TD-EVAL: Revisiting Task-Oriented Dialogue Evaluation by Combining Turn-Level Precision with Dialogue-Level Comparisons

Emre Can Acikgoz*, **Carl Guo***, **Suvodip Dey***, **Akul Datta**, **Takyoung Kim**, **Gokhan Tur**, **Dilek Hakkani-Tür**
University of Illinois Urbana-Champaign
{acikgoz2, carlguo2, sdey, gokhan, dilek}@illinois.edu

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 evaluation framework that unifies fine-grained turn-level analysis with holistic dialogue-level comparisons. At turn level, we evaluate each response along three TOD-specific dimensions: *conversation cohesion*, *backend knowledge consistency*, and *policy compliance*. Meanwhile, we design *TOD Agent Arena* that uses pairwise comparisons to provide a measure of dialogue-level quality. Through experiments on MultiWOZ 2.4 and τ-Bench, we demonstrate that TD-EVAL effectively identifies the conversational errors that conventional metrics miss. Furthermore, TD-EVAL exhibits better alignment with human judgments than traditional and LLM-based metrics. These findings demonstrate that TD-EVAL introduces a new paradigm for TOD system evaluation, efficiently assessing both turn and system levels with a plug-and-play framework for future research[1].

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 Large Language Models (LLMs) have significantly enhanced the capabilities and flexibility of modern TOD systems (Hudeček and Dusek, 2023; Acikgoz et al., 2025). 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.

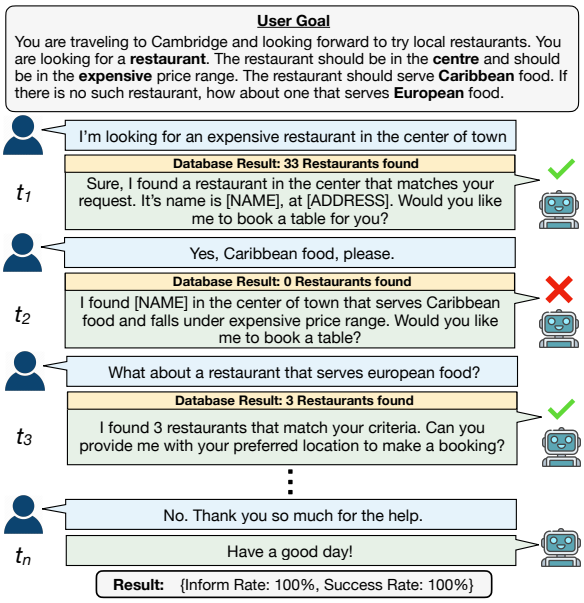

Figure 1: Sample conversation between a user and TOD agent with a MultiWOZ goal. Commonly used evaluation metrics in TOD, Inform and Success, fail to detect the turn-level error and assign a perfect score.

To automate TOD evaluation, previous work has converged to various offline metrics. For example, metrics like *Inform* and *Success* rates are utilized to estimate task completion in MultiWOZ (Budzianowski et al., 2018). Traditionally, the overall performance of TOD systems is evaluated at the dialogue-level using task completion metrics, ignoring turn-level performance. While a few metrics assess turn-level quality. For example, Joint Goal Accuracy (JGA) (Budzianowski et al., 2018) for dialogue understanding and BLEU (Pa-

* indicates equal contribution.
[1]Project Page: https://emrecanacikgoz.github.io/TD-Eval/

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 113–132
Aug 25–27, 2025.

pineni et al., 2002) for response generation. However, these metrics rely heavily on ground-truth annotations, making them impractical for evaluating arbitrary conversations without labeled data.

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 the fact that an intermediate step is completely wrong. On the other hand, task completion relies 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 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 providing a more reliable and human aligned evaluation of conversational quality compared to other metrics. TD-EVAL introduces three turn-level metrics: *conversation cohesion*, *backend knowledge consistency*, and *policy compliance*, evaluated using an LLM judge *together with its 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 (Hurst et al., 2024) as the default. The proposed framework complements the per-turn judgments with a dialogue-level comparison method: a *TOD Agent 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, TOD Agent Arena 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.
- 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, validating with human evaluations.
- We perform a comprehensive evaluation of state-of-the-art LLMs on TD-EVAL and provide insights into their strengths and limitations. We publicly release our code, system-response data, and human evaluations, along with a Huggingface Leaderboard for TOD Agents.

