Modeling Long-Term Memory for Multi-Session Task-Oriented Dialogue Systems via Memory-Active Policy

Anonymous ACL submission

Abstract

Existing Task-Oriented Dialogue (TOD) systems generally focus on single-session dialogues and overlook the study of multi-session interactions, leading to the inability to track long-term memory to obtain target-related information from previous dialogue sessions for more efficiently personalized interaction in TOD. To address this challenge, we introduce a **MS-TOD** dataset, the first multi-session TOD dataset designed to retain long-term memory across sessions, enabling fewer turns and more efficient task completion. Based on this new dataset, we propose a **Memory-Active Policy (MAP)** that improves multi-session dialogue efficiency by reducing turns through a two-stage approach. Specifically, we first introduce Memory-Guided Dialogue Planning, which retrieves relevant history through intent descriptions, utilizes a memory judger to identify key QA units, and employs a reader to generate responses based on reconstructed memory. Next, the Proactive Response Strategy is designed to detect and correct errors or omissions, ensuring efficient and accurate task completion. We evaluate MAP on our MS-TOD dataset, focusing on response quality and effectiveness of the proactive strategy. Experimental results show that MAP enhances multi-session TOD performance by improving turn efficiency and task success through long-term memory integration while maintaining comparable performance in single-session multi-turn tasks.

1 Introduction

Task-oriented dialogue (TOD) systems (Wang et al., 2021; He et al., 2022; Bang et al., 2023a; Swamy et al., 2023a) have traditionally focused on single-session scenarios, overlooking the fact that real world interactions often span multiple sessions over extended periods. Although Large Language Models (LLMs) have been explored to enhance TOD performance (Xu et al., 2024a,b; Chung et al., 2023; Heck et al., 2023a), these works mainly focus on

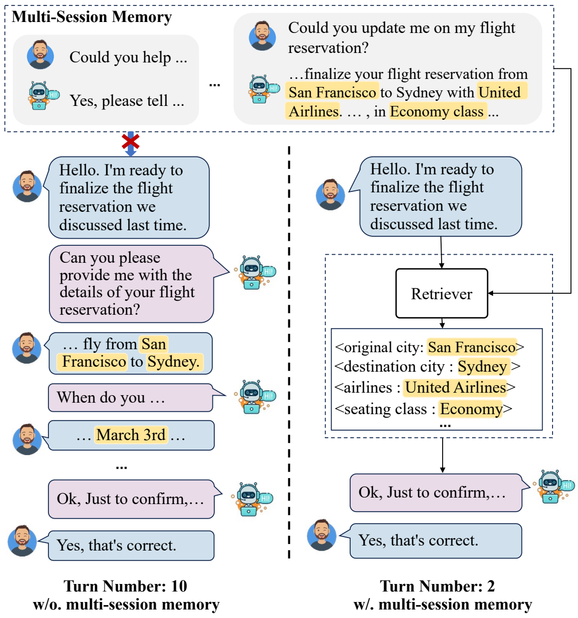

Figure 1: Task-oriented dialogue, with (*right*) vs without (*left*) multi-session memory; the latter demands more turns of conversation.

single-session conversations and do not adequately consider personalization and long-term engagement. Moreover, existing TOD datasets (Stacey et al., 2024; Liu et al., 2024; Budzianowski et al., 2018; Rastogi et al., 2020a) are confined to single-session settings, leaving a gap in publicly available benchmarks for evaluating long-term memory retention across multi-sessions.

To illustrate the limitations of single-session systems, consider Figure 1, which compares two dialogue snippets from a TOD system **without** multi-session memory and **with** multi-session memory. In the first case, the user must repeatedly restate details (e.g., flight times, seat preferences) in every new session, resulting in inefficiency and user frustration. By contrast, when multi-session memory is integrated, the system seamlessly retrieves itineraries and preferences from earlier sessions, eliminating redundant interactions and providing a more personalized experience.

To bridge this gap, we introduce the Multi-session Task-oriented Dialogue Dataset (MS-TOD), which consists of hundreds of users, each engaging in over 20 sessions with diverse task goals sourced from SGD (Rastogi et al., 2020a). MS-TOD supports comprehensive evaluation of TOD systems to retrieve long-term context, maintain consistent task slots, and adapt responses to individual user preferences. However, effectively leveraging long-term memory in multi-session TOD remains challenging. While existing approaches in open-domain multi-session conversations focus on retrieving dialogue history or summaries (Lu et al., 2023; Zhong et al., 2024; Joko et al., 2024; Li et al., 2024a; Du et al., 2024), multi-session TOD system face additional demands: they must recall critical slot-value pairs, track evolving user intents, and proactively resolve missing or outdated information while minimizing redundant user queries.

Therefore, we propose the **Memory-Active Policy (MAP)** to incorporate long-term memory in multi-session TOD tasks. MAP consists of two core phases: (1) **Memory-Guided Dialogue Planning**, where an LLM generates an intent hypothesis and aligns it with structured memory entries to align user goals across sessions. Furthermore, a memory judger assesses relevant memory units and refines task slot descriptions for precise, context-aware responses. (2) **Proactive Response Strategy**, which iteratively detects missing or mismatched slots by comparing predicted responses with task goals, actively engaging users to resolve incomplete slots, thereby reducing redundancy and ensuring smooth, goal-oriented interactions. Experimental results on MS-TOD demonstrate that MAP effectively improves dialogue coherence, response quality, task success rate, and dialogue efficiency in multi-session TOD.

The main contributions include:

- We introduce MS-TOD, the first multi-session task-oriented dialogue dataset for evaluating dialogue systems in long-term interactions.
- We propose MAP framework, which integrates long-term memory into TOD systems for efficient task completion in minimal dialogue turns.
- Experiments show that MAP outperforms baselines in most metrics, validating its active memory mechanism.

Settings	GPT-4 Score	Slot Acc.
No Retrieval (Direct Prompting)		
Current Session Context	2.60	0.13
Full Conversation Context	4.76	0.61
Retrieval-Augmented Generation		
BM25-Based Retrieval	5.90	0.53
Embedding-Based Retrieval	7.01	0.67
Hybrid Retrieval	7.04	0.68
Oracle (Upper Bound)		
Oracle	**8.51**	**0.82**

Table 1: Evaluation of confirmation-type response generation under different prompting and retrieval strategies.

2 Preliminary Experiments

To investigate the effectiveness of different strategies for handling dialogue history in multi-session task-oriented response generation, We conduct a preliminary study comparing direct prompting (Swamy et al., 2023b; Xu et al., 2024a) with retrieval-augmented generation (RAG) (Huang et al., 2024; Lu et al., 2023) in multi-session TOD.

Because standard TOD datasets lack multi-session dependencies, we construct a test set specifically for *confirmation-type* response generation (details in Section 3). Our pipeline includes (1) **Retrieval.** We explore three strategies for retrieving relevant historical dialogues: *sparse retrieval* (BM25 (Robertson and Zaragoza, 2009)), *dense retrieval* (text-embedding-small-3[1]), and a *hybrid* approach that combines both to leverage their complementary strengths. (2) **Response Generation.** GPT-4o-mini then generates confirmation-type responses by incorporating the retrieved dialogue history and task goal information.

As shown in Table 1, RAG consistently outperforms direct prompting. For instance, *dense retrieval* achieves 0.67 slot accuracy and a 7.01 GPT-4 score, surpassing full-context prompting (0.61 and 4.76, respectively). *Hybrid retrieval* further improves slot accuracy to 0.68 and the GPT-4 score to 7.04, demonstrating the value in combining sparse and dense strategies. Oracle retrieval (using ground-truth context) reaches 0.88 and 8.51, underscoring **the need for more accurate retrieval strategies in multi-session TOD**.

[1]OpenAI. text-embedding-3-small. 2025. OpenAI, https://platform.openai.com/docs/guides/embeddings.

