level by combining BLEU, *Inform*, and *Success* (Budzianowski et al., 2018), ignoring turn-level performance. Although dialogue understanding metrics like Joint Goal Accuracy (JGA) and slot accuracy are computed at turn-level to evaluate the correctness of predicted dialogue states, they do not measure the quality of the system's responses directly.

In typical TOD systems, each user turn is processed by converting the dialogue context into a query to a backend database. The system then decides on the next response by combining these database results with the user's request. However, as discussed, commonly used TOD metrics only capture the summative outcome of a dialogue and may not penalize intermediate system misdirections. For example, in Figure 1, the system hallucinates in the second turn (t_2) and misinforms the user, suggesting that there is a restaurant that matches the user's request despite the empty response from the database. However, the mistake is overwritten in subsequent turns, resulting in 100% Inform and Success (indicating perfect task completion), despite an intermediate step being completely wrong. On the other hand, task completion rely on delexicalizing responses by replacing entity attributes with placeholders (e.g., [NAME], [ADDRESS]) (Budzianowski et al., 2018). This obscures critical errors (e.g., incorrect hotel names) while preserving the slot structure. Moreover, classic turn-level evaluations often depend on manual annotation, which is both costly and time-consuming. As underscored by recent studies (Li et al., 2024; Xu et al., 2024), more realistic and comprehensive evaluation approaches are required in the LLM era to capture both intermediate correctness and overall response quality, exposing how current automatic metrics often miss critical errors in TOD.

In this work, we propose TD-EVAL (**T**urn and **D**ialogue-level **Eval**uation), an easy-to-use evaluation framework that combines turn-level performance with overall dialogue-level response comparisons as illustrated in Figure 3. By integrating both levels, TD-EVAL enables local error analysis for troubleshooting and global model-to-model comparisons for reliable performance benchmarking, while maintaining strong alignment with human evaluations. TD-EVAL introduces three turn-level metrics: *conversation cohesion*, *backend-knowledge consistency*, and *policy compliance*, evaluated using an LLM judge *together with justifications*. These metrics help identify subtle errors often missed by traditional automatic evaluation methods. The LLM judge model is flexible and can be easily configured using any open-source or proprietary model, with GPT-4o as the default. The proposed framework complements the per-turn judgments with a dialogue-level comparison method: a *Conversational Task Completion Agent Arena* (CTCA Arena), where entire dialogues from competing agents are systematically ranked via an Elo-based pairwise evaluation. Unlike general-purpose chatbot arenas (Zheng et al., 2023) that focus on open-domain or chit-chat dialogues, CTCA emphasizes task-oriented interactions requiring domain-specific database integration, measuring success by the agent's ability to fulfill service-oriented goals rather than linguistic fluency.

The main contributions of our work are summarized as follows:

- We propose TD-EVAL, a two-level framework that combines three turn-level metrics with pairwise comparisons at the dialogue level for the holistic evaluation of TOD, while requiring no labeled annotations.
- Our framework uses LLM-based judging, capturing domain-specific errors overlooked by standard automatic metrics (e.g., Inform, Success).
- We demonstrate the effectiveness of TD-EVAL on MultiWOZ 2.4 and τ-Bench, as validated by human evaluations, which uncover critical errors overlooked by the automatic metrics used in these benchmarks.
- We perform a comprehensive evaluation of 10 state-of-the-art LLMs on TD-EVAL and provide insights into their strengths and limitations, together with a TD-EVAL Leaderboard.

We publicly release our evaluation framework, system-response data, and human evaluations to foster future TOD research and reproducibility.

2 Preliminaries

While various aspects can contribute to the *satisfactory* evaluation dialogue systems, we focus on task completion scenarios, reviewing **automatic metrics** used in TOD and analyzing their limitations in the context of modern dialogue systems.

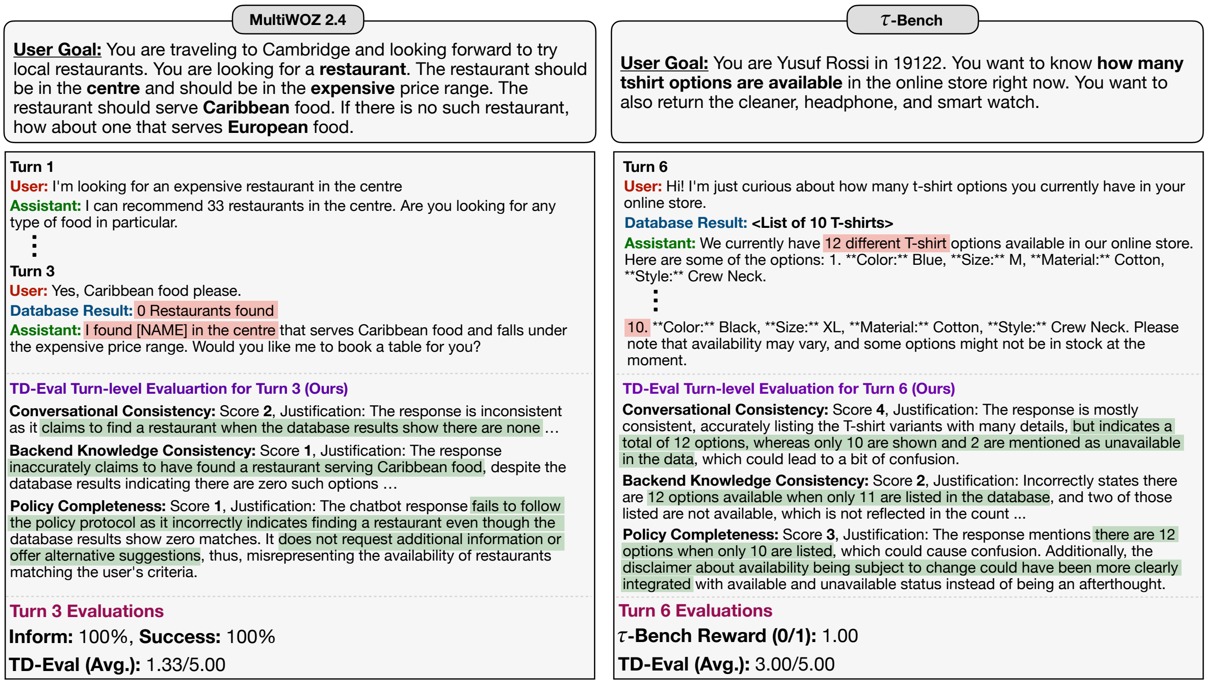

Figure 2: **Comparison of TD-EVAL with Automatic Evaluations. Left:** In MultiWOZ 2.4, a non-existent restaurant is suggested but received a 100% score from Inform and Success, while TD-EVAL's turn-level analysis assigns a lower score. **Right:** On τ-Bench, the agent claims 12 T-shirts (only 10 exist), yet string matching treats it as correct; TD-EVAL flags this missmatch. We highlight errors in red and correct elements in green.

