MemGuide: Intent-Driven Memory Selection for Goal-Oriented Multi-Session LLM Agents

Yiming Du[1,2*], Bingbing Wang[3*], Yang He[4], Bin Liang[1,2†], Baojun Wang[5], Zhongyang Li[6], Lin Gui[7], Jeff Z. Pan[8], Ruifeng Xu[3†], Kam-Fai Wong[1,2†]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] MoE Key Laboratory of High Confidence Software Technologies, China
[3]Harbin Institute of Technology, Shenzhen, China
[4]The Hong Kong University of Science and Technology, Hong Kong, China
[5]Huawei Noah's Ark Lab, Shenzhen, China
[6]Microsoft AI,
[7]King's College London, UK
[8]The University of Edinburgh, UK
{ydu, kfwong}@se.cuhk.edu.hk, bin.liang@cuhk.edu.hk, yhecb@connect.ust.hk, bruli@microsoft.com,
bingbing.wang@stu.hit.edu.cn, xuruifeng@hit.edu.cn,
http://knowledge-representation.org/j.z.pan/

Abstract

Modern task-oriented dialogue (TOD) systems increasingly rely on large language model (LLM) agents, leveraging Retrieval-Augmented Generation (RAG) and long-context capabilities for long-term memory utilization. However, these methods prioritise semantic similarity over task intent, degrading multi-session coherence. We propose **MemGuide**, a two-stage intent-driven memory selection framework: (1) **Intent-Aligned Retrieval** retrieves goal-consistent QA-formatted memory units; (2) **Missing-Slot Guided Filtering** reranks units by slot-completion gain via a chain-of-thought reasoner and fine-tuned LLaMA-8B filter. We also introduce the **MS-TOD**, the first multi-session TOD benchmark with 132 diverse personas, 956 task goals, and annotated intent-aligned memory targets. Evaluations on MS-TOD show that MemGuide boosts task success rate by 11% (88%→99%) and reduces dialogue length by 2.84 turns, and matches single-session performance.

Code and Dataset: —
https://github.com/Elvin-Yiming-Du/MS_TOD_Memory

Introduction

Modern task-oriented dialogue (TOD) systems increasingly integrate large language models (LLMs) to enhance generalization, context understanding, and response generation (Nguyen et al. 2025; Xu et al. 2024a,b; Chung et al. 2023; Hudeček and Dusek 2023). To utilize historical dialogue states across turns, two dominant strategies have emerged: retrieval-augmented generation (RAG), which supplements the LLM with relevant task descriptions or prior dialogue states (Xu et al. 2024a,b), and the use of long-context models, which encode the entire history directly as input (Nguyen et al. 2025). However, they are both limited to surface-level semantic similarity, often ignoring the task-specific intent and slot-level continuity crucial for coherent multi-session TOD.

While memory-augmented methods (Lu et al. 2023) have emerged, they are often evaluated on single-session benchmarks that lack support for long-term goal tracking and multi-session memory supervision. Unlike open-domain set-

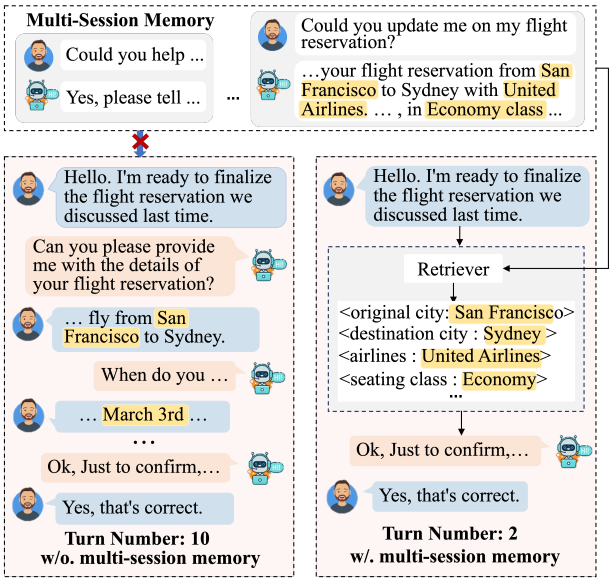

Figure 1: Task-oriented dialogue, without (*left*) vs. with (*right*) multi-session memory; the former demands more turns of conversation.