2 Preliminaries

While various aspects can contribute to the *satisfactory* evaluation of 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.

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; Ye et al., 2022) contains two automatic task completion metrics, **Inform** and **Success**. In τ-Bench (Yao et al., 2025), authors defined a binary **Reward** metric that checks the action and output correctness.

Inform and Success. Inform measures the system's ability to provide correct entities from database search results in response to user requests. In addition to entity correctness, Success checks

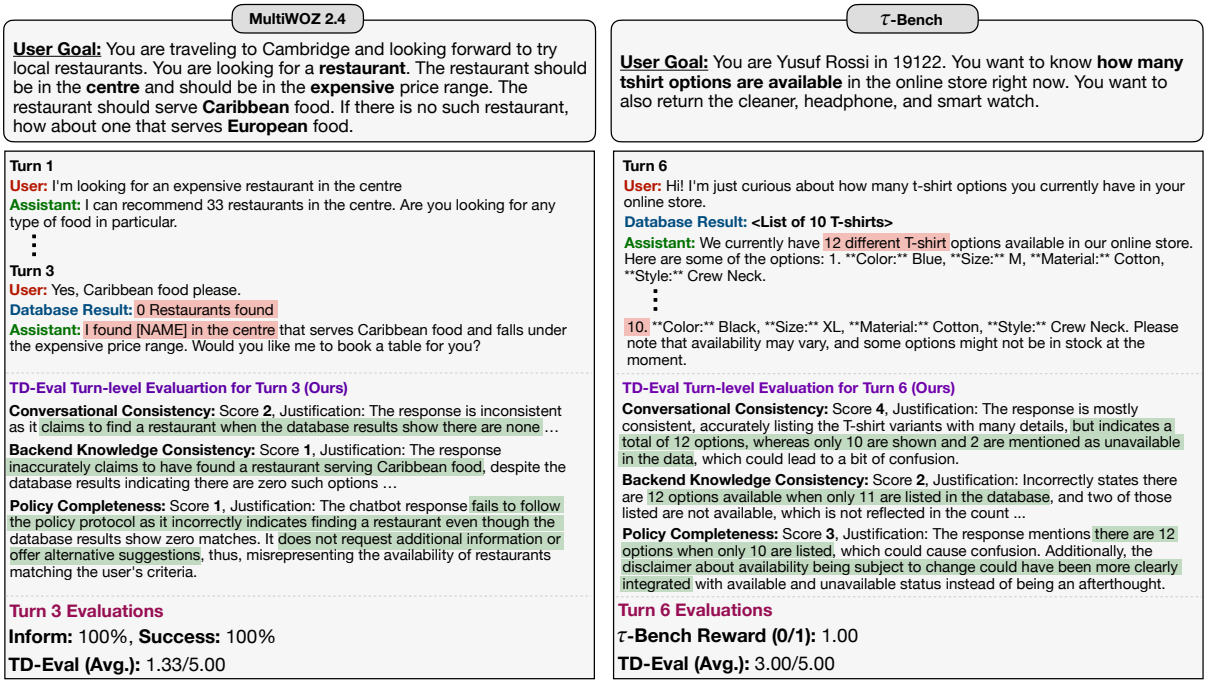

Figure 2: **Comparison of TD-EVAL with Automatic Evaluations. Left:** In MultiWOZ 2.4, a non-existent restaurant is suggested but received a 100% score from Inform and Success, while TD-EVAL's turn-level analysis assigns a lower score. **Right:** On τ-Bench, the agent claims 12 T-shirts (only 10 exist),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

whether the system returns all the requested attributes. For each dialogue, Inform and Success are reported as binary scores (1 or 0), where 1 denotes success and 0 denotes failure. For example, if a user asks to `book a table at a suggested restaurant` and the system fails to confirm the reservation, the Success for that dialogue is 0.

τ-Bench Reward. τ-Bench (Yao et al., 2025) defined a binary reward metric that captures both action correctness and output completeness. 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.