2.1 Automatic Metrics in TOD Evaluation

Traditionally, the responses generated by TOD systems are evaluated primarily for task completion. MultiWOZ (Budzianowski et al., 2018) contains two automatic task completion metrics, **Inform** and **Success** (Budiwati et al., 2021). In τ-Bench (Yao et al., 2024), authors defined a binary **Reward** metric that checks the action and output correctness.

Inform and Success. The Inform rate measures the system's ability to provide correct entities from database search results in response to user requests. Given ground-truth entities S_d and retrieved entities S'_d for domain d, Inform is 1 if $S'_d \subseteq S_d$ for all $d \in D$; otherwise, it is 0. For example, if a user requests a cheap Italian restaurant, but the system suggests an expensive Italian restaurant, the Inform rate for that dialogue is 0. The Success rate evaluates whether the system fully satisfies user goals. It is the fraction of dialogues where all user requests are successfully fulfilled. For instance, if a user asks to `book a table at a suggested restaurant` and the system fails to confirm the reservation, the Success rate for that dialogue is 0.

τ-Bench Reward. τ-Bench framework defined a binary reward metric that captures both **action correctness** and **output completeness** in a task-oriented dialogue. Specifically, let $r_{\text{action}} \in \{0, 1\}$ be an indicator that the final state of the database matches the unique outcome of ground truth, and let $r_{\text{output}} \in \{0, 1\}$ be an indicator that the system's final response to the user contains all the required information. Then the overall reward is computed as $r = r_{\text{action}} \times r_{\text{output}} \in \{0, 1\}$. A successful dialogue must both perform the correct database updates (e.g., returning the correct items) and provide all necessary details (e.g., item prices) to the user.

Critically, all three metrics are calculated at the *end of the dialogue*, providing a summary of the system's overall ability to understand and fulfill user requests.

2.2 Limitations of These Metrics

Lack of Turn-Level Granularity. Inform, Success, and τ-Bench rewards are calculated once the dialogue concludes and can miss specifics about how the system performs at each turn. As a result, errors early in the conversation are unpenalized if the system eventually corrects itself. For instance, as illustrated in Figure 1, a system might initially hallucinate a restaurant name but later rectify its mistake, yielding a misleadingly high Inform and Success score.

Binary Nature and Oversimplification. Each of these metrics treat correctness as an all-or-nothing concept. Any deviation from the exact requested entity is considered equally wrong, whether it is entirely incorrect (e.g., a different type of venue in a different location) or partially incorrect (e.g., correct area but wrong cuisine). This approach ignores the difference between small and large errors, which can pose challenges when evaluating LLM-based systems that may produce partially correct responses. On the other hand, LLMs exhibit a variety of failure modes that traditional systems do not. They may hallucinate details, misinterpret subtle intentions, or produce inconsistent responses across turns. Because these metrics focus only on the final outcome, they fail to capture these unique conversational flaws.

String-matching Vulnerabilities. τ-Bench's reward-based evaluations rely on simplistic substring matching can produce misleading evaluations. These methods simply check for the presence of target substrings in model-generated responses, which can lead to misleading evaluations. For example, a model answering a numerical question incorrectly may still contain the correct number elsewhere in its response, either by coincidence or in an unrelated context, yet it still may be scored as correct like in Figure 2 (right).

These limitations collectively mean that automated metrics can provide incomplete or inflated estimates of true performance, especially for LLM-driven TOD systems where user experience depends on the quality and consistency of responses at every turn.

3 TD-EVAL

To comprehensively evaluate LLM performance in completing TOD tasks, assessments are crucial at both the turn level and the dialogue level (Siro et al., 2022). We propose a two-step evaluation protocol, TD-EVAL, as shown in Figure 3. The first step introduces a turn-level metric to evaluate TOD responses across three key dimensions: *Conversation Cohesion*, *Backend Knowledge Consistency*, and *Policy Compliance*. We chose these dimensions to specifically target the TOD domain. The second step extends this evaluation to from turn to dialogue-level with a pairwise ranking method; for each dialogue pair, it indicates a preference for the stronger conversation following prompted policies. *This dual-step protocol offers a holistic view of TOD performance, capturing both localized (per turn) and global (whole-dialogue) qualities through a scalable, automated with LLM-as-judge approach.*

3.1 Turn-Level Evaluation

As illustrated in the left side of Figure 3, turn-level evaluation focuses on analyzing each system's response on each turn within the context of the dialogue, ensuring that intermediate errors are identified and penalized. The evaluation process involves feeding the user query, conversation history, and system response into our framework, which uses an LLM as the primary evaluator. Each response is rated on a 5-point scale with three different dimensions: (i) conversation cohesion, (ii) backend knowledge consistency, and (iii) policy compliance, using prompts specifically designed to elicit accurate judgments.

Conversation Cohesion. This dimension evaluates how well the agent's response aligns with the preceding dialogue context, ensuring the system remains relevant to both the user query and the dialogue history while maintaining a coherent flow. A key motivation for measuring conversation cohesion is that it helps identify whether the agent remains on-topic and avoids illogical transitions, which are critical for a smooth user experience. To measure this, we prompt the LLM judge with the entire conversation context (including the user query, dialogue history, database results, and the system's response) and ask it to assign a score from 1 (Very Bad) to 5 (Very Good), along with its justification (See Figure 6 for prompt).

Backend Knowledge Consistency. This dimension assesses how accurately the agent integrates external information (e.g., database results) into its responses, reflecting the system's ability to retrieve and incorporate factual details during the dialogue. Maintaining correct backend knowledge is crucial to building user trust and ensuring that the agent can handle task-oriented requests (e.g., providing real-time data or service details) without misinformation. We feed the same conversation context plus database results into the LLM judge with the prompt of instructions presented in Figure 7, which checks the factual correctness and topical consistency of the response.