*These authors contributed equally.
†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tings that focus on free-form recall (Zhong et al. 2024), task-oriented dialogue demands structured slot tracking, evolving intent management, and consistent state maintenance. Crucially, real-world users frequently interact with assistants across multiple sessions to accomplish complex goals, yet most existing TOD models and datasets (Budzianowski et al. 2018; Rastogi et al. 2020; Stacey et al. 2024; Liu et al. 2024) are confined to single-session settings, highlighting a fundamental gap between current TOD systems and the demands of persistent, goal-driven dialogue settings. As shown in Figure 1, traditional single-session TOD systems require users to restate details (e.g., flight times, seat preferences) in every session, leading to inefficiency and frustration.

To address this, we propose **MemGuide**, a novel framework that utilizes long-term memory in the multi-session TOD task. Unlike prior methods that rely solely on semantic similarity for memory selection, MemGuide incorporates task intent and slot-level guidance to enhance memory relevance and response quality. It consists of two core phases: (1) **Intent-Aligned Retrieval**, where an LLM generates an intent hypothesis based on the current dialogue state and retrieves memory units aligned with the predicted goal, ensuring retrieved history is both semantically similar and goal-consistent to support cross-session task tracking robustly. (2) **Missing-Slot Guided Filtering**, which first employs a chain-of-thought (CoT) slot reasoner to detect missing slot values, followed by a filtering module that removes irrelevant or redundant Question Answering (QA) memory units to distill slot-level content for response generation.

These two phases enable MemGuide to convert long-term user history into actionable context, supporting minimal-turn, task-consistent dialogue generation across sessions.

To enable systematic evaluation, we construct the **Multi-Session Task-oriented Dialogue (MS-TOD)**, a new benchmark comprising 132 simulated speakers, each engaging in over 20 sessions covering diverse task goals derived from Schema-Guided Dialogue (SGD) (Rastogi et al. 2020). MS-TOD supports evaluation of slot continuity and long-term memory retrieval across sessions. Unlike open-domain benchmarks focused on retrieving dialogue summaries (Zhong et al. 2024; Li et al. 2024a; Du et al. 2024), multi-session TOD introduces additional challenges. Systems must recall key slot-value pairs, track evolving user intents, and proactively resolve missing or outdated information, while minimizing redundant interactions. To support automatic evaluation, we introduce a **proactive response generation module** to simulate user engagement and evaluate system performance in resolving missing information. Experimental results demonstrate that MemGuide significantly improves dialogue coherence, response quality, task success rate, and overall efficiency in multi-session TOD. The main contributions include:

- We propose **MemGuide**, a two-stage framework that distills and leverages cross-session memory for efficient, minimal-turn task completion.
- We introduce **MS-TOD**, the first multi-session TOD dataset and benchmark task for evaluating long-term memory integration across sessions.

Settings	GPT-4 Score	Slot Acc.
No Retrieval (Direct Prompting)		
Current Session Context	2.60	0.13
Full Context	4.76	0.61
Retrieval-based Methods		
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.

- We demonstrate that MemGuide consistently outperforms strong baselines across multiple metrics, demonstrating the effectiveness of intent-aware memory retrieval and slot-guided filtering.

Related Work

Task-Oriented Dialogue Dataset

TOD datasets are typically constructed via either Machine-to-Machine (M2M) (Shah et al. 2018; Rastogi et al. 2020) or Wizard-of-Oz (WOz) setups (Wen et al. 2017; Budzianowski et al. 2018). M2M datasets (e.g., SGD, STAR) provide schema-driven task flows, while WOz-based datasets (e.g., MultiWOZ, FRAMES) offer more natural but annotation-heavy dialogues. Recent efforts aim to improve realism and domain diversity (Hu et al. 2023; Dai et al. 2022; Xu et al. 2024b; Li et al. 2024b), yet existing benchmarks primarily assume single-session tasks. There remains a notable gap in datasets designed for multi-session TOD, where tracking long-range goals and user intents is essential.