2.2 Limitations of Traditional 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 not penalized if the system eventually corrects itself. For instance, as illustrated in Figure 1 (left), a system might initially hallucinate a restaurant name but later rectify its mistake, resulting in a misleadingly high 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. Furthermore, 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 mainly on substring matching. 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 coinci-

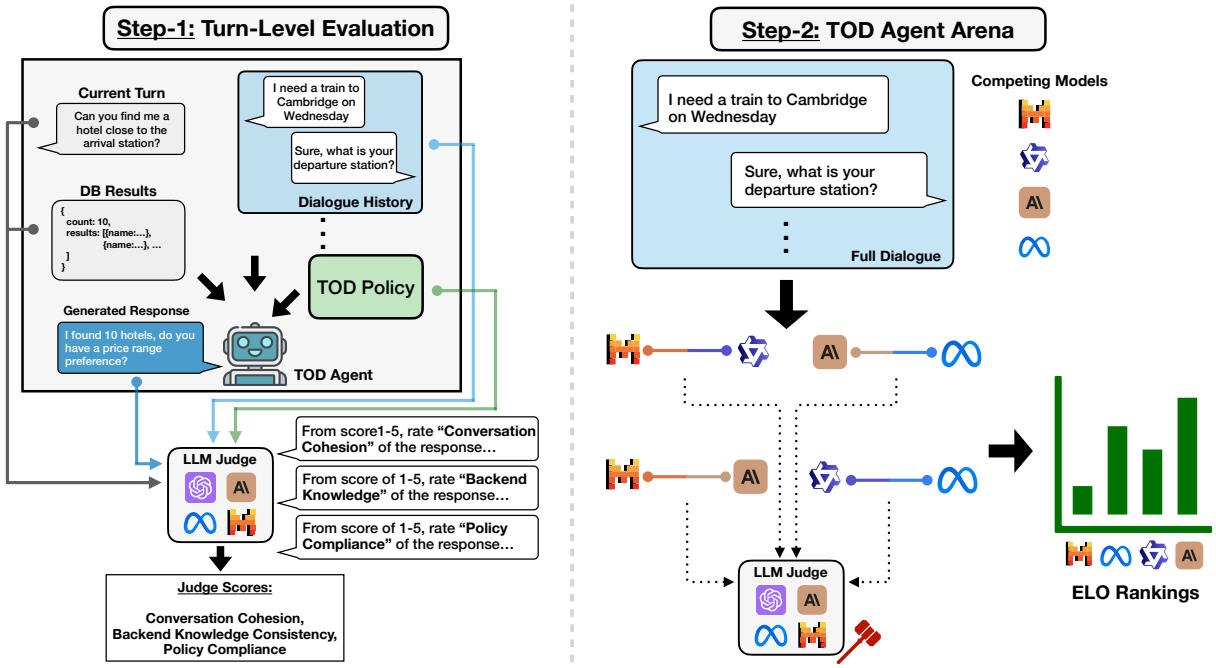

Figure 3: **Overview of the TD-EVAL Framework.** The left part illustrates the pipeline for turn-level evaluations, where a system is assessed across three criterion by an LLM-based judge. The right part depicts the second step, in which models compete within TOD Agent Arena and ranked based on win and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

dence 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 of TOD systems.

3 TD-EVAL

To comprehensively evaluate LLM performance on TOD tasks, it is critical to assess both turn-level and dialogue-level performances (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 from turn to dialogue-level with a pairwise ranking method. For each dialogue pair, the LLM 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 LLM-as-judge approach.*

3.1 Turn-Level Evaluation

As illustrated in the left-hand side of Figure 3, turn-level evaluation focuses on analyzing the response of each system 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, database result, and system response into our framework, which uses a user-selected LLM as the primary evaluator. Each response is rated on a 5-point scale across the three defined dimensions, 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 to 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 current 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 10).

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. 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 and database results into the LLM judge with instructions (See Figure 11).

Policy Compliance. A core requirement in TOD is following predefined, domain-specific policy protocols (e.g., when to request a user's booking details or how many suggestions to offer). To evaluate this, we check whether an agent's responses align with a policy prompt (Figure 12), 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 TOD Agent 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) does not 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 **TOD Agent 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 criteria, 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 with Elo. 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 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 (Figure 13), 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. 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 Human Evaluation

4.1 Human Evaluation Process

We conducted a comprehensive two-step human evaluation to validate TD-EVAL as an effective metric. The study involved 10 annotators with academic backgrounds in Computer Science at the graduate level. All annotators were proficient in English; several have prior experience in NLP research, making them well-suited for nuanced evaluation of TOD conversations. The study was conducted using fully synthetic conversations generated from MultiWOZ and τ-Bench goals with a user simulator. The data generation process followed the experimental setup described in Section 5.1. Annotators were presented with the dialogue history and the corresponding database re-