Policy Compliance. A core requirement in TOD is following predefined, domain-specific policy protocols (e.g., when to request a user's booking details

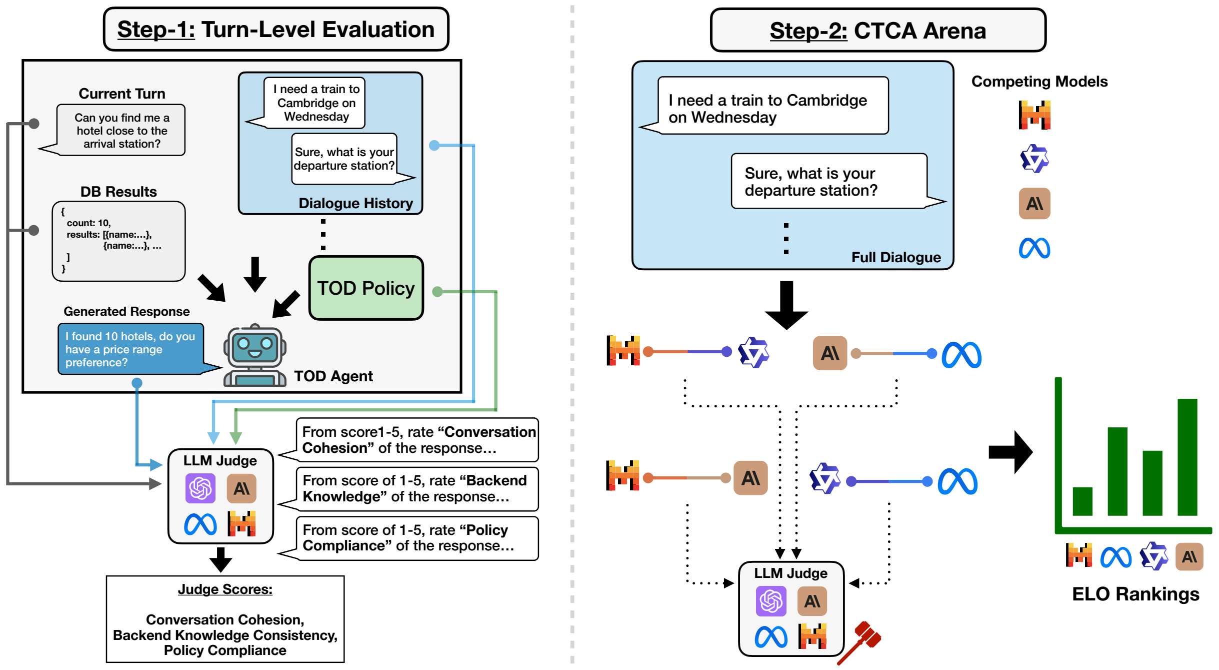

Figure 3: **Overview of the TD-EVAL Framework.** The left part illustrates the pipeline for turn-level evaluations, where a system is assessed across three criterion by an LLM-based judge. The right part depicts the second step, in which models compete within CTCA Arena and ranked based on win and loss rates.

or how many suggestions to offer). To evaluate this, we check whether an agent's responses align with a policy prompt as in Figure 8, covering the domain's possible slots (e.g., time, price range), a set of policy rules (e.g., "wait for key information before making a reservation"), and a scoring guide. At each turn, the LLM judge references these instructions—along with the dialogue context and database outputs—to assign a discrete score. We create domain-specific prompts (e.g., for *restaurant*, *train*, *hotel*, *attraction*, *taxi*), each reflecting unique policy objectives. This structured approach ensures evaluations reveal how well the system follows the exact flows required in real-world services, addressing common failures of LLMs to adhere to detailed policies.

3.2 Dialogue-Level Evaluation

Motivation for CTCA Arena. While turn-level metrics are useful for spotting localized issues, they can be too granular for capturing the overall user experience in extended conversations, especially in service-oriented conversations that require database (DB) interaction. For instance, if two models achieve close per-turn Likert scores, it can be difficult to discern which one ultimately yields a more coherent, user-friendly dialogue end-to-end. Moreover, turn-level assessments may not reflect human preferences in extended interactions. A slight advantage (e.g., an average score of 4.2 vs. 4.0) doesn't necessarily show how well the agent meets user goals or recovers from mistakes over multiple turns.

Differences from Existing Arenas. Unlike general-purpose chatbot arenas (Zheng et al., 2023), which often focus on open-domain or chit-chat dialogues, our **CTCA Arena** specifically targets task-oriented scenarios that require integration with domain-specific DBs. This emphasis on DB-driven tasks introduces unique challenges: the system must retrieve, update, and reason about information stored in external data sources while maintaining coherent and contextually relevant multi-turn conversations. By centering our comparisons on these, our arena offers a distinct perspective on conversational quality, where success is measured not only by linguistic fluency but also by the agent's ability to fulfill service-oriented goals.

Evaluation of Pairwise Comparisons. To address these limitations, we employ a pairwise ranking methodology inspired by MT-Bench (Zheng et al., 2023), which utilizes an Elo rating system. The Elo system provides a setup for evaluating relative performance through pairwise comparisons, as shown in the right side of Figure 3. In our adaptation, all models begin with an initial rating of 1000, and their scores are dynamically adjusted

Model	Conversational Consistency			Backend Knowledge			Policy Compliance			Overall
	MultiWOZ	Tau-Bench	Avg.	MultiWOZ	Tau-Bench	Avg.	MultiWOZ	Tau-Bench	Avg.	Avg.
o1	4.4722	**4.7680**	4.6201	4.3623	4.3198	4.3412	**4.4179**	**4.6091**	**4.5135**	**4.4916**
GPT-4o	3.9362	4.6601	4.2982	4.0841	4.1256	4.1049	4.0326	4.3568	4.1947	4.1993
GPT-4o-mini	3.7133	4.4404	4.0769	3.9239	3.9068	3.9154	4.0285	4.2628	4.1457	4.0460
GPT-3.5-turbo	3.1984	4.4066	3.8025	3.5774	4.0506	3.8140	3.3899	3.5106	3.4503	3.6889
Claude-3.5-Sonnet	**4.6518**	4.6030	**4.6274**	**4.5447**	4.1315	4.3381	4.3930	4.4798	4.4364	4.4673
Llama-3.1-405B-Instruct	4.3129	4.5663	4.4396	4.3741	**4.4623**	**4.4182**	4.1673	4.4469	4.3071	4.3883
Llama-3.3-70B-Instruct	4.1749	3.9977	4.0863	4.1352	3.0003	3.5678	3.9003	3.8061	3.8532	3.8358
Mistral-Large	4.2076	4.2553	4.2315	4.2795	4.3088	4.2942	4.1085	4.2724	4.1905	4.2387
Qwen2.5-72B-Instruct	4.2172	4.4093	4.3133	4.4085	4.4256	4.4171	4.0301	4.2046	4.1174	4.2826

Table 1: Turn-level Results. The best results are highlighted in **bold**, while the second-best results are underlined.