Task-Oriented Dialogue Systems

Traditional TOD systems adopt modular pipelines for NLU, DST, and response generation (Wu et al. 2019a; Peng et al. 2018), later unified into end-to-end models trained on annotated dialogues (Wen et al. 2017; Wang et al. 2020). With the rise of LLMs, recent work explores their use in zero-shot and fine-tuned TOD (Madotto et al. 2021; Bang, Lee, and Koo 2023), often achieving strong results on intent recognition and slot filling. In parallel, long-term memory (LTM) methods such as ChitChat (Li et al. 2024c), MemoryBank (Zhong et al. 2024), and LoCoMo (Maharana et al. 2024) support extended context retention through summarization or heuristic filtering, but lack structured memory aligned with task goals. Most assume single-session dialogues and overlook challenges in maintaining multi-session goal continuity. This work addresses these gaps by introducing MemGuide for long-range, goal-aware tracking.

Preliminary Experiments

To motivate our framework, we first examine the limitations of direct prompting in multi-session TOD and explore the potential of retrieval-based strategies. Since existing TOD

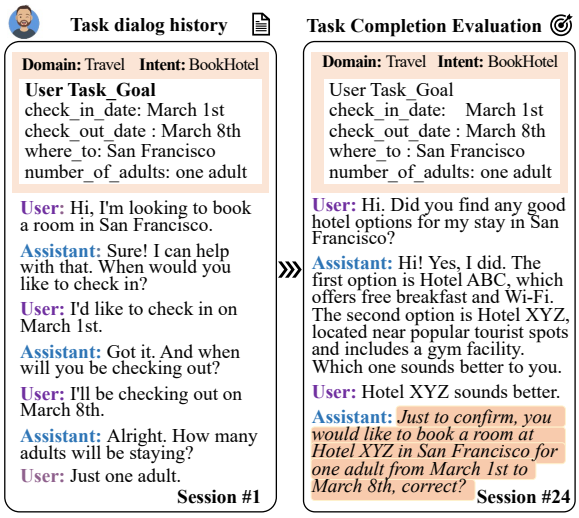

Figure 2: An Example of MS-TOD dataset.

datasets lack long-term dependencies, we construct an evaluation set focused on *confirmation-type* response generation. This evaluation set adopts the same input–output formulation, evaluation criteria, and sample set as the MS-TOD benchmark to ensure consistency and comparability.

To clarify the evaluation setting, we formulate a confirmation-style response generation task. Formally, given the current user utterance u_t, the dialogue context $c = \{u_1, r_1, \ldots, u_{t-1}, r_{t-1}\}$ from the current session, and history dialogue sessions H in the corresponding memory bank (retrieved or concatenated), the model generates a confirmation-style response r_t. The task goal g is held out for reference during evaluation. This setting mimics a common scenario in TOD where the system determines whether the dialogue contains sufficient information to proceed with task execution. For all settings, we use GPT-4o-mini as the unified generator, allowing a fair comparison across input strategies. We compare two strategies:

(1) **Retrieval-based methods**, including sparse (BM25 (Robertson and Zaragoza 2009)), dense (text-embedding-small-3[1]), and a hybrid retrieval. Each selects top-k history sessions H relevant to the current utterance u_t with the model input $x = [H; c; u_t]$. (2) **Direct prompting**, where the full dialogue history is concatenated with the current user utterance u_t without retrieval. The input is $x = [H; C; u_t]$. The model generates a confirmation response $r_t = \text{LLM}(x)$.