Metric	Gwet AC1	Randolph's κ
Conversation Cohesion	0.49	0.42
Backend Knowledge	0.66	0.63
Policy Compliance	0.64	0.58
Overall	0.56	0.54

Table 1: Inter-Annotator Agreement scores from the first step of human evaluation. "Overall" combines scores from all three sub-metrics.

sults, and were asked to provide both turn-level and dialogue-level ratings for the three sub-metrics (conversation cohesion, backend knowledge, and policy compliance) on a 5-point Likert scale, ranging from"Very Bad" to "Very Good". Detailed evaluation guidelines is provided in Figure 8. Refer to Appendix A.1 for further details.

In the first step of human evaluation, all 10 participants annotated the same 5 conversations (3 from MultiWOZ and 2 from τ-Bench). We observed a strong skew in the collected ratings, with nearly 90% of scores falling in the 4–5 range (see Figure 6). To account for this prevalence bias, we used Gwet's AC1 (Gwet, 2002) and Randolph's κ (Randolph, 2005) to calculate agreement, both of which offer more stable agreement measures than traditional κ statistics in such imbalanced settings (Finch and Choi, 2024). As shown in Table 1, the inter-annotator agreement scores are between 0.4 and 0.7, indicating a substantial level of agreement among annotators. For the second step, we followed up with 9 of the original annotators. Each annotator was assigned 10 distinct conversations (6 from MultiWOZ and 4 from τ-Bench), totaling of 90 annotated dialogues covering 490 turns.

4.2 Baseline Metrics

4.2.1 Traditional Metrics

For MultiWOZ, we use the traditional Success rate (Budzianowski et al., 2018), which is a stricter measure that subsumes Inform. For τ-Bench, we use τ-Bench reward (Yao et al., 2025).

4.2.2 LLM-as-judge

We use the state-of-the-art LMUnit (Saad-Falcon et al., 2024) metric as our LLM-as-judge baseline. LMUnit evaluates a dialogue response against natural language unit tests using a unified scoring model, outputting a score between 1 and 5. To ensure a fair comparison with TD-EVAL, we designed natural language unit test cases that capture our three sub-metrics (conversation consistency, backend knowledge correctness, and policy adherence). LMUnit scores are obtained via an API call, which takes the dialogue history, system response, and corresponding unit tests as inputs. We handcrafted unit tests for both the dialogue level and the three turn level sub-metrics, shown in Figure 9. The turn-level sub-metrics are calculated as the mean score of the corresponding unit tests. We refer to this extension of LMUnit as $\text{LMUnit}_{\text{TD}}$. Further details on the LMUnit test cases and API calls are provided in Appendix A.2.

Granularity	Metric	Gwet AC1	Randolph's κ
Turn-level	$\text{LMUnit}_{\text{TD}}$	0.43	0.41
	TD-EVAL	0.56	0.50
Dialogue-level	Traditional	0.41	0.34
	$\text{LMUnit}_{\text{TD}}$	0.44	0.40
	TD-EVAL	0.57	0.52

Table 2: Comparison of different evaluation methods with human scores, using combined "Overall" score from all three metrics.

4.3 Result

Table 2 compares human ratings (from the second step of human evaluation) with scores from traditional metrics, $\text{LMUnit}_{\text{TD}}$, and TD-EVAL, at both the turn and dialogue levels. Traditional metrics are shown only at the dialogue level, as they are computed over entire conversations. Additionally, we modified $\text{LMUnit}_{\text{TD}}$ to produce dialogue-level evaluations as well, details of which are provided in Appendix A.2. To assess alignment with human judgments, we again use Gwet's AC1 (Gwet, 2002) and Randolph's κ (Randolph, 2005). We compute the overall turn-level score by aggregating the three sub-metric scores and measuring agreement across all turns. In Table 2, it can be seen that TD-EVAL shows stronger alignment with human judgments compared to traditional Success rate and τ-Bench reward. Moreover, across both turn-level and dialogue-level evaluations, TD-EVAL consistently outperforms $\text{LMUnit}_{\text{TD}}$. Thus, TD-EVAL offers a more reliable and human-aligned assessment of conversational quality.