Ranking	Model	Conversational Task Completion Agent Arena	Votes	Wins	Losses	Ties	Organization	License
1	Claude-3.5-Sonnet	1279.66	1488	1237	216	35	Anthropic	Proprietary
2	GPT-4o	1107.40	1488	861	581	46	OpenAI	Proprietary
3	GPT-4o-mini	1037.46	1488	743	707	38	OpenAI	Proprietary
4	Mistral-Large	988.89	1488	659	785	44	Mistral	Mistral Research
5	Llama-3.1-405B-Instruct	961.04	1488	708	718	62	Meta	Llama
6	Llama-3.3-70B-Instruct	901.34	1488	591	840	57	Meta	Llama
7	GPT-3.5-turbo	724.21	1488	265	1217	6	OpenAI	Proprietary

Table 2: Dialogue-level results and CTCA Arena leaderboard.

based on LLM-based judgments of their responses. The process works as follows; first, two competing models generate responses for the same dialogue context as "Conversation A" and "Conversation B". Following the provided judge prompt (See Figure 9), our LLM judge evaluates these conversations to determine between three options only: (i) "Conversation A" if Conversation A was better, (ii) "Conversation B" if Conversation B was better, or (iii) "Equal" if they were roughly equivalent. Based on this comparison, the models' Elo ratings are updated, with the magnitude of adjustment reflecting the outcome's predictability—unexpected victories (such as a lower-rated model outperforming a higher-rated one) result in larger rating changes. This iterative process continues until we obtain a stable ranking that reflects each model's relative conversational capabilities. Thus, the dialogue-level arena adds a broader perspective that complements turn-level scores, ensuring that multi-turn dynamics, user goals, and overall satisfaction are adequately captured in our evaluation.

4 Experiments

This section presents the experimental setup and results for TD-EVAL. We first describe the dataset and models we used, followed by quantitative results and ablation analysis.

4.1 Experimental Settings

Evaluation Instances. To assess the performance of TD-EVAL, we utilize two popular conversational dialogue benchmarks: MultiWOZ 2.4 (Ye et al., 2022) and τ-Bench (Yao et al., 2024), both originally evaluated using automatic metrics. We selected 100 dialogues from MultiWOZ 2.4 with human validation and used all 165 dialogues from τ-Bench, covering the retail and airline domains. Finally, we report the accuracy of our LLM-based judge separately on each benchmark dataset.

Models. We evaluate nine different LLMs including both closed-source and opened-source models: o1, GPT-4o, GPT-4o-mini, GPT-3.5-turbo (OpenAI), Llama-3.3-70B-Instruct, Llama-3.1-405B-Instruct (Meta), Claude-3.5-Sonnet (Anthropic), Mistral-Large (Mistral Research), Qwen 2.5-72B-Instruct (Alibaba). Each model is prompted to produce system responses given multi-turn dialog contexts and database records, using the same policy instructions described in Section 3.1. To evaluate the generated responses, we employ GPT-4o as default both in LLM-based Judge and User Simulator.

4.2 Main Results

Turn-level Evaluation Results. Table 1 shows the average Likert scores (1–5) on *conversation cohesion*, *backend knowledge*, and *policy compliance*, along with an overall average across these three dimensions. The o1 model achieves the highest overall score of 4.4916, demonstrating strong performance in policy compliance due to its advanced reasoning capabilities in following instructions. Following that, Claude-3.5-Sonnet ranks second, with Llama-3.1-405B-Instruct placing third

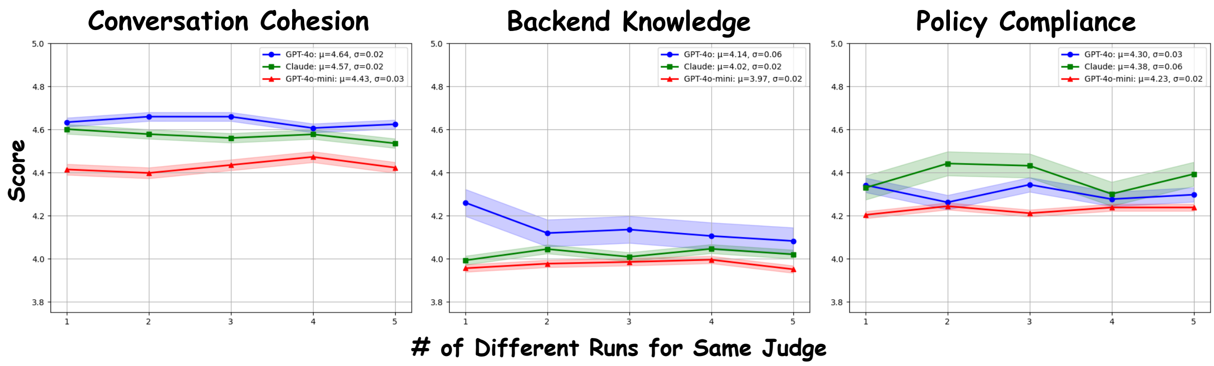

Figure 4: **LLM Judge Consistency Across Multiple Runs.** Evaluation results for three metrics using the same LLM judge over five runs: conversation cohesion (left), backend knowledge (middle), and policy compliance (right).

as the top open-source model. Notably, Llama-3.1-405B-Instruct achieves the highest score in *Backend Knowledge* (4.4182), outperforming proprietary models. Table 1 indicates that while certain models perform well in cohesion and factual correctness, they may fall short in strictly enforcing policies. This highlights the need for a comprehensive evaluation across multiple dimensions.