As Table 1 shows, retrieval-based methods consistently outperform direct prompting. For instance, *dense retrieval*-based method achieves 0.67 slot accuracy and 7.01 GPT-4 score, surpassing full-context prompting (0.61 and 4.76). This significant gap highlights how direct prompting struggles with context limitations and the "lost-in-the-middle" (Liu et al. 2023), where irrelevant history overwhelms key information, motivating our development of MemGuide for advanced intent-aligned retrieval.

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

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

Table 2: MS-TOD dataset statistics for evaluation.

Dataset

MS-TOD is a multi-session benchmark for evaluating long-term memory through user-specific memory banks, with over 20 sessions from a single user. This supports assessment of memory retrieval, slot tracking, and intent continuity. For evaluation, we include held-out sessions with manually annotated confirmation-type responses (Figure 2). MS-TOD is built in four stages: (1) multi-session dialogue generation; (2) confirmation-type response annotation; (3) QA-style memory bank construction; and (4) human validation.

Multi-Session Dialogue Generation

We begin by generating multi-session dialogues for each task goal sampled from the SGD dataset (Rastogi et al. 2020). For every task, we synthesize *three* temporally ordered sessions using GPT-4, each conditioned on the slot-filling status of the previous session. This simulates how users revisit and revise the same task across time. Specifically, Session 1 presents an incomplete task with missing slots; Session 2 introduces updates; and Session 3 concludes with final confirmation. This staged construction reflects real-world dialogue dynamics while avoiding overfitting, as most SGD tasks involve fewer than ten slots.

Confirmation-Type Response Annotation

To evaluate long-term task fulfillment in dialogue systems, we annotate the final session of each task with *confirmation-type responses*. Each marks the utterance confirming task completion and associated slot-value goal with manually labels (confirmation/non-confirmation). These annotations serve two purposes: (1) **Supervising Memory Selection**, indicating when to trigger memory retrieval; and (2) **Supporting Evaluation**, evaluating if the system recalls goal-relevant content and generates accurate confirmations.

Memory Bank Construction

Since multi-session interactions are organized around individuals, we group dialogues into *individual memory banks* (Figure 2), where each bank stores temporally ordered sessions. During construction, we ensure that intents within the same bank are distinct and non-conflicting, enabling consistent memory usage and avoiding cross-intent interference.

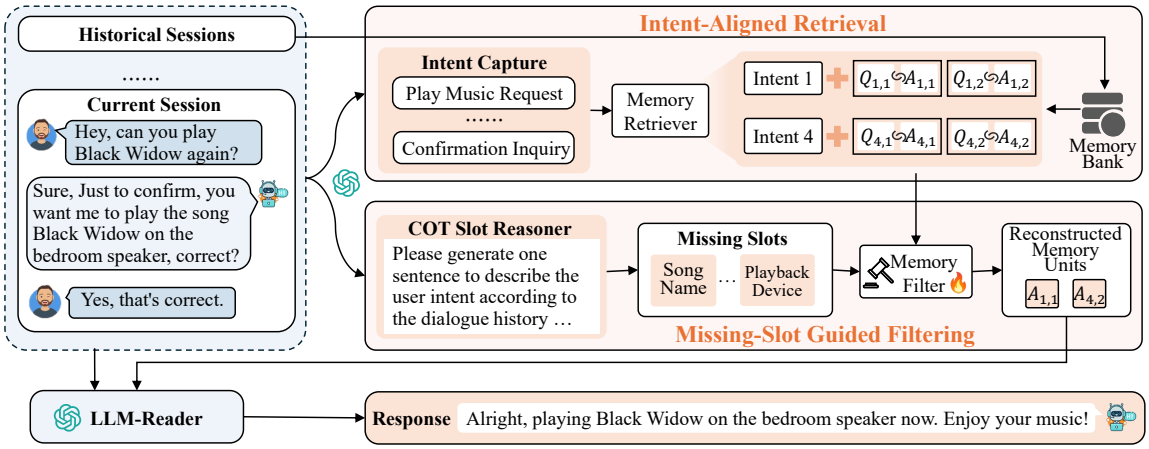

Figure 3: Overflow of our MemGuide framework, which comprises Intent-Aligned Retrieval and Missing-Slot Guided Filtering.