5 Experiments

5.1 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., 2025), both originally evaluated using automatic metrics. We

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

Ranking	Model	TOD 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 4: Dialogue-level results and TOD Agent Arena Leaderboard.

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 9 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 the default model in both the LLM-based Judge and User Simulator.

User Simulators. In our evaluation framework, we employed online user simulators to facilitate interactive assessment of dialogue agents. For experiments on MultiWOZ 2.4, we utilized the AutoTOD (Xu et al., 2024) user simulator and for τ-Bench evaluations, we leveraged the benchmark's official user simulator as provided by the authors (Yao et al., 2025). Both simulators were powered by GPT-4o as the underlying LLM.

5.2 Main Results

Turn-level Evaluation Results. Table 3 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 placing third as the top open-source model. Notably, Llama-3.1-405B achieves the highest score in *backend knowledge* (4.4182), outperforming proprietary models. Table 3 indicates that while certain models perform well in cohesion and factual correctness, they may fail to strictly enforce policies. This highlights the need for a comprehensive evaluation in multiple dimensions.

Dialogue-level Evaluation Results. Table 4 summarizes the pairwise Elo rankings under our TOD Agent 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 dialogue-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

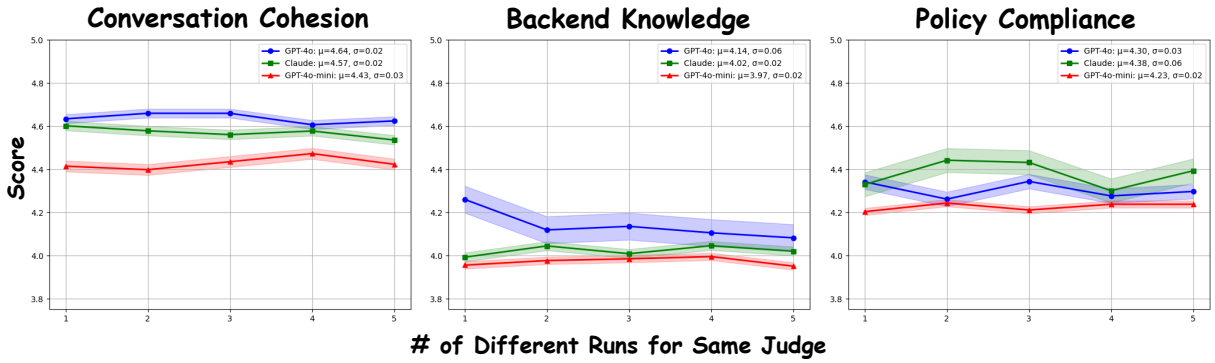

Figure 4: **LLM Judge Consistency Across Multiple Runs.** Evaluation results for three metrics using the same LLM judge over five runs: conversation cohesion (left), backend knowledge (middle), and policy compliance (right).

performance. In contrast, GPT-3.5-turbo sits at the bottom, aligning with its lower per-turn scores.

5.3 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.

TD-EVAL scores are robust to different LLM judges. The TD-EVAL scoring remains stable across diverse LLM judges. In Figure 5 (Appendix B), we compare different agents with three different LLM judges. We observe that all three judges produce closely aligned absolute scores. Refer to Appendix B for further details.

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) 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 (Engelbrech et al., 2009; Higashinaka et al., 2010; Ultes and Minker, 2014; Georgila, 2024) and dialogue level evaluation approaches (Bodigutla et al., 2020; Xu et al., 2024), developing a comprehensive evaluation metric capable of assessing the overall interaction quality remains an open problem, particularly for TOD systems built on LLMs. 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 for TOD systems based on LLMs.

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 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 intend to 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 used alongside existing metrics in 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.

Acknowledgement

This work was supported in part by Other Transaction award HR0011249XXX from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program and has benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program, through which leading foundation models hosted by Microsoft Azure and access to Azure credits were provided to conduct the research.

References

Emre Can Acikgoz, Cheng Qian, Hongru Wang, Vardhan Dongre, Xiusi Chen, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. 2025. A desideratum for conversational agents: Capabilities, challenges, and future directions. *arXiv preprint arXiv:2504.16939*.

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.

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.

Klaus-Peter Engelbrech, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling user satisfaction with hidden markov model. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, page 170–177, USA. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2024. ConvoSense: Overcoming monotonous commonsense inferences for conversational AI. *Transactions of the Association for Computational Linguistics*, 12:467–483.