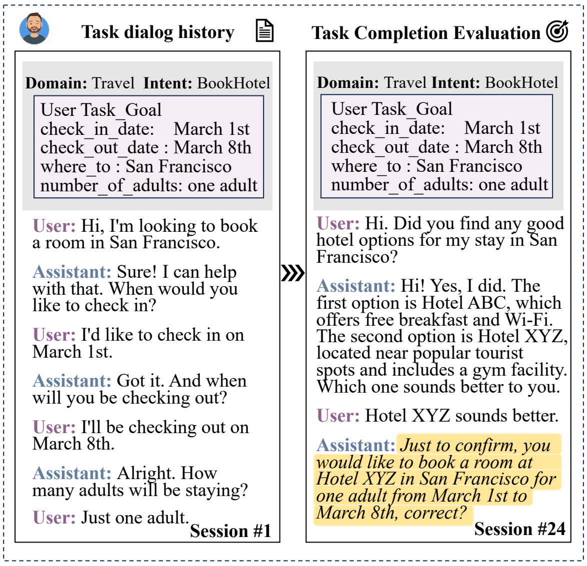

Figure 2: An Example of MS-TOD dataset.

Attribute	Evaluation
Domains	16
Intentions	19
Task goals	956
Dialogues	2,861
Utterances	18,530
Avg. slots per task goal	4.24
Number of individuals	132
Avg. intentions per individual	5.45
Avg. sessions per individual	21.67
Avg. Utterances per individual	140.38

Table 2: MS-TOD dataset statistics for evaluation.

3 Dataset

To systematically evaluate TOD systems in scenarios requiring multi-session long-term memory integration, we develop the MS-TOD dataset, derived from the Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020a). MS-TOD comprises two subsets: a training subset for training the memory judger (Section 4.2) and an evaluation subset designed to assess multi-session memory activation and TOD response generation shown in Figure 2.

3.1 Data Generation

Multi-Session Dialogue Construction. Because existing TOD corpora typically feature single-session interactions lacking structured multi-session dependencies, we create *three* dialogue sessions for each task in the SGD dataset. Compared with single-session dialogues, this design more closely simulates how users revisit and refine the same task at different times and in different contexts. We chose three sessions—rather than a higher number—to strike a balance between capturing realistic user behavior and avoiding repetitive dialogue data, particularly given that SGD tasks involve fewer than ten task slots. As a result, three sessions offer sufficient coverage of task variations without overpopulating the dataset. More details can be found in Appendix A.1.

Confirmation-Type Response Annotation. In the final session of each task, we introduce confirmation-type annotations to mark utterances indicating the completion of long-term or recurring tasks. These annotations serve two primary functions: (1) **Guiding Memory Activation**: Highlighting key dialogue points to trigger long-term memory activation, summaries, or confirmations; and (2) **Supporting System Evaluation**: They enable evaluation of the system's ability to recognize and record cross-session information or long-term goals during dialogue strategy assessment.

3.2 Individual Memory Bank Construction

Since multi-session interactions occur at the individual level, we group sessions into *Individual Memory Banks* (Figure 2), each storing an individual's historical dialogues for maintaining continuity and adapting responses. Each bank contains over 20 sessions spanning more than six distinct user intentions (Table 2), plus a dedicated evaluation session per intention requiring confirmation responses. Task goals are also provided to guide system outputs, supporting effective memory activation and task handling in diverse scenarios.

To refine these memory banks, we employ a GPT-4-based generator that extracts high-level intent descriptions and creates task-specific QA pairs (Appendix A.2). These structured QA pairs enable efficient retrieval of relevant contexts, allowing the system to selectively activate memories and adapt dynamically to user needs for multi-domain, intention-aware TOD. More details on the dataset can be found in Appendix B.

4 Memory-Active Policy

To address the need for long-term memory and multi-session context in TOD, the Memory-Active Policy (MAP) combines memory-driven dialogue planning with a proactive policy strategy as shown in Figure 3.

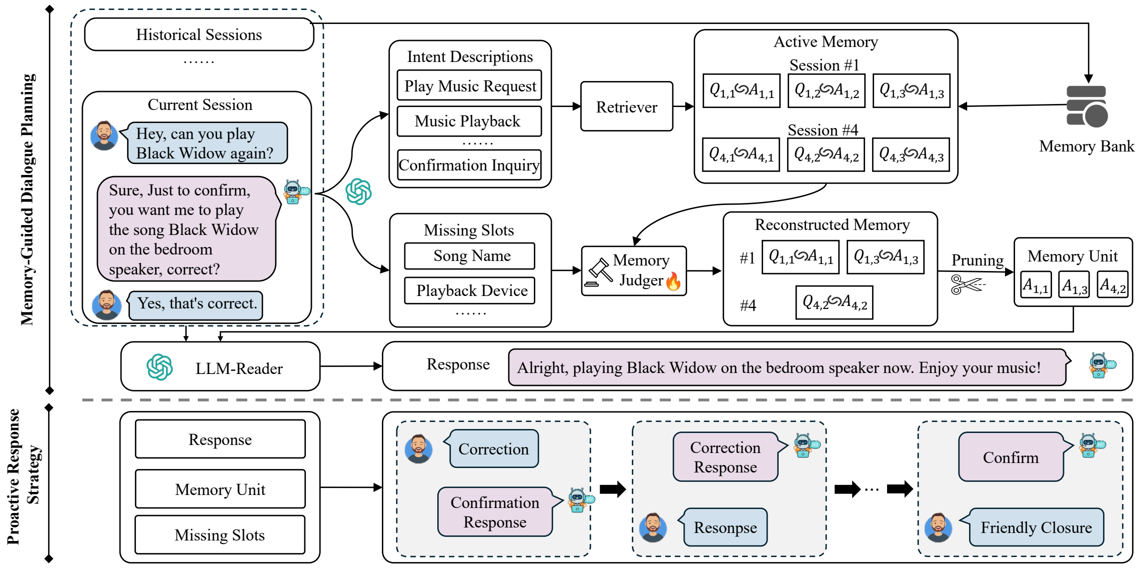

Figure 3: Overflow of our MAP framework, which comprises Memory-Guided Dialogue Planning and Proactive Response Strategy.

4.1 Definition

The objective of this task is to generate a natural language response r based on the provided dialogue context c and individual memory bank M. The dialogue context c represents the ongoing interaction, comprising chronologically ordered user utterances u_j and system responses r_j. The individual memory bank M consists of memory representations from multiple dialogue sessions, where each session provides an intent description k_i and a corresponding set of QA pairs v_i. Formally, we define: $M = \{(k_i, v_i)\}_{i=1}^{N}, v_i = \{(q_{ij}, a_{ij})\}_{j=1}^{n}$. Here, k_i represents the session's high-level intent, while v_i stores detailed task-related information. This structured design facilitates efficient retrieval and utilization of long-term user-specific memory.

The response r is generated by a large language model LLM that integrates c and M, ensuring semantic coherence, memory relevance, and task slot accuracy. Formally, the optimal response is obtained by maximizing the conditional probability distribution:

$$r^* = \arg\max_{r \in \mathcal{R}} P(r \mid c, M), \quad (1)$$

where $\mathcal{R}$ denotes the set of all possible responses. This approach emphasizes understanding the dialogue context and leveraging individual memory to produce coherent and relevant responses.

4.2 Memory-Guided Dialogue Planning

Memory-Guided Dialogue Planning consists of two key steps: (1) Intent Capture and Memory Activation, where the system identifies and retrieves relevant memories aligned with the user's intentions; (2) Memory Judgement and Refinement, which detects missing task slots, and re-ranks relevant memories to ensure optimal information recall for response generation.