Each bank contains over 20 sessions spanning at least six task intents (Table 2), with one held-out evaluation session per intent for confirmation-type assessment. For each completed session, we use GPT-4 to generate an intent description along with a set of QA pairs, each capturing a slot-specific fact for retrieval. This QA-style format is motivated by prior work (Chen et al. 2023) showing that question–answer structures facilitate more accurate and efficient retrieval compared to unstructured text. To maintain causal consistency, memory access is limited to sessions prior to the current evaluation point.

Human Validation

To ensure coherence, correctness, and usability, we apply a structured multi-stage validation process involving three annotators experienced in NLP. The process includes: (1) Verifying each session preserves the intended goal and correct slot values; (2) Removing dialogues with excessive redundancy across sessions; (3) Verifying that confirmation-type utterances match expected slot-value goals; (4) Removing sessions that fail to complete task goals or lack confirmation turns; and (5) Excluding episodes with unnatural repetition of similar intents. We additionally conduct inter-annotator agreement evaluation to ensure labeling consistency.

MemGuide

MemGuide is a two-stage framework for multi-session TOD, as shown in Figure 3. It first performs ***Intent-Aligned Retrieval***, extracting the current user intent via an LLM and retrieving QA-formatted memory units from the memory bank that align with this intent using semantic similarity. Then, ***Missing-Slot Guided Filtering*** identifies unfilled task slots through a CoT reasoner and re-ranks retrieved memories with a fine-tuned LLaMA-8B filter based on their ability to fill these slots. Finally, ***Response Generation*** leverages the filtered memories to produce proactive responses, reducing conversational turns and boosting task success.

Given a dialogue context $c = \{u_1, r_1, \ldots, u_{t-1}, r_{t-1}, u_t\}$ and a user-specific long-term memory bank M, the goal is to generate the next system response r_t. The memory bank M is a collection of memories from past sessions, structured as $M = \{(k_i, V_i)\}_{i=1}^{N}$ and t is the number of turns. Each entry consists of a high-level intent description k_i (e.g., "book a flight to San Francisco") represents the high-level intent and a corresponding set of QA-formatted memory units $V_i = \{(q_{i,j}, a_{i,j})\}_{j=1}^{n}$, where $q_{i,j}$ is the j-th question about a task slot within the i-th intent (e.g., "What is the departure date?") and $a_{i,j}$ is the corresponding answer (e.g., "July 28, 2025"). The optimal response r should coherently continue the conversation while strategically utilizing information from M to progress towards task completion. It is generated by an LLM conditioned on the context and a carefully selected memory subset $M_{sel} \subseteq M$.

Intent-Aligned Retrieval

The stage aims to broadly identify past sessions relevant to the current user goal by retrieving memory entries from the long-term bank that share a consistent high-level intent, ensuring thematic alignment.

Current Intent Extraction. Given the dialogue context c, we use GPT-4o-mini to generate a high-level intent description d_{int}, which summarizes the user's objective in the current session. The model is prompted to generate a short, command-like phrase summarizing the user's goal (e.g., "find a local Italian restaurant"). This distilled phrase serves as a current intent key, k_{cur}, providing a canonical task representation that is standardized for retrieval.

Semantic Retrieval. We then retrieve memory units from the bank M, which contain stored intent keys $\{k_i\}_{i=1}^{N}$. and corresponding QA pairs that are semantically closest to the extracted current intent k_{cur}. We use an embedding model (e.g., text-embedding-3-small) to compute dense vector representations for all memory units. The relevance score between k_{cur} and each memory unit is calculated using cosine similarity. The top-K memory units (k_i, V_i) with the highest scores are selected to form the candidate memory set, M_{cand}. This procedure ensures that we only consider memories from sessions with a shared high-level objective, guaranteeing thematic alignment for subsequent stages.