Kallirroi Georgila. 2024. Comparing pre-trained embeddings and domain-independent features for regression-based evaluation of task-oriented dialogue systems. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 610–623, Kyoto, Japan. Association for Computational Linguistics.

Kilem L. Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters.

Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In *Spoken Dialogue Systems for Ambient Environments*, pages 48–60, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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 Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Dong, Adithya Sagar, Xifeng Yan, and Paul Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests. *Preprint*, arXiv:2412.13091.

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.

Stefan Ultes and Wolfgang Minker. 2014. Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 208–217, Philadelphia, PA, U.S.A. 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 R Narasimhan. 2025. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*.

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.

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.

A Additional Human Evaluation Details

This section provides the additional details related to human evaluation described in Section 4.

A.1 Human Evaluation Setup

Human evaluation was conducted through Qualtrics, an online survey platform. Ten annotators were recruited for evaluation via convenience sampling, where the authors selected participants from their personal and professional networks. We include a snapshot of the survey interface that was used for human evaluation in Figure 7. We provide detailed instructions to human annotators, shown in Figure 8, as reference when evaluating conversations. We observed a strong skew in the collected ratings, with nearly 90% of scores falling in the 4–5 range, which is shown in Figure 6.

A.2 LMUnit Scoring

The natural language unit tests and API calls [2] to compute the LMUnit scores is shown in Figure 9. LMUnit (Saad-Falcon et al., 2024) scores are returned as floating point numbers with decimal points, while Traditional and TD-EVAL scoring is integer-based. To address this, LMUnit scores are rounded to the nearest integer after averaging the unit test scores. Turn-level scores are calculated based on the three sub-metrics to calculate agreement with TD-EVAL scores. In the dialogue level, there is only one score that broadly measures goal completion performance of the TOD agent for the overall conversation. To mirror traditional metrics, we change LMUnit to evaluate based on goal completion rather than the original sub-metrics. This score is also calculated on a [1, 5] scale.

A.3 Additional Results

In this section, we perform an extended analysis of the turn and dialogue-level results in Table 2 to find more fine-grained human agreement results of each evaluation metric. Table 5 shows the turn-level agreement results split by TD-EVAL sub-metrics. Table 6 lays out the dialogue-level agreement results split by the dataset (MultiWOZ or τ-bench). Besides the traditional Success rate, we also report AutoTOD's online Success rate (Xu et al., 2024) that operates on raw, non-delexicalized responses. In Table 6, $\text{Success}_{\text{OR}}$ refers to the traditional Success rate computed with delexicalized responses, while $\text{Success}_{\text{AT}}$ denotes the online Success rate. For both turn and dialogue level, TD-EVAL consistently shows greater alignment with human judgment.

Metric	Sub-Metric	Gwet AC1	Randolph's κ
$\text{LMUnit}_{\text{TD}}$	Conversation Cohesion	0.45	0.43
	Backend Knowledge	0.42	0.40
	Policy Compliance	0.41	0.39
	Overall	0.43	0.41
TD-EVAL	Conversation Cohesion	0.56	0.51
	Backend Knowledge	0.55	0.50
	Policy Compliance	0.56	0.50
	Overall	0.56	0.50

Table 5: Extended comparison of turn-level performance of $\text{LMUnit}_{\text{TD}}$ and TD-EVAL with human ratings, split by TD-EVAL sub-metrics.

Dataset	Metric	Gwet AC1	Randolph's κ
MultiWOZ	$\text{Success}_{\text{OR}}$	0.40	0.33
	$\text{Success}_{\text{AT}}$	0.47	0.34
	$\text{LMUnit}_{\text{TD}}$	0.49	0.47
	TD-EVAL	0.63	0.59
τ-Bench	Reward	0.43	0.33
	$\text{LMUnit}_{\text{TD}}$	0.34	0.30
	TD-EVAL	0.47	0.42

Table 6: Extended comparison of dialogue-level. Scores are split up by dataset type. $\text{Success}_{\text{OR}}$ indicates the traditional Success rate computed with delexicalized responses, while $\text{Success}_{\text{AT}}$ denotes the online Success rate without any delexicalization of system response.

B Additional Ablation Study

For this ablation study, 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. Furthermore, the relative ranking of the models evaluated remains consistent between 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.