Intent Representation and Memory Activation. Given the dialogue context c_i, we use LLM (GPT4o-mini) to generate a high-level intent description k_i, which summarizes the user's objective in the current session. The intent description k_i is then used to retrieve relevant memory units from the long-term memory M, represented as $M = \{(k_j, v_j)\}_{j=1}^{m}$, where k_j is an intent-related key and v_j is the corresponding structured information, such as paired questions and answers. Using an embedding model, k_i is mapped to a dense representation and compared with k_j to activate the most relevant memory units v_i. These activated memory units v_i, containing structured information such as task-related questions and answers, are then used to guide subsequent dialogue processing.

Memory Judgement. Accurately recognizing the task goal is crucial for invoking relevant memory and formulating an effective dialogue policy. To refine retrieved memories, we employ a Chain-of-Thought (CoT) (Wei et al., 2022) mechanism, which identifies the task goal and generates missing task slot queries. The retrieved QA pairs are then evaluated by a memory judger, trained on LLaMA 3.1-8B[2], to assess their relevance. Given a dialogue context c_i, a missing query q_{miss}, and the relevant memory QA pairs v_j. The output indicates if the QA pairs under intent k_i sufficiently answer q_{miss}.

Let j index memory units and u index QA pairs within the j-th unit. The memory judger evaluates each QA pair $(q_{j,u}, a_{j,u})$ as:

$$s_{j,u} = P(y = 1 | (c_i, q_{\text{miss}}, q_{j,u}, a_{j,u})) \quad (2)$$

where $y = 1$ indicates that the QA pair contributes to the task goal, while $y = 0$ indicates irrelevance. The memory judger LLM_{MJ} is trained using a cross-entropy loss function:

$$\mathcal{L} = -\sum_{(q_{j,u}, a_{j,u})} [y \log s_{j,u} + (1 - y) \log(1 - s_{j,u})] \quad (3)$$

This formulation ensures the judger assigns higher scores to QA pairs that are more relevant to the missing query q_{miss}. Next, all retrieved memory QA pairs are re-ranked based on updated scores. For each QA pair, the final score $s_{\text{f},ju}$ is calculated by combining the previous retrieval score $s_{\text{prev},ju}$ and the judger's score s_{ju} using a weighted sum:

$$s_{\text{f},ju} = \alpha \cdot s_{\text{prev},ju} + (1 - \alpha) \cdot s_{ju}, \quad (4)$$

where α is a weight parameter balances retrieval relevance and judger evaluation. The top 5 QA pairs $v_{selected}$ with the highest $s_{\text{final},ju}$ scores are selected to ensure contextual relevance and task alignment in subsequent dialogue steps.

Memory Refinement mainly contains memory pruning and memory reconstruction. Memory pruning filter activated memory units $\{v_j\}$ by removing redundant components (e.g., auxiliary questions $q_{j,u}$) tied to intent k_i, retaining only core answers $A_{\text{core}} = \{a_{j,u}\}$. We reconstruct answers A_{core} into the dialogue context c through concatenation ensuring noise removal prior to contextualization.

4.3 Proactive Response Strategy

The response generation phase synthesizes the dialogue context c and pruned memory A_{core} (from memory reconstruction) into a confirmation response r. Using $\text{LLM}_{\text{Reader}}$, the system evaluates if integrated memory supports task completion:

$$r = \text{LLM}_{\text{Reader}}(c, A_{\text{core}}) \quad (5)$$

where r serves dual purposes: (1) providing task guidance and (2) explicitly verifying memory relevance to user goals (see Appendix A.3 for details).

To provide more comprehensive responses to user queries, we propose a proactive dialogue policy. Based on the generated response r, we identify missing or incorrect slots within the dialogue. This results in a set of slots, denoted as $L = \{l_1, l_2, ..., l_n\}$, where each l_i represents a missing or erroneous slot. We design an agent to simulate the user, explicitly informing it of the slot set L. The user agent then interacts with our dialogue model in an interactive conversation to address the identified slots.

At each dialogue turn, a supervisor (played by an LLM) evaluates whether the conversation accurately fulfills the slot information requirements. If a slot s_i is successfully resolved during the interaction, it is removed from L. Mathematically, the update to the slot set is expressed as:

$$S \leftarrow L \; \{l_i\} \quad (6)$$

The interaction continues for multiple turns until the slot set becomes empty, $L = \emptyset$, ensuring all missing or erroneous slots are resolved.

5 Experiments

5.1 Experimental Setups

Evaluation Settings. Our evaluation primarily focuses on **GPT-4 score**[3], **Joint Goal Accuracy (JGA)**, **Dialogue Turn Efficiency (DTE)**, and **Success Rate (S.R.)** as key performance metrics. DTE reflects the system's efficiency by measuring the number of turns required to complete a task, where a lower value indicates more effective interactions. To assess memory activation, we include **Recall@k** to evaluate the retrieval of relevant long-term historical context. Additionally, we conduct human evaluation to assess response accuracy, informativeness, and coherency. For further insights into task completion accuracy and response quality, we report **Slot Accuracy**, **BLEU** (Papineni et al., 2002), and **ROUGE** (Lin, 2004)

Baselines We conduct comparisons with state-of-the-art conversational approaches using different large language models, such as LLaMA3-8B (Touvron et al., 2024), Qwen2.5-7B (Team, 2024c), Mistral-7B (Team, 2024a), and GPT-4o-mini (Team, 2024b). Furthermore, we conduct a

[2]Meta. (2024). Llama 3.1: A Family of Open and Efficient Multilingual Language Models. Meta AI. Retrieved from https://llama.meta.com/

[3]GPT4-as-the-judge prompts can be found in Appendix A.4

Model	Setting	GPT4	JGA	DTE	S.R.
LLaMA3-8B	w/o MAP	4.89	**0.64**	5.37	0.82
	w/ MAP	**6.39**	0.63	**3.46**	**0.92**
Qwen-7B	w/o MAP	6.26	0.66	4.93	0.83
	w/ MAP	**6.81**	**0.66**	**4.31**	**0.87**
Mistral-7B	w/o MAP	6.20	0.73	2.52	1.00
	w/ MAP	**6.48**	**0.80**	**1.21**	**1.00**
GPT4o-mini	w/o MAP	6.93	0.67	6.03	0.88
	w/ MAP	**7.14**	**0.70**	**3.19**	**0.99**

Table 3: Performance comparison of task-oriented dialogue models with and without long-term memory integration. The w/o MAP setting uses full-context prompting, feeding the entire dialogue history as input, while w/ MAP leverages memory active policy to retrieve and utilize relevant long-term memory.

comparison with task-oriented dialogue methods in the context of dialogue state tracking (DST), including BERT-DST (Chao and Lane, 2019), AutoTOD (Xu et al., 2024a), and LDST (Feng et al., 2024), to evaluate the adaptability of our approach in task-specific dialogue scenarios. To evaluate the effectiveness of memory activation, we compare our method against various retrieval methods, including BM25 (Robertson and Zaragoza, 2009), T5 (Raffel et al., 2020), BERT-base, BERT-large (Devlin et al., 2018), nv-embed-v2 (Lee et al., 2024), bge-large-en-v1.5 (Liu et al., 2023), and text-embedding-3-small (OpenAI, 2023).

5.2 Main Results

Overall Performance. We conduct the experiments comparing full context prompting and our MAP framework in the metric of GPT4, JGA, DTE, and S.R. As shown in Table 3, MAP demonstrates consistent performance gains over baseline prompting methods. For instance, applying MAP to Mistral-7B increases JGA from 0.73 to 0.80 and S.R. from 0.83 to 0.87. Notably, LLaMA3-8B, Qwen-7B, and GPT-4o-mini also show significant improvements in both JGA and S.R. when integrated with MAP. In terms of response quality, **GPT-4 scores** rise notably for all models; for example, LLaMA3-8B achieves the largest gain, from 4.89 to 6.39. Regarding **DTE**, MAP considerably shortens the required turns, with reductions of 35.6% for LLaMA3-8B, 12.6% for Qwen-7B, 52.0% for Mistral-7B, and 47.1% for GPT-4o-mini. These results demonstrate that **integrating long-term memory enhances both response quality and conversation efficiency.**

Human Evaluation. We conduct a human evaluation to further assess the effectiveness of the MAP structure, as presented in Table 5. The evaluation focuses on confirmation-type responses and multi-turn dialogues adopting a proactive response strategy. Accuracy, informativeness, and coherence serve as evaluation metrics, with their average, denoted as A.I.C., representing overall performance. The results indicate that the MAP structure consistently enhances response quality, reinforcing the primary experimental findings. Further details are provided in Appendix C.2.