Dialogue-level Evaluation Results. Table 2 summarizes the pairwise Elo rankings under our CTCA Arena. Claude-3.5-Sonnet dominates the rankings with an Elo rating of 1514.42, winning 571 out of 600 head-to-head matchups in turn-level evaluations. This aligns with its strong turn-level performance, indicating its effectiveness in end-to-end conversations. Although Mistral-Large did not rank at the very top in turn-level scores, it unexpectedly places second in the dialogue-level evaluation (Elo = 1068.67), implying that it manages to recover from localized errors and provides more user-friendly end-to-end interactions. GPT-4o and Meta-Llama-3.1-405B Instruct both maintain solid Elo ratings (1055.40 and 1020.36, respectively), consistent with their strong turn-level performance. In contrast, GPT-3.5-turbo (Elo = 648.35) sits at the bottom, aligning with its lower per-turn scores.

4.3 Ablation Studies

TD-EVAL demonstrates consistent scoring in multiple runs. A common concern with LLM-based evaluation is the potential variability in assessment outcomes. To investigate this, we conducted a reliability analysis by *running the same judge five times* (GPT-4o) on identical responses from three agents (Claude-3.5-Sonnet, GPT-4o-mini, and GPT-4o) on τ-Bench. Figure 4 visualizes score variation over multiple runs together with standard deviations. The nearly flat curves and low standard deviations demonstrate that TD-EVAL yields highly consistent judgments, underscoring the robustness of our evaluation protocol.

Meric	Human vs. Human	Human vs. LLM
Conversation Cohesion	0.50	0.46
Backend Knowledge	0.56	0.33
Policy Compliance	0.49	0.38
Overall	0.52	0.39

Table 3: Human-Human and Human-LLM Judgement Agreement with Randolph's Kappa.

TD-EVAL scores are robust to different LLM judges. To evaluate the robustness of our LLM-based judge, we compare GPT-4o, GPT-4o-mini, and Claude-3.5-Sonnet with three different LLM judges (GPT-4o, Llama-405B, and Claude-3.5-Sonnet), shown in Figure 5. We observe that all three judges produce closely aligned absolute scores. Additionally, the relative ranking of the evaluated models remains consistent across different judges. Interestingly, Llama-405B sometimes assigns higher scores than its proprietary counterparts, possibly reflecting different reasoning preferences. Overall, our results demonstrate that the scoring remains stable across diverse LLM judges.

5 Human Agreement Evaluation

To validate TD-EVAL as a metric, we performed a human evaluation with five engineering students proficient in English. Each annotators reviewed 8 randomly selected conversations from MultiWOZ and τ-bench, generated by different models. Using the given dialogue history and database results as context, they provided turn-level and dialogue-level ratings on a 5-point Likert scale, from "Very Bad" to "Very Good". In total, we collected annotations for 40 conversations (20 from each dataset), covering 323 dialogue turns, with each conversation

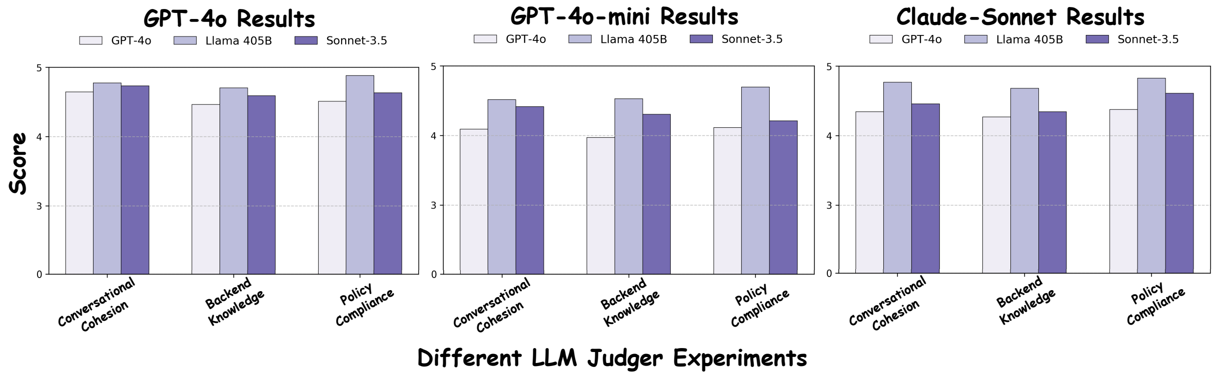

Figure 5: **Different LLM Judge Experiments.** Evaluation scores of GPT-4o (left), GPT-4o-mini (center), and Claude-Sonnet-3.5 (right) under different LLM judges (GPT-4o, Llama-405B, and Claude-3.5-Sonnet).

evaluated by two annotators. Detailed annotation guidelines and additional information are provided in Appendix A.

Table 3 presents the results of our human evaluation. To measure agreement, we used Randolph's Kappa (Randolph, 2005), as each conversation was evaluated by a subset of annotators. Kappa scores are computed for both turn-level and dialogue-level (overall) assessment. Our findings indicate a strong agreement for both Human-Human (i.e., between human evaluators) and Human-LLM (i.e., between human annotators and the TD-EVAL evaluation framework), demonstrating the robustness and reliability of the TD-EVAL framework.

6 Related Work

6.1 LLM-Based TOD Evaluation

Recent research on TOD has leveraged LLMs through various ways, including fine-tuning (Su et al., 2022) and prompting strategies (Hu et al., 2022), with particular emphasis on challenges related to database access and logical reasoning (Hudeček and Dusek, 2023). With these challenges, comprehensive evaluation remains a key bottleneck in TOD research. While early works—such as Walker et al. (1997)—established the foundational evaluation paradigms, the advancement of TOD evaluation metrics has lagged behind that of open-domain dialogue systems, which has introduced diverse approaches such as reference-based (Lowe et al., 2017) and reference-free (Li et al., 2021; Dey and Desarkar, 2023) methods. In the TOD domain, automated evaluation metrics often fail to capture nuanced multi-turn errors (Mehri et al., 2019). To address these limitations, TD-EVAL introduces multi-perspective LLM-based evaluation metrics specifically tailored to TOD scenarios.