[2] contextual.ai/lmunit/

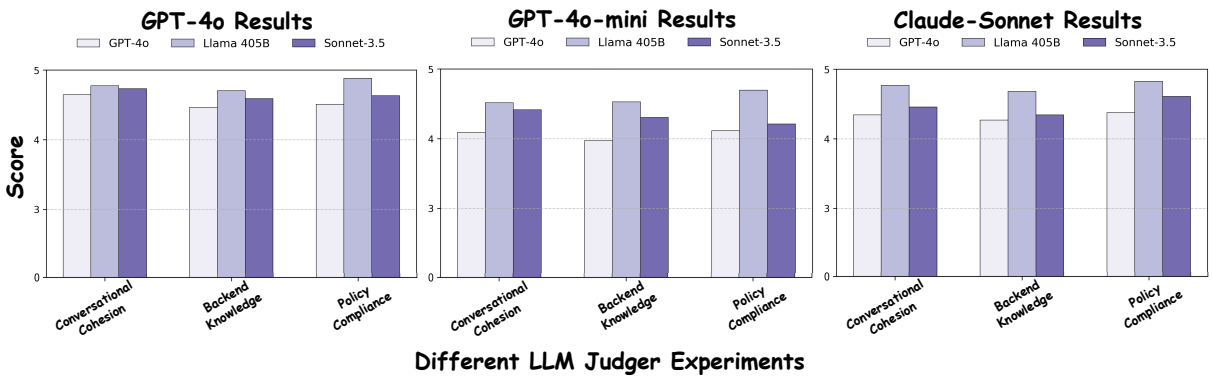

Figure 5: **Different LLM Judge Experiments.** Evaluation scores of GPT-4o (left), GPT-4o-mini (center), and Claude-Sonnet-3.5 (right) under different LLM judges (GPT-4o, Llama-405B, and Claude-3.5-Sonnet).