Comparision with Traditional TOD Models. Since no dedicated multi-session TOD model is available, we validate MAP using DST models as baselines. AutoTOD, which retains a full dialogue pipeline, allows evaluation across all four metrics, while LDST and BERT-DST are limited to JGA. As shown in Table 4, MAP outperforms these models, achieving the highest GPT-4 score of 7.14 and a JGA of 0.698, significantly surpassing AutoTOD at 0.440. Additionally, MAP improves efficiency, reducing DTE to 3.19 turns compared to 7.8 for AutoTOD. These results highlight the advantages of **multi-session memory integration and a proactive response strategy in improving both accuracy and efficiency.**

Model	GPT4	JGA	DTE	S.R.
Bert-DST*	-	0.067	-	-
LDST*	-	0.234	-	-
AutoTOD†	6.49	0.440	7.80	0.81
MAP	**7.14**	**0.698**	**3.19**	**0.99**

Table 4: Performance comparison of traditional TOD models and MAP. Models marked with * focus on DST, predicting slot-value pairs without handling dialogue management or task execution, making them unsuitable for evaluating S.R., GPT-4 score, and DTE. † indicates simplified AutoTOD.

Model	Confirmation		Multi-Turn	
	w/o MAP	w/ MAP	w/o MAP	w/ MAP
LLaMA3-8B	1.64	**1.99**	1.60	**2.03**
Qwen-7B	1.46	**1.88**	1.48	**1.77**
Mistral-7B	1.79	**1.99**	2.04	**2.18**
GPT4o-mini	1.86	**2.27**	1.72	**1.85**

Table 5: Human evaluation results based on the average A.I.C., which is the mean of Accuracy, Informativeness, and Coherence. w/ denotes with, w/o denotes without.

5.3 Ablation Study

Memory Judger for Improved Retrieval. We explore the impact of the Memory Judger on Recall@k by filtering out irrelevant memory units,

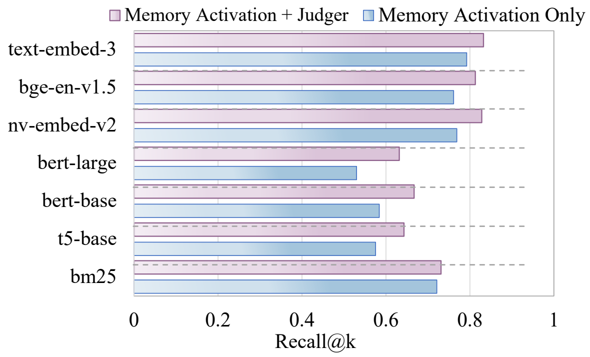

Figure 4: Impact of memory judger on memory activation performance across different embedding models.

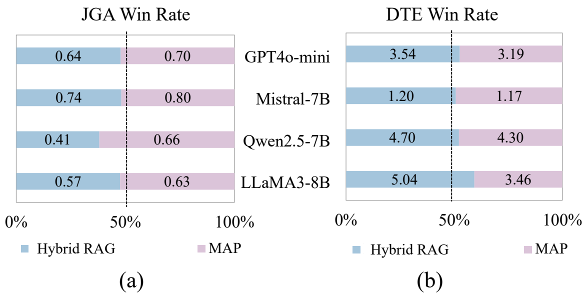

Figure 5: Comparison of Hybrid RAG vs. MAP across four LLMs on two metrics: (a) JGA Win Rate and (b) DTE Win Rate.

with k=5. As shown in Figure 4, integrating the Memory Judger into our Memory Activation module improves performance by 9.4%. Specifically, BM25's score rises from 0.721 to 0.731, while `text-embedding-3-small` improves from 0.792 to 0.832. Retrieval quality improves by 7.7%, with `nv-embed-v2` achieving 0.828. These results underscore Memory Judger's efficiency in enhancing retrieval quality with limited memory units.

Memory Judger and Memory Refinement substantially enhance multi-session dialogue performance. By integrating Memory Judgement and Memory Refinement, MAP outperforms the Hybrid RAG baseline in dialogue state tracking (JGA) and task completion efficiency (DTE), as shown in Figure 5. For instance, MAP achieves a JGA of 0.74 on Qwen2.5-7B, surpassing Hybrid RAG at 0.41. Similarly, on Mistral-7B, MAP reaches 0.64, compared to Hybrid RAG at 0.57 (Figure 5(a)). In terms of DTE, MAP reduces the the required turns for task completion, achieving a DTE of 3.19 on GPT-4o-mini, compared to 4.30 for Hybrid RAG (Figure 5(b)). These findings demonstrate the effectiveness of Memory Judger and Refinement in MAP, enhancing dialogue state tracking and dialogue efficiency across LLMs.

QA memory improves performance in existing task-oriented dialogue datasets. To validate the generalizability of QA memory within MAP, we evaluate it on two standard dialogue state tracking benchmarks: SGD and MultiWOZ2.2. Despite sharing the same DST task, these datasets differ in annotation protocols and domain complexity, leading to distinct sets of published baselines as shown in Table 6. For SGD evaluation, compared with fine-tuned LDST (Feng et al., 2023) and the SGD Baseline (Rastogi et al., 2020b), GOLOMB (Gulyaev et al., 2020), SGP-DST (Ruan et al., 2020), TS-DST (Du et al., 2022), we achieved a state-of-the-art JGA of 0.846 and achieves comparable performance with LDST in the metric of AGA. For MultiWOZ 2.1 evaluation, compared with SGD Baseline, TRADE (Wu et al., 2019a), DS-DST (Zhang et al., 2020), TripPy (Heck et al., 2020), TOATOD (Bang et al., 2023b), SDP-DST (Lee et al., 2021), and LDST, MAP* achieves a JGA of 0.879, substantially outperforming other methods. We attribute this phenomenon to the fewer task domains in MultiWOZ, which enables our QA memory to effectively capture relationships among task attributes, thereby demonstrating that QA memory is a powerful method for improving task-oriented dialogue state tracking.

Dataset	Methods	JGA	AGA
SGD	SGD Baseline	0.254	0.906
	GOLOMB	0.465	0.750
	SGP-DST	0.722	0.913
	TS-DST	0.786	0.956
	LDST	0.845	**0.994**
	MAP*	**0.846**	0.965
MultiWOZ 2.2	SGD Baseline	0.420	-
	TRADE	0.454	-
	DS-DST	0.517	-
	TripPy	0.530	-
	TOATOD	0.638	-
	SDP-DST	0.576	0.985
	LDST	0.607	**0.988**
	MAP*	**0.879**	0.976

Table 6: Results of different methods on SGD and MultiWOZ 2.2. MAP* is a single-session variant of MAP, where the Memory Judger is disabled while retaining the QA memory.

5.4 Case Study

In our case study, we compared four methods for generating confirmation responses: (1) Direct Prompting with the full conversation history, (2) Hybrid RAG retrieving relevant dialogue history,

Session 23 Evaluation -	Intention: The user intends to listen to songs.
Dialogue Context	User: Can you play 'Drive' again? Assistant:
Confirmation Response (Direct Prompting with the Full Conversation History)	Sure! Just to confirm, you'd like me to play the song "Drive" on the **bedroom speaker** again. Should I start it now? *Also, would you like me to check out some tickets for ... (Extra Noise)*
Confirmation Response (Hybrid RAG with Retrieved History)	Sure! Let me play "Drive" for you once more. By the way, have you decided on the number of tickets you'd like to buy for Bill Callahan's concert? It's coming up on March 10th in Los Angeles! *(Wrong Answer)*
Confirmation Response (Hybrid RAG with Summary)	Sure! I've started playing "Drive" again. Is there anything else you'd like? *(Key Slot Omission)*
Confirmation Response (MAP)	Sure! Just to confirm, you'd like me to play the song "Drive" on the **bedroom speaker** again. Should I start it now? *(Right Answer)*

Figure 6: Case study.