Missing-Slot Guided Filtering

While the retrieved memories in M_{cand} are thematically aligned with the user's intent, their immediate utility for the current dialogue turn can vary significantly. To address this, this stage performs a fine-grained filtering, prioritizing memory units that are most likely to resolve the immediate information needs of the dialogue.

Information Gap Identification. To guide the filtering process, we first must precisely identify what information is required to advance the current task. We leverage an LLM configured as a CoT reasoner (Wei et al. 2022) to analyze the dialogue context c and the overall goal corresponding with intent key k_{cur}, so as to enumerate a list of essential task slots that have not yet been filled or confirmed. The CoT prompt is structured to guide the LLM through a logical sequence. First, it enumerates all required slots for the intent, then checks the dialogue history to determine which of these slots have already been filled or confirmed, and finally outputs only those slots that remain unresolved. The result is a list of hypothesized missing slots, $L_{miss} = \{slot_1, slot_2, \dots\}$. For example, in a flight booking task where the user has only specified a destination, L_{miss} might be identified as $\{\text{departure date}, \text{return date}, \text{seat preference}\}$.

Re-ranking by Marginal Slot-Completion Gain. Having identified the information gaps represented by L_{miss}, we then re-rank each QA pair $(q_{i,j}, a_{i,j})$ within the candidate set M_{cand} based on its potential to fill one of these gaps. This process is driven by the principle of selecting information with the highest marginal slot-completion gain, the expected improvement in downstream task performance.

To operationalize this, we fine-tune a smaller, efficient LLaMA-8B model (Meta AI 2024) to act as a specialized filter. This filter estimates the probability that a given QA pair provides an answer for one of the missing slots. For each QA pair $(q_{i,j}, a_{i,j})$ from M_{cand}, the model computes:

$$s_{i,j} = P(y = 1 \mid c, L_{\text{miss}}, q_{i,j}, a_{i,j}) \tag{1}$$

where $y = 1$ signifies that the answer $a_{i,j}$ successfully fills a slot presented in L_{miss}. To supervise this filter, we construct a training dataset using the same pipeline as MS-TOD memory bank generation. Specifically, for each held-out session, we simulate missing-slot contexts and label each QA pair from prior sessions as positive (if it fills a missing slot) or negative (otherwise). Detailed training configurations and dataset are provided in Appendix. The model is optimized using a standard binary cross-entropy loss function:

$$\mathcal{L} = -\sum_{i,j} [y_{i,j} \log s_{i,j} + (1 - y_{i,j}) \log(1 - s_{i,j})] \tag{2}$$

To balance the initial semantic relevance from the initial semantic retrieval (denoted as $s^{\text{pre}}_{i,j}$) with the slot-filling utility score $s_{i,j}$ from our filter, we compute a final score:

$$s_{\text{final},ij} = \alpha \cdot s^{\text{pre}}_{i,j} + (1 - \alpha) \cdot s_{i,j} \tag{3}$$

where α is a hyperparameter. We select the top-K (e,g,. K=5) QA pairs with the highest $s_{\text{final},ij}$ scores.

Response Generation

The final response is generated using the top-K memorys $A_{\text{core}} = \{a_1, a_2, \dots, a_K\}$, omitting auxiliary questions $q_{i,j}$. An LLM reader receives the dialogue context c, the core facts A_{core}, and missing slots L_{miss} as prompt inputs.

$$r = \text{LLMReader}(\text{prompt}(c, A_{core}, L_{\text{miss}})) \tag{4}$$

The prompt instructs the model to: (1) continue the conversation naturally, and (2) proactively address L_{miss} using A_{core}, e.g., by confirming or suggesting stored values. For example, if memory provided a preferred airline, the system might respond, "I see you're flying to Montreal. Last time you flew with Air Canada. Would you like to book with them again?" This proactive strategy, directly informed by our two-stage memory selection, minimizes redundant questions and accelerates task completion. All prompt templates used in our method are provided in Appendix.

Experiments

To enable fine-grained assessment of long-term memory utilization and task completion, we evaluate MemGuide on sessions from the MS-TOD benchmark that