6.2 Multi-Turn TOD Evaluation

The inherent multi-turn nature of TOD presents additional challenges for evaluation, largely due to the error accumulation across turns (Liao et al., 2018, 2021). Although prior studies have explored both turn-level and dialogue-level evaluation approaches (Bodigutla et al., 2020; Dey et al., 2022; Kim et al., 2022; Xu et al., 2024), developing a comprehensive evaluation metric capable of assessing the overall interaction quality remains an open problem. To resolve this gap, TD-EVAL provides a systematic framework designed to evaluate TOD interactions at both the turn and dialogue levels, providing a more holistic assessment of system performance.

7 Conclusion

We present TD-EVAL, a simple yet powerful framework for TOD evaluation that combines fine-grained turn-level checks with a holistic dialogue-level ranking. By adopting an LLM-as-judge paradigm, TD-EVAL goes beyond standard metrics to reveal subtle yet critical errors, such as inconsistent database usage and policy violations, which often remain undetected by final-turn or dialogue-level summaries. Through ELO-based ranking and targeted turn-level scoring, our experiments on MultiWOZ 2.4 and τ-Bench demonstrate TD-EVAL's strong alignment with human judgments. This work opens a new path for LLM-driven TOD evaluation—one that is both flexible and transparent—ensuring greater accountability and accuracy in developing next-generation dialogue systems. We release our framework, system responses, and human evaluations to foster reproducibility and community adoption.

8 Limitations

While metrics in TD-EVAL cover core aspects occurring in general TOD scenarios, it still remains open questions to design more flexible, fine-grained evaluation metrics that can cover diverse scenarios during multi-turn interactions. Furthermore, practitioners should consider that the performance of LLM-based evaluation can be improved when appending qualified few-shot demonstrations or tailored scoring rubrics in specific service domains. Lastly, it should be noted that conventional evaluation metrics are still useful to evaluate TOD in specific aspects, thus TD-EVAL should be utilized with existing metrics with a complementary relationship.

9 Ethics Statement

We conduct our experiments using the publicly available MultiWOZ and τ-Bench datasets, adhering fully to their terms of use. Since we employ LLMs to generate evaluations with justifications, the risk of producing harmful, biased, or discriminatory statements is minimal. However, we acknowledge the potential ethical concerns associated with this work.

References

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3897–3909, Online. Association for Computational Linguistics.

Sari Dewi Budiwati, Tirana Fatyanosa, Mahendra Data, Dedy Rahman Wijaya, Patrick Adolf Telnoni, Arie Ardiyanti Suryani, Agus Pratondo, and Masayoshi Aritsugi. 2021. To optimize, or not to optimize, that is the question: TelU-KU models for WMT21 large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 387–397, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Suvodip Dey and Maunendra Sankar Desarkar. 2023. Dial-M: A masking-based framework for dialogue evaluation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 77–84, Prague, Czechia. Association for Computational Linguistics.

Suvodip Dey, Ramamohan Kummara, and Maunendra Desarkar. 2022. Towards fair evaluation of dialogue state tracking by flexible incorporation of turn-level performances. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 318–324, Dublin, Ireland. Association for Computational Linguistics.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. 2022. Mismatch between multi-turn dialogue and its evaluation metric in dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 297–309, Dublin, Ireland. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. *arXiv preprint arXiv:2402.10466*.

Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 801–809, New York, NY, USA. Association for Computing Machinery.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle

Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 165–177, Stockholm, Sweden. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k[free]): An alternative to fleiss' fixed-marginal multirater kappa.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2018–2023, New York, NY, USA. Association for Computing Machinery.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan 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, Bangkok, Thailand. Association for Computational Linguistics.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Appendix

A Additional Details on Human Evaluation

Human evaluation is conducted through Qualtrics, an online survey platform. The instructions provided to human annotators is shown in Fig. 10. Examples of the evaluation samples on qualtrics are shown in Fig. 11. For evaluation, dialogue samples were selected from 5 LLMs with different parameter sizes. These choices were made to evaluate judgement quality for both good and bad quality responses. The models chosen were GPT-4o, Claude Sonnet 3.5, Mistral Large, GPT-4o mini, and Llama 70B. Dialogue samples were evenly split, with 4 from MultiWOZ and 4 from Tau-Bench. Each annotator was randomly assigned two model samples to evaluate. Survey samples were split between MultiWOZ and τ-Bench due to their divergent approaches to response generation. MultiWOZ dialogues were generated with fixed "ground truth" user utterances, while τ-Bench samples were made by online generation of both user and TOD agent.

Evaluate the **conversation consistency** of the following task-oriented dialogue chatbot response on a 5-point scale from "Very Bad" to "Very Good".
The prompt will include the **dialogue history**, **current user query**, **database results**, **chatbot response**.

Conversation Consistency Definition
Conversation consistency refers to the degree to which the chatbot's response aligns with the context of the conversation, including:

- **Relevance**: The response directly relates to the dialogue history and the current user query.
- **Topic Consistency**: The response remains on-topic with the dialogue history and the user query.
- **Coherence**: The response logically continues the progression of the dialogue.

Scoring Guide:

- **Very Good (5)**: Response is completely consistent with the previous conversation context, with no inconsistencies or errors.
- **Good (4)**: Response is mostly consistent with the context. Only minor improvements are needed.
- **Fair (3)**: Response is somewhat consistent but contains noticeable inconsistencies or lacks depth in addressing the context.
- **Bad (2)**: Response shows limited consistency with the conversation context and requires significant improvement.
- **Very Bad (1)**: Response is incoherent or completely inconsistent with the conversation context.

Always include the score first, then a rationale on a new line. The rationale should be at most 2 sentences. Follow the template below:
Score: [YOUR SCORE NUMBER HERE]
Justification: [YOUR RATIONALE HERE]

Evaluate the conversation consistency of the following dialogue response
Dialogue History
{dialogue_history}
User Query
{user_query}
Database Results
{db_result}
Chatbot Response
{agent_response}

Figure 6: The GPT-4o prompt used for evaluating Conversational Consistency on MultiWOZ.

Evaluate the **backend knowledge consistency** of the following task-oriented dialogue chatbot response on a 5-point scale from "Very Bad" to "Very Good".
The prompt will include the **dialogue history**, **current user query**, **database results**, **chatbot response**.

Backend Knowledge Consistency Definition
Backend knowledge consistency refers to how well the chatbot's response aligns with the information provided in the policy or database results, considering:

- **Accuracy**: The response accurately reflects the information in the database results.
- **Topic Consistency**: The response stays on-topic with the database results and the dialogue context.
- **Coherence**: The response logically incorporates and progresses based on the database results.