(3) Hybrid RAG retrieving a summary of the conversation, and (4) MAP with intention-based QA memory. As shown in Figure 6, Direct Prompting introduces excessive noise, while both Hybrid RAG settings often miss key task slots due to retrieval inaccuracies. In contrast, MAP consistently generates responses with accurate, task-relevant slots and fewer errors. These findings indicate that **leveraging intention-based QA memory within MAP is more effective in preserving crucial task information and minimizing errors in confirmation response generation compared to other baselines**. More details are provided in Appendix D.

6 Related Works

6.1 Task-Oriented Dialogue Dataset

To advance research in TOD modeling, numerous datasets have been developed, categorized primarily by how dialogue utterances are curated: Machine-to-Machine (M2M) (Shah et al., 2018) and Wizard-of-Oz (WOz) (Kelley, 1984). M2M datasets like SGD (Rastogi et al., 2020a) define service schemas with intents, slots, and constraints, while STAR (Mosig et al., 2020) enhances this by outlining ideal dialogue flows and incorporating realistic user behavior. WOz-based datasets like WOZ (Wen et al., 2017) and FRAMES (Asri et al., 2017) have demonstrated the effectiveness of the WOz setup. MultiWOZ (Budzianowski et al., 2018) stands out for its user-friendly interface for annotators and well-defined user goals, resulting in a diverse and semantically complex dataset.

Recent TOD datasets aim to reflect more realistic interactions (Zhang et al., 2022; Hu et al., 2023; Dai et al., 2022). Notable contributions include an employee-oriented dataset by Xu et al. (2024b), featuring expert-validated HR schemas and diverse user profiles, and OB-MultiWOZ by Li et al. (2024b), which enhances TOD sessions with QA-style dialogues supported by external knowledge. These datasets expand the scope of TOD research beyond traditional customer-centric scenarios. Despite the progress in developing diverse TOD datasets, there remains a significant gap in the availability of multi-session TOD datasets.

6.2 Task-Oriented Dialogue Systems

TOD systems have evolved significantly. Initially, they followed a modular pipeline with distinct Natural Language Understanding (NLU), Dialogue State Tracking (DST), policy learning, and generation components (Wu et al., 2019b; Peng et al., 2018). Recently, end-to-end models have emerged, integrating these modules into a single framework trained on annotated dialogues (Wen et al., 2017; Wang et al., 2020). While simplifying structure, this approach still depends on large datasets and retains some modular traits.

With LLMs excelling in NLP, interest in their integration into TOD systems has grown (Raffel et al., 2020; Ouyang et al., 2022). LLMs enhance NLU and DST by extracting user intents and entities (Zhao et al., 2022; Gupta et al., 2022; Madotto et al., 2021, 2020). While Hudeček and Dušek (2023) explored direct LLM use without fine-tuning, performance lags behind supervised models. Conversely, fine-tuning LLMs for TOD tasks shows significant gains (Bang et al., 2023a; Hosseini-Asl et al., 2020). In contrast to existing methods, this work introduces a memory-active policy that integrates long-term memory into multi-session TOD systems, dynamically tracking user intents and preserving critical information to improve multi-turn and long-duration conversations.

7 Conclusion

This study introduces a multi-session TOD task and presents the MT-TOD dataset which features diverse multi-session task goals and structured individual memory banks. Then, MAP, a multi-session TOD framework that integrates memory-guided dialogue planning and a proactive response strategy, is designed for efficient task completion in minimal dialogue turns. Experimental results demonstrate that our MAP significantly reduces the number of dialogue turns, enhances response quality, and improves task success rate, outperforming both direct prompting and other long-term retrieval methods.

Limitation

While our model demonstrates effectiveness on the current dataset, several limitations remain. First, our experiments are limited to locally deployable LLM models, and we have not explored the potential benefits of scaling to larger models, which may yield further improvements. Second, our approach does not incorporate external knowledge bases or internet search functionality, which could enhance contextual understanding and factual accuracy. Lastly, the model's generalizability to broader domains and more complex real-world scenarios remains untested, necessitating further evaluation across diverse datasets and tasks. Future work will address these limitations by expanding model scalability, integrating external knowledge sources, and conducting more comprehensive evaluations.

Ethics Statement

Our research improves multi-session task-oriented dialogue systems through memory-augmented processing while adhering to ethical guidelines. All datasets are publicly available and free of personally identifiable information, with no collection of user-sensitive data or involvement of human subjects. To evaluate model responses, three research assistants with relevant expertise conduct human assessments, each compensated $20 per hour, above the local average for similar roles. While improving AI-driven dialogue, we acknowledge risks such as misinformation and biases, which we address through rigorous evaluation, emphasizing transparency, fairness, and accountability. We advocate for responsible deployment and ongoing bias mitigation research to ensure ethical and equitable AI dialogue systems.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023a. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7355–7369.

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

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040*.

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

Yinpei Dai, Wanwei He, Bowen Li, Yuchuan Wu, Zheng Cao, Zhongqi An, Jian Sun, and Yongbin Li. 2022. Cgodial: A large-scale benchmark for chinese goal-oriented dialog evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4097–4111.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Du, Luyi Cheng, Bo Xu, Zhijun Wang, Sufen Wang, Junyi Yuan, and Changqing Pan. 2022. Ts-dst: A two-stage framework for schema-guided dialogue state tracking with selected dialogue history. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 152–164, Bangkok, Thailand. Association for Computational Linguistics.

Hao Feng, Wei Zhang, Jing Liu, and Maosong Sun. 2024. Ldst: A llama-based dialogue state tracking framework. *arXiv preprint arXiv:2403.12345*.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards llm-driven dialogue state tracking. *Preprint*, arXiv:2310.14970.

Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker. *Preprint*, arXiv:2002.02450.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi, and Yonghui Wu. 2022. Show, don't tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4541–4549.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022. Space-3: Unified dialog model pre-training for task-oriented dialog understanding and generation. *arXiv preprint arXiv:2209.06664*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gašić. 2023a. Chatgpt for zero-shot dialogue state tracking: A solution or an opportunity? *arXiv preprint arXiv:2306.01386*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gašić. 2023b. Chatgpt for zero-shot dialogue state tracking: A solution or an opportunity? *Preprint*, arXiv:2306.01386.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *Preprint*, arXiv:2005.02877.

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

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi 3 woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Heyan Huang, Puhai Yang, Wei Wei, Shumin Shi, and Xian-Ling Mao. 2024. Ostod: One-step task-oriented dialogue with activated state and retelling response. *Knowledge-Based Systems*, 293:111677.

Vojtěch Hudeček and Ondřej Dušek. 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.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P De Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 796–806.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. *Preprint*, arXiv:2109.07506.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Miaoran Li, Baolin Peng, Jianfeng Gao, and Zhu Zhang. 2024b. Opera: Harmonizing task-oriented dialogs and information seeking experience. *ACM Transactions on the Web*, 18(4):1–27.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

Jing Liu, Yu Wang, Wenhao Zhang, Lei Li, Zhiyuan Liu, and Maosong Sun. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024. Toad: Task-oriented automatic dialogs with diverse response styles. *arXiv preprint arXiv:2402.10137*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*.

OpenAI. 2023. Text embedding models. *OpenAI API*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Preprint*, arXiv:1909.05855.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Yu-Ping Ruan, Zhen-Hua Ling, Jia-Chen Gu, and Quan Liu. 2020. Fine-tuning bert for schema-guided zero-shot dialogue state tracking. *Preprint*, arXiv:2002.00181.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024. Lucid: Llm-generated utterances for complex and interesting dialogues. *arXiv preprint arXiv:2403.00462*.