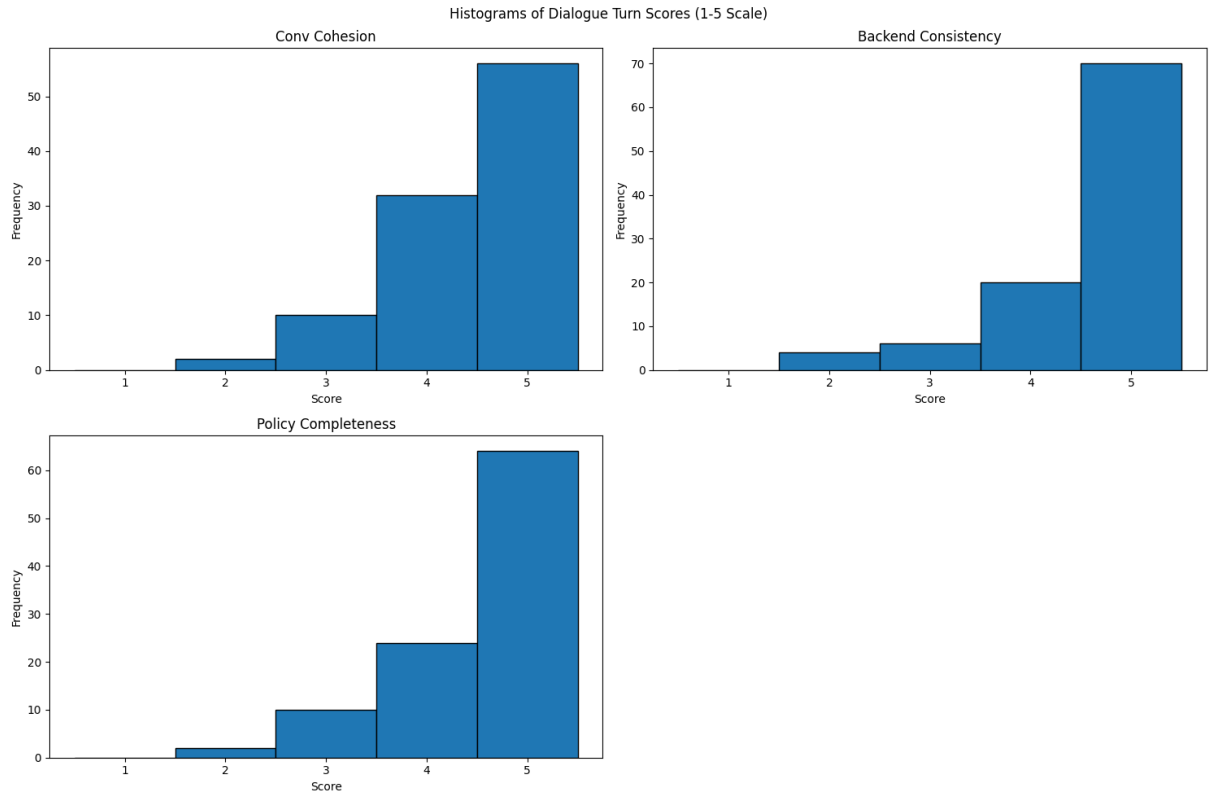

Figure 6: **Human Evaluation Scores Distribution** This is the distribution of all scores from 10 annotators during the inter-annotator agreement evaluation stage, separated by the three metrics. For all three TD-Eval metrics, the vast majority of LLM responses were classified as 4 (Good) or 5 (Very Good), resulting in class distribution imbalance.

Agent: I found the restaurant "Golden Wok" for you. It is located at 191 Histon Road, Chesterton, in the north area. The restaurant serves a variety of traditional Chinese meals, including sweet and sour dishes, rice, and noodles. The price range is moderate, and you can contact them at 01223350688.
User: I would like to book a table for 2 people at Golden Wok at 18:15 on Monday, please.

Backend Result:
Thought: I need to call an API.
Function: book_restaurant
Args: {"name": "golden wok", "people": "2", "day": "monday", "time": "18:15"}
Output:

- {'reference': '8ODLVWDR'}

Thought: I don't need an API and want to respond to the user.

Agent: I have successfully booked a table for 2 people at Golden Wok at 18:15 on Monday. Your reference number is 8ODLVWDR.

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	○	○	○	○	○

Figure 7: **Human Evaluation Interface:** This is the interface showing an example evaluation question in the human evaluation study. We conduct the evaluation on Qualtrics.

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

Turn-Level Unit Tests for Conversation Cohesion

1. **Accuracy**: Does the response directly relate to the dialogue history and the current user query?
2. **Topic Consistency**: Does the response remain on-topic with the dialogue history and the user query?
3. **Coherence**: Does the response logically continue the progression of the dialogue?

Turn-Level Unit Tests for Backend Knowledge Consistency

1. **Accuracy**: Does the response accurately reflect the information in the database results?
2. **Topic Consistency**: Does the response stay on-topic with the database results and the dialogue context?
3. **Coherence**: Does response logically incorporate and progress based on the database results?

Turn-Level Unit Tests for Policy Compliance

1. **Number of Suggestions**: Does the response provide suggestions only when the database results are few enough to do so?
2. **Information Gathering**: Does the response request required, relevant information from the user before offering suggestions or booking services?
3. **Appropriate Timing**: Does the response avoid premature actions (i.e. make a booking or suggest a service) too early in the conversation, before the necessary information is gathered?

API call template for turn-level score

{
“query": [CONV_HISTORY] \n “Database Result:" [DB_RESULT]
“response": [AGENT_RESPONSE]
“unit_test": [UNIT_TEST_QUESTION]
}

Dialogue-Level Unit Test for Goal Completion

1. **Goal Completion:** Does the conversation achieve it’s intended goal?

API call template for dialogue-level score

{
“query": “Goal: " [CONV_GOAL]
“response": [CONV_HISTORY]
“unit_test": “Does the conversation achieve it’s intended goal?"
}

Figure 9: Natural language unit tests for computing the $\text{LMUnit}_{\text{TD}}$ scores at the turn and dialogue level. For the turn level, we calculate scores for the three sub-metrics. Scores from each unit test are between [1, 5]. For the dialogue metric, we only ask whether the conversation achieved it’s intended goal, which is supplied in both MultiWOZ and τ-bench datasets. We provide the API call template to LMUnit as additional context.

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 10: The GPT-4o prompt used for evaluating Conversational Consistency on MultiWOZ 2.4.

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 11: The GPT-4o prompt used for evaluating Backend Knowledge on MultiWOZ 2.4.

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 12: The GPT-4o prompt used for evaluating Policy Compliance Consistency on MultiWOZ 2.4 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 13: The GPT-4o prompt used for dialogue-level evaluation in TOD Agent Arena.