Scoring Guide:

- **Very Good (5)**: Response is completely consistent with the database results, with no inconsistencies or errors.
- **Good (4)**: Response is mostly consistent with the database results. Only minor improvements are needed.
- **Fair (3)**: Response is sufficiently consistent with the database results but contains noticeable inconsistencies or lacks depth in addressing the results.
- **Bad (2)**: Response shows limited consistency with the database results and requires significant improvement.
- **Very Bad (1)**: Response is incoherent or completely inconsistent with the database results.

Always include the score first, then a rationale on a new line. The rationale should be at most 2 sentences. Follow the template below:
Score: [YOUR SCORE NUMBER HERE]
Justification: [YOUR RATIONALE HERE]

Evaluate the backend knowledge consistency of the following dialogue response
Dialogue History
{dialogue_history}
User Query
{user_query}
Database Results
{db_result}
Chatbot Response
{agent_response}

Figure 7: The GPT-4o prompt used for evaluating Backend Knowledge on MultiWOZ.

Evaluate the **policy compliance** of the following task-oriented dialogue chatbot response on a 5-point scale from "Very Bad" to "Very Good". The prompt will include the **policy protocol**, **dialogue history**, **current user query**, **database results**, **chatbot response**.

Policy Compliance Definition

Policy compliance refers to how well the chatbot adheres to the expected policy protocol, specifically:

- **Number of Suggestions**: Providing suggestions only when the database results are few enough to do so.
- **Information Gathering**: Requesting required, relevant information (slots) from the user before offering suggestions or booking services.
- **Appropriate Timing**: Avoiding premature actions, such as making a booking or suggesting a service too early in the conversation, before the necessary information is gathered from the user.
- **Alignment with Policy**: Avoiding actions that do not align with the suggested flow of interaction in the policy, when available.

Domain Possible Slots (May not be exhaustive)

The current predicted domain for this turn in the conversation is "Restaurant". You should check if domain-relevant slots have been filled in the current conversation. The possible slots for all domains are shown below (these may not be totally exhaustive):

Restaurant

- bookpeople: the number of people included in the restaurant reservation
- booktime: the time for the reservation at the restaurant
- food: the type of cuisine the restaurant serves
- name: the name of the restaurant
- pricerange: the price range of the restaurant

Scoring Guide:

- **5 (Very Good)**: Response fully follows policy protocol with no errors or omissions.
- **4 (Good)**: Response mostly follows policy protocol, with only minor room for improvement.
- **3 (Fair)**: Response sufficiently follows policy protocol but has clear areas where it could improve in completeness or timing.
- **2 (Bad)**: Response does not adequately follow policy protocol, though there may be partial adherence.
- **1 (Very Bad)**: Response fails to follow policy protocol and is incomplete or incoherent.

Always include the score first, then a rationale on a new line. The rationale should be at most 2 sentences. Follow the template below:

Score: [YOUR SCORE NUMBER HERE]

Justification: [YOUR RATIONALE HERE]

Policy Protocol

The chatbot response should depend on the database results and dialogue history:

1. If the database results return a number: Indicate the number of entries that match the user's query and request additional information if needed to narrow down the results.
2. If the database results return values: If vital details are missing based on the dialogue history, request additional information. Otherwise, provide the relevant entries to the user.

Evaluate the policy completeness of the following dialogue response

Dialogue History

{dialogue_history}

User Query

{user_query}

Database Results

{db_result}

Chatbot Response

{agent_response}

Figure 8: The GPT-4o prompt used for evaluating Policy Compliance Consistency on MultiWOZ restaurant domain.

Compare these two AI assistant conversations and determine which one is better.
Consider the following aspects:

1. **Conversation Consistency**:
    - **Relevance**: The response directly relates to the dialogue history and the current user query.
    - **Topic Consistency**: The response remains on-topic with the dialogue history and the user query.
    - **Coherence**: The response logically continues the progression of the dialogue.
2. **Backend Knowledge Consistency**:
    - **Accuracy**: The response accurately reflects the information in the database results.
    - **Topic Consistency**: The response stays on-topic with the database results and the dialogue context.
    - **Coherence**: The response logically incorporates and progresses based on the database results.
3. **Policy Compliance**:
    - **Number of Suggestions**: Providing suggestions only when the database results are few enough to do so.
    - **Information Gathering**: Requesting required, relevant information (slots) from the user before offering suggestions or booking services.
    - **Appropriate Timing**: Avoiding premature actions, such as making a booking or suggesting a service too early in the conversation.
    - **Policy Protocol**: The chatbot response should depend on the database results and dialogue history:
        (a) If the database results return a number: Indicate the number of entries that match the user's query and request additional information if needed to narrow down the results.
        (b) If the database results return values: If vital details are missing based on the dialogue history, request additional information. Otherwise, provide the relevant entries to the user.

Conversation A:
{conv_a_formatted}

Conversation B:
{conv_b_formatted}

Which conversation was better? Answer with only:
CONVERSATION_A if Conversation A was better
CONVERSATION_B if Conversation B was better
EQUAL if they were roughly equivalent

Figure 9: The GPT-4o prompt used for dialogue-level evaluation in Service-Oriented Dialogue Arena.

Thank you for participating in this annotation task. We have provided 8 dialogues where an AI agent is responding to user requests, and we want you to evaluate the responses. This process should take roughly 45-60 minutes. Progress will be saved automatically, so you can complete it in multiple sessions, but **make sure you are using the same browser each time**. Please rate individual dialogue responses of AI agents from 1 (worst) to 5 (best) on the following qualities: **Conversation Consistency**, **Backend Knowledge Consistency**, and **Policy Compliance**, the metric definitions are below. In addition, please rate the full dialogue in terms of task completion and response coherence.

Conversation Consistency

How much an agent's response align with the context of conversation context.

- **Relevance**: The response directly relates to the dialogue history and the current user query.
- **Topic Consistency**: The response remains on-topic with the dialogue history and the user query.
- **Coherence**: The response logically continues the progression of the dialogue.

Scoring Scale

1. **Very Good**: Response is completely consistent with the previous conversation context, with no inconsistencies or errors.
2. **Good**: Response is mostly consistent with the context. Only minor improvements are needed.
3. **Fair**: Response is somewhat consistent but contains noticeable inconsistencies or lacks depth in addressing the context.
4. **Bad**: Response shows limited consistency with the conversation context and requires significant improvement.
5. **Very Bad**: Response is incoherent or completely inconsistent with the conversation context.