Sandesh Swamy, Narges Tabari, Chacha Chen, and Rashmi Gangadharaiah. 2023a. Contextual dynamic prompting for response generation in task-oriented dialog systems. *arXiv preprint arXiv:2301.13268*.

Sandesh Swamy, Narges Tabari, Chacha Chen, and Rashmi Gangadharaiah. 2023b. Contextual dynamic prompting for response generation in task-oriented dialog systems. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3102–3111, Dubrovnik, Croatia. Association for Computational Linguistics.

Mistral AI Team. 2024a. Mistral 7b: A high-performance language model. *arXiv preprint arXiv:2402.12345*.

OpenAI Team. 2024b. Gpt-4: Openai's advanced language model. *OpenAI Research*.

Qwen Team. 2024c. Qwen-2.5: Advanced large language model with enhanced capabilities. *arXiv preprint arXiv:2409.12345*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Awanish Batra, Simon Randriamihaja, et al. 2024. Llama 3: Open and efficient foundation language models. *arXiv preprint arXiv:2401.00778*.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134.

W Wang, Z Zhang, J Guo, Y Dai, B Chen, and W Luo. 2021. Task-oriented dialogue system as natural language generation. arxiv 2021. *arXiv preprint arXiv:2108.13679*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019a. Transferable multi-domain state generator for task-oriented dialogue systems. *Preprint*, arXiv:1905.08743.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019b. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and He-Yan Huang. 2024a. 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.

Weijie Xu, Zicheng Huang, Wenxiang Hu, Xi Fang, Rajesh Cherukuri, Naumaan Nayyar, Lorenzo Malandri, and Srinivasan Sengamedu. 2024b. Hr-multiwoz: A task oriented dialogue (tod) dataset for hr llm agent. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 59–72.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *Preprint*, arXiv:1910.03544.

Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022. A slot is not built in one utterance: Spoken language dialogs with sub-slots. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 309–321.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *arXiv preprint arXiv:2201.08904*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

A Prompts

A.1 Prompt of dialogue generation

We designed a multi-session dialogue prompt (as shown in Figure 7) that generates multi-session dialogue data based on input dialogue intent, task goal, and target session count. Additionally, during the generation process, we annotate whether each utterance is a confirmation response. These annotations, after manual verification, will be used in the main experiment for confirmation-type response generation.

A.2 Prompt of Task Slot Query Generation

During the evaluation process, we design a prompt (as shown in Figure 8) that generates a query corresponding to the missing task attributes based on the current dialogue context and task objectives. The input to this prompt is the dialogue context history and the generated task objectives. This query is then used as input to the memory judger to assist in selecting QA memory units that align with the task objectives.

Attribute	Train
Domains	16
Intentions	22
Task goals	4,534
Dialogues	13,441
Utterances	89,152
Avg. slots per task goal	4.49
Number of individuals	565
Avg. intentions per individual	6.24
Avg. sessions per individual	23.79
Avg. Utterances per individual	157.80

Table 7: MS-TOD Subset Statistics for Memory Judger Training.

A.3 Prompts of Confirmation Response Generation

In the evaluation process, we employed a confirmation-type response generation approach to assess the integration performance of multi-session memory in task-oriented dialogues. We designed the prompt as shown in Figure 9, which leverages the dialogue context, task objectives, and activated memory units to generate responses.

A.4 Prompts of GPT4 Evaluation

During the evaluation process, we employed a GPT-4 prompt (as shown in Figure 11) to assess the quality of confirmation-type responses. This prompt evaluates the response holistically from four perspectives: requirement alignment, content accuracy, language quality, and comparison to the reference answer. The input to this prompt includes the dialogue history, task objectives, the reference response, and the model-generated response. This design ensures that the evaluation of the response is not solely based on the dataset's reference reply but also takes into account multiple factors such as whether the task objectives are met and the overall quality of the response. Such an evaluation approach is more comprehensive.

A.5 Prompts of Dialogue State Tracking

we used a prompt modified from (Heck et al., 2023b) (as shown in Figure 10) that generates the dialogue state for each user turn in the dialogue. Let

$$A_1 = P \oplus \text{system} : M_1 \oplus \text{user} : U_1$$

$$A_t = A_{t-1} \oplus system : M_t \oplus \text{user} : U_t, \quad \forall t \in [2, T]$$

where P is the task description which provides the model with instructions for how to process a dialogue between a system M and a user U. In contrast to (Heck et al., 2023b), P does not include the detailed description for slots to challenge ChatGPT's ability to understand the meaning of the slots. Apart from that, ChatGPT often generated answers with excessively detailed explanations, deviating from the expected response format. To address this issue, a prompt that includes "No explanation!" as an instruction to ChatGPT not to provide detailed explanations was introduced (Feng et al., 2023) and we added this to our prompt.

B Dataset

B.1 Dataset for Memory Judger

To ensure that the memory judger generalizes across different domains and scenarios, we generated the training dataset(as shown in Table 7) using the same method described in the main text. The dataset spans 16 domains, 4,534 task goals, and 13,411 dialogues, involving a total of 565 individuals, each with an average of 6.24 intentions. Beyond training the memory judger, this dataset can also serve as an alternative evaluation set for broader benchmarking.

B.2 Dataset Structure

MS-TOD encompasses multiple individual task-oriented dialogue datasets, each consisting of several sessions. We present an example of one session (as shown in Figure 12) from an individual. This session includes a $session\_id$, where a larger value indicates a more recent timestamp. The domain represents the specific field or area of the dialogue. The $reference\_dialogue\_id$ corresponds to the $dialogue\_id$ in the original SGD dataset that shares the same task objective. The $exist\_confirmation$ indicates whether the session contains a confirmation-type response and whether it is an evaluation target. The intent represents the specific purpose or goal of the dialogue. The content stores the actual dialogue text. The $task\_goal$ includes task slots and their corresponding attribute values. Each individual contains dozens of session data structured as described above.

B.3 Intent-driven QA Memory

For each historical session, we generated an intent description and the corresponding QA memory (as shown in Figure 13) for the objectives of that intent description. The QA memory consists of multiple QA pairs, where each query is a question about a task attribute under that intent, and the answer is the slot value corresponding to that task attribute.

Activation Module	Recall@3	Recall@5	Recall@10
bm25	0.642	0.721	0.842
t5-base	0.443	0.575	0.773
bert-base	0.463	0.584	0.785
bert-large	0.401	0.530	0.730
nv-embed-v2	0.668	0.769	0.896
bge-large-en-v1.5	0.681	0.761	0.888
text-embed3-small	**0.702**	**0.792**	**0.905**

Table 8: Performance evaluation of activation modules on memory retrieval

C Supplementary Experimental Results

C.1 Memory Activation Comparision

Table 8 compares the performance of different activation modules on memory retrieval. **text-embed3-small** achieves the highest recall across all thresholds, with 0.702 at Recall@3, 0.792 at Recall@5, and 0.905 at Recall@10, demonstrating superior retrieval capability. Among other models, **nv-embed-v2** and **bge-large-en-v1.5** also perform well, while traditional retrieval methods like **BM25** remain competitive at Recall@10 but lag behind embedding-based methods at lower recall levels. **T5-base** and **BERT-based models** exhibit lower recall, suggesting that general pre-trained models are less effective for specialized memory retrieval. These results highlight **text-embed3-small** as the most effective choice for long-term memory activation in multi-session dialogues.

C.2 Human Evaluation Details

Table 11 presents the results of human evaluation, including accuracy, informativeness, and coherency scores. Accuracy is rated on a scale of 0 to 1, while informativeness and coherency are rated from 0 to 3. The average scores in 5 are computed using a weighted sum with weights of 1, 1/3, and 1/3. All evaluations were conducted in a blind review manner to compare the response quality of w/o MAP and w/ MAP. Additionally, the Confirmation-type Response type assesses the response quality after memory-guided dialogue planning, while the multi-turn evaluation focuses on dialogues under the proactive response strategy, continuing until task completion or forced termination.

Model	Setting	Slot Accuracy	BLEU	ROUGE
LLaMA3-8B	w/o MAP	**0.62**	**10.47**	28.59
	w/ MAP	0.56	9.86	**30.39**
Qwen-7B	w/o MAP	0.48	10.33	29.77
	w/ MAP	**0.55**	**10.90**	**31.28**
Mistral-7B	w/o MAP	0.59	**10.09**	**28.42**
	w/ MAP	**0.56**	6.66	24.64
GPT4o-mini	w/o MAP	0.61	**20.30**	**43.49**
	w/ MAP	**0.68**	13.6	35.20

Table 9: Performance comparison of task-oriented dialogue models with and without long-term memory integration: Slot Accuracy, BLEU, and ROUGE metrics.

Model	Slot Accuracy	BLEU	ROUGE
AutoTOD	0.61	3.34	24.07
MAP	**0.68**	**5.47**	**25.03**

Table 10: Performance comparison on Slot Accuracy, BLEU, and ROUGE.

C.3 Additional Evaluation Metrics

Table 9 compares the performance of task-oriented dialogue models with and without memory-augmented processing (MAP) across Slot Accuracy, BLEU, and ROUGE metrics. The results reveal a trade-off between structured slot accuracy and response fluency. In most models, MAP slightly reduces slot accuracy, as seen in LLaMA3-8B, which drops from 0.62 to 0.56, and Mistral-7B, which decreases from 0.59 to 0.56. However, GPT4o-mini benefits from MAP, achieving the highest slot accuracy of 0.68. BLEU scores generally decline, suggesting that MAP shifts responses away from verbatim accuracy towards greater contextual adaptability. Mistral-7B drops from 10.90 to 6.66, and LLaMA3-8B decreases from 10.47 to 9.86. Conversely, ROUGE scores improve with MAP in several cases. LLaMA3-8B increases from 28.59 to 30.39, and Qwen-7B rises from 29.77 to 31.28, indicating enhanced informativeness and coherence. However, Mistral-7B experiences a slight decrease in ROUGE from 28.42 to 24.64. Overall, the results suggest that MAP enhances response informativeness while slightly compromising slot accuracy and BLEU, highlighting a trade-off between structured information retention and more natural, contextually aware responses.

Table 10 presents the performance comparison between AutoTOD and MAP on Slot Accuracy, BLEU, and ROUGE. The results indicate that MAP consistently outperforms AutoTOD across all three metrics, demonstrating its effectiveness in enhancing dialogue quality. Slot Accuracy improves from 0.61 to 0.68, indicating better tracking of task-specific information. BLEU increases from 3.34 to 5.47, reflecting more precise and fluent responses. ROUGE also shows a slight improvement, rising from 24.07 to 25.03, suggesting that MAP enhances informativeness and coherence. These results highlight the advantages of memory-augmented processing, which enables more accurate and contextually relevant dialogue generation.

D Case Study Detail

Figure 14 presents four different configurations of conversation contexts not shown in the main paper. Specifically, (1) Full conversation history includes every session from the dialogue history as prompt input to the reader. (2) Retrieval-based methods retrieve the dialogue sessions most relevant to the current session (Session 23) and append them to the reader's context (3) Retrieving a summary compiles a summary of past sessions (Sessions 1 to 22) for inclusion alongside the current context. Finally, (4) MAP integrates QA memory with the Session 23 context to generate responses. By illustrating these detailed contexts, Figure 14 provides further insights into how each approach manages multi-session dialogue.

Model	Setting	Confirmation-type Response			Multi-Turn		
		Accuracy	**Informativeness**	**Coherency**	**Accuracy**	**Informativeness**	**Coherency**
GPT4o-mini	w/o MAP	0.62	1.83	1.90	0.81	1.92	2.44
	w/ MAP	**0.65**	**2.38**	**2.48**	**0.87**	**1.93**	**2.74**
LLaMA	w/o MAP	0.56	1.47	1.74	0.78	1.64	2.36
	w/ MAP	**0.61**	**1.98**	**2.16**	**0.88**	**2.51**	**2.71**
Qwen	w/o MAP	0.43	1.24	1.85	0.82	1.60	2.02
	w/ MAP	**0.54**	**1.70**	**2.30**	**0.92**	**1.93**	**2.47**
Mistral	w/o MAP	0.58	1.63	1.99	0.89	2.49	2.72
	w/ MAP	**0.61**	**2.06**	**2.08**	**0.93**	**2.74**	**2.85**

Table 11: Comparison of different models on human evaluation metrics: accuracy, informativeness, and coherence. The results are presented for both confirmation-type responses and multi-turn dialogue settings, comparing standard inference (‘w/o MAP‘) with memory-augmented processing (‘w/ MAP‘).

Prompts of the Dataset Generation

User Prompt:
"""
Help me generate an English conversation under the {dialogue_intent} intent, where {task_goal}. The conversation should be between a user and an assistant, and it should be split into {task_goal_length} sessions at different points in time, with continuity and connection between the sessions and each session should not less than 6 turns. Additionally, the final session must include a assistant response containing a complete confirmation-type utterance before the user confirms, and this utterance should be marked with ‘is_confirmation‘ set to ‘True‘. and the user must provide a final confirmation response at the end of the final session. For all other sessions, the conversation should end with an assistant's polite declarative statement.
"""
System Prompt:
""" You are dialogue generator assistant.
The sessions should be clearly separated, and the conversation should be formatted as follows:
Each turn should be a dictionary entry.
The conversation should be in the format of a list of sessions, where each session is a list of dictionaries representing each turn.
Each dictionary entry should have two keys: speaker (either 'user' or 'assistant') and text (the spoken dialogue).
Except for final session, each session should be a seperate dialogue and include a complete dialogue structure, beginning with a greeting from the user and ending with an assistant's polite declarative statement.
Feel free to expand the dialogue with additional relevant details, but avoid redundant expressions or repeating the same phrases.
Reponse me with a json format

```
{
    "sessions": [
        [
            {
                "speaker": "xx",
                "text": "xx"
            },
            {
                "speaker": "xx",
                "text": "xx"
            }
        ]
    ]
}
```
"""

Figure 7: Prompts of the Dataset Generation

Prompts of the Task Slot Querying Generation

"""
Please help me generate questions, based on the provided {conversation history}, that correspond to unanswered attributes in the task goal {task_attributes}.
1. The questions should start with 'What,' 'When,' 'Why,' 'How,' or 'Where.'
2. Ensure that the generated questions are in third person.
fill the following json: { [Question], }
"""

Figure 8: Prompts of the Task Slot Querying Generation

Prompts of Confirmation Response Generation

""" You are an dialogue assistant.
Generate a confirmation response based on the userś utterance. Include any relevant task goals [TASK GOALS] identified in the dialogue or related memory [MEMORY]. If [MEMORY] is unavailable, construct your response accurately and comprehensively using the provided conversation details. Ensure your reply acknowledges the userś request clearly and incorporates relevant information from both the dialogue and the related memory units [MEMORY].
[TASK GOAL]
{task_goal}

[MEMORY]
{memory}
"""