Backend Knowledge Consistency

How well an agent's response aligns with information provided by backend database results.

- **Accuracy**: The response directly reflects the information in the database results.
- **Topic Consistency**: The response stays on-topic with the database results and the dialogue context.
- **Coherence**: The response logically incorporates and progresses based on the database results.

Scoring Scale

1. **Very Good**: Response is completely consistent with the database results, with no inconsistencies or errors.
2. **Good**: Response is mostly consistent with the database results. Only minor improvements are needed.
3. **Fair**: Response is sufficiently consistent with the database results but contains noticeable inconsistencies or lacks depth in addressing the results.
4. **Bad**: Response shows limited consistency with the database results and requires significant improvement.
5. **Very Bad**: Response is incoherent or completely inconsistent with the database results.

Policy Compliance

How well an agent's response adheres to the expected policy protocol.

- **Number of Suggestions**: Providing suggestions only when the database results are small enough to do so.
- **Information Gathering**: Requesting required, relevant information (slots) from the user before offering suggestions or booking services.
- **Appropriate Timing**: Avoiding premature actions, such as making a booking or suggesting a service too early in the conversation.
- **Alignment with Policy**: Avoiding actions that do not align with the suggested flow of interaction in the policy, when available.

Expected Policy

The chatbot response should depend on the database results and dialogue history:

- If the database results return a number larger than 10: Indicate the number of entries that match the user's query and request additional information if needed to narrow down the results.
- If the database results return values less than 10: If vital details are missing, request additional information. Otherwise, provide the relevant entries to the user

Scoring Scale

1. **Very Good**: Response fully follows policy protocol with no errors or omissions.
2. **Good**: Response mostly follows policy protocol, with only minor room for improvement.
3. **Fair**: Response sufficiently follows policy protocol but has clear areas where it could improve in completeness or timing.
4. **Bad**: Response does not adequately follow policy protocol, though there may be partial adherence.
5. **Very Bad**: Response fails to follow policy protocol and is incomplete or incoherent.

Figure 10: The evaluation instructions provided as reference for human evaluation on Qualtrics platform.

Read the dialogue context below:

User: I want to find a moderately priced restaurant.
Agent: We have found 31 moderately priced restaurants, here's a sample: shanghai family restaurant, lan hong house, restaurant alimentum, saint johns chop house, de luca cucina and bar, to narrow down the results, could you please provide more information such as the area or type of food you're looking for?
User: Yes I would like the restaurant to be located in the center of the attractions.

Database Result:
Count: 44

- {'address': "little saint mary's lane", 'area': 'centre', 'id': '40', 'name': "little saint mary's church", 'phone': '01223366202', 'postcode': 'cb21qy', 'pricerange': 'free', 'type': 'architecture', 'entrancefee': 'free'}
- {'address': '98 king street', 'area': 'centre', 'id': '5', 'name': 'broughton house gallery', 'phone': '01223314960', 'postcode': 'cb11ln', 'pricerange': 'free', 'type': 'museum', 'entrancefee': 'free'}
- {'address': 'trinity lane', 'area': 'centre', 'id': '48', 'name': 'old schools', 'phone': '01223332320', 'postcode': 'cb21tt', 'pricerange': 'free', 'type': 'architecture', 'entrancefee': 'free'}
- {'address': 'anglia ruskin university, east road', 'area': 'centre', 'id': '56', 'name': 'ruskin gallery', 'phone': '01245493131', 'postcode': 'cb11pt', 'pricerange': 'free', 'type': 'museum', 'entrancefee': 'free'}
- {'address': '1 wheeler street', 'area': 'centre', 'id': '36', 'name': 'kambar', 'phone': '01223842725', 'postcode': 'cb23qb', 'pricerange': 'moderate', 'type': 'nightclub', 'entrancefee': '5 pounds'}

Agent: We have found 44 attractions in the centre area, but it seems none of them are restaurants, here's a sample of the results: little saint mary's church, broughton house gallery, old schools, ruskin gallery, kambar, to find a restaurant, could you please provide more information such as the type of food you're looking for or would you like me to try and find restaurants in the centre area again?

Rate the Conversation Consistency, Backend Knowledge Consistency, and Policy Compliance from 1 to 5 for the above response.

	Very Bad	Bad	Fair	Good	Very Good
Conversation Consistency	○	○	○	○	○
Backend Knowledge Consistency	○	○	○	○	○
Policy Compliance	○	○	○	○	○

Read the dialogue context below:

User: Yes I am looking for an expensive restaurant in the South of town.
Agent: there are 5 expensive restaurants in the south area, here are a few options: the good luck chinese food takeaway, taj tandoori, chiquito restaurant bar, peking restaurant, and frankie and bennys, which one would you like to book or would you like more information.
User: Sorry, Actually I need an expensive restaurant in the north. the first on your list would be great.
Agent: I've searched for the Good Luck Chinese Food Takeaway in the north area, but it seems it's not available, I found it earlier in the south area, would you like to try another expensive restaurant in the north or go back to the original option in the south.
User: Sounds great, can you book a table for 5 at 18:00?
Agent: I've booked a table for 5 people at 18:00 at the Good Luck Chinese Food Takeaway in the north area, your reference code is 4T9R2A1B.
User: Actually, I won't be needing a reservation just yet. Could you verify the type of food Saigon City serves?
Agent: Saigon City serves Vietnamese food, would you like to book a table there or need more information about the restaurant, such as address or phone number.
User: Yes, could I get the restuarant's phone number and postcode?
Agent: The phone number for Saigon City is 01223873455 and the postcode is N1 9GB.
User: No, that's it. You've been very helpful, thank you!
Agent: You're welcome, it was my pleasure to assist you, have a great day and I hope you enjoy your dining experience at Saigon City or any other restaurant you choose to visit.

Response Coherence: Rate your satisfaction with the chatbot's answers to the user's questions. How would you rate your satisfaction with bot's responses in terms of being **easy to understand and helpful?**

Very Bad	Bad	Fair	Good	Very Good
○	○	○	○	○

Figure 11: Example turn level (left) and dialogue level (right) evaluation questions in Qualtrics. Includes previous dialogue context, sample of database results, current turn user query and agent response. Includes full dialogue as context for evaluation.