Figure 9: Prompt of Confirmation Response Generation

Prompt of Dialogue State Tracking on MultiWOZ 2.2

```
"""Consider the following list of concepts, called "slots" provided to you as a json list.

"slots": {
"attraction-area",
"attraction-name",
"attraction-type",
"bus-day",
"bus-departure",
"bus-destination",
"bus-leaveat",
"hospital-department",
"hotel-area",
"hotel-bookday",
"hotel-bookpeople",
"hotel-bookstay",
"hotel-internet",
"hotel-name",
"hotel-parking",
"hotel-pricerange",
"hotel-stars",
"hotel-type",
"restaurant-area",
"restaurant-bookday",
"restaurant-bookpeople",
"restaurant-booktime",
"restaurant-food",
"restaurant-name",
"restaurant-pricerange",
"taxi-arriveby",
"taxi-departure",
"taxi-destination",
"taxi-leaveat",
"train-arriveby",
"train-bookpeople",
"train-day",
"train-departure",
"train-destination",
"train-leaveat",
}

Now consider the following dialogue between two parties called the "system" and "user". Can you tell me which of the "slots" were updated by the "user" in its latest response to the "system"? Present the updates in JSON format. If no "slots" were updated, return an empty JSON list. If you encounter "slots" that were requested by the "user" then fill them with "?". If the user informed that he did not care about a "slot", fill it with "dontcare". Return the output in JSON format and no explanation!
{dialogue}
"""
```

Figure 10: Prompt of Dialogue State Tracking on MultiWOZ 2.2

Prompts of GPT4 Evaluation

""" You are a strict and objective evaluator. Your task is to assess the quality of the final predicted response using the provided conversation context, the user's target goal attributes, and a reference answer. Your evaluation should be fair, professional, and reflect an expert judgment of the response's quality.

[Dialogue Context]

$\{\{conversation_{h}istory\}\}$

[Task Goal]

$\{\{task\_goal\}\}$

[reference_answer]

$\{\{reference\_anwser\}\}$

[predict_answer]

$\{\{predict\_answer\}\}$

Evaluation Criteria:

Requirement Alignment: Does the final predict_answer meet the user's task goal?

Content Accuracy: Is the information in the final response correct, clear, and logically organized?

Language Quality: Is the language fluent, coherent, and readable? Are there any obvious grammatical or word choice errors?

Comparison to Reference Answer: Compared to the reference answer, how does the final response differ in terms of completeness, professionalism, and clarity?

Overall Score: Assign a score from 1 to 10 (10 being the best), considering all of the above factors.

The evaluation must be structured in the following JSON format:

‘‘‘json
{
"Rationale": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
}
‘‘‘ """

Figure 11: Prompts of GPT4 Evaluation

MS-TOD dialogue session structure


```
{
  "session_id": 9,
  "domain": "Travel",
  "reference_dialogue_id": "66_00101",
  "exist_confirmation": true,
  "intent": "ReserveHotel",
  "content": [
    {
      "speaker": "user",
      "utterance": "Hi again, I'm ready to finalize the booking for Aloft
         Portland Airport At Cascade Station.",
      "is_confirmation": false
    },
    {
      "speaker": "assistant",
      "utterance": "Just to confirm, you are booking 1 room at Aloft Portland
         Airport At Cascade Station, Portland, from March 5th to March 7th.
         The room is a standard king room with free Wi-Fi and a 24-hour
         cancellation policy. Is that correct?",
      "is_confirmation": true
    },
    {
      "speaker": "user",
      "utterance": "Yes, that is correct.",
      "is_confirmation": false
    },
    {
      "speaker": "assistant",
      "utterance": "Excellent! Your room has been successfully booked. You
         will receive a confirmation email shortly.",
      "is_confirmation": false
    },
    {
      "speaker": "user",
      "utterance": "Thank you so much for your help!",
      "is_confirmation": false
    },
    {
      "speaker": "assistant",
      "utterance": "You're welcome! Have a great stay in Portland.",
      "is_confirmation": false
    }
  ],
  "task_goal": {
    "hotel_name": "Aloft Portland Airport At Cascade Station",
    "location": "Portland",
    "check_in_date": "March 5th",
    "check_out_date": "March 7th",
    "number_of_rooms": 1
  }
}
```


Figure 12: MS-TOD Session Structure.

MS-TOD Intent Description and QA Memory


```
{
  "9": {
    "intent_description": "The user's intent is to finalize and confirm a
        hotel booking for a specific room at Aloft Portland Airport At Cascade
        Station, including details about the stay dates and room type.",
    "qa_summary": [
      {
        "Question": "What type of room did the user book?",
        "Answer": "The user booked a standard king room."
      },
      {
        "Question": "When is the user's reservation?",
        "Answer": "The user's reservation is from March 5th to March 7th."
      },
      {
        "Question": "Where is the user's reservation located?",
        "Answer": "The user's reservation is located at Aloft Portland Airport
             At Cascade Station."
      },
      {
        "Question": "What amenities are included in the user's reservation?",
        "Answer": "The user's reservation includes free Wi-Fi."
      },
      {
        "Question": "What is the cancellation policy for the user's booking?",
        "Answer": "The cancellation policy for the user's booking is 24 hours
            ."
      }
    ]
  }
}
```


Figure 13: Intent description and QA Memory in MT-TOD.

Session ID	Dialogue History	Summary	Intention-driven QA Memory
Session 1	User:Hi there! I'd like to book an appointment at Blow Dry Bar. Assistant : Sure! When would you like to schedule your appointment? …	The user wants to book an appointment at Blow Dry Bar on March 3rd at 3:15 pm. The assistant confirms the date and is checking the availability for the requested time.	**Intention Description**: The user intends to book a specific appointment. "Question": "What type of appointment is the user looking to book?", "Answer": "The user is looking to book an appointment at Blow Dry Bar." …
Session 3	User:Hey, can you play the song 'Drive' on the **bedroom speaker**? Assistant:Sure, I can play 'Drive' for you. Do you want me to start the song now? …	The user requested to play the song \"Drive\" on the **bedroom**. The assistant confirmed and started the song, asking if the user needed anything else. The user declined further assistance and expressed gratitude.	**Intention Description:** The user intends to listen to the song. "Question": "Where was the song 'Drive' played?", "Answer": "The song 'Drive' was played on the bedroom speaker."
Session 12	User:Hi there! I'm interested in buying a ticket for the Bill Callahan concert? Assistant:Great choice! The Bill Callahan concert is happening on March 10th in Los Angeles. How many tickets would you like to purchase? …	The user expresses interest in buying a ticket for the Bill Callahan concert on March 10th in Los Angeles. They request one ticket, and the assistant confirms availability and asks if the user would like to proceed with the purchase.	**Intention Description**: The user's intent is to purchase a single ticket "Question": "Where will the concert be held?", "Answer": "The concert will be held in Los Angeles."
Session 16	User:Hi, I need to book a ride. Assistant:Sure, I can help with that. Where would you like to go? …	The user wants to book a ride to 1580 Clayton Road #1 for three people and has requested a shared ride.	**Intention Description**: The user's intent is to arrange a shared transportation service.
Session 23 Evaluation	User: Can you play 'Drive' again? Assistant:		**Intention Guess**: The user intends to listen to songs.
Confirmation Response (Direct Prompting with the Full Conversation History)	Sure! Just to confirm, you'd like me to play the song "Drive" on the **bedroom speaker** again. Should I start it now? *Also, would you like me to check out some tickets for the Bill Callahan concert on March 10th in Los Angeles, or are you planning to book a ride somewhere else? Let me know! (Extra Noise)*		
Confirmation Response (Hybrid RAG with Retrieved History)	Sure! Let me play "Drive" for you once more. By the way, have you decided on the number of tickets you'd like to buy for Bill Callahan's concert? It's coming up on March 10th in Los Angeles! *(Wrong Answer)*		
Confirmation Response (Hybrid RAG with Summary)	Sure! I've started playing "Drive" again. Is there anything else you'd like? *(Wrong Answer)*		
Confirmation Response (MAP)	Sure! Just to confirm, you'd like me to play the song "Drive" on the **bedroom speaker** again. Should I start it now? *(Right Answer)*		

Figure 14: Comparison of confirmation response generation across four approaches: (1) Direct Prompting with the full conversation history, (2) Hybrid RAG retrieving relevant dialogue history, (3) Hybrid RAG retrieving a summary of the conversation, and (4) MAP with intention-based QA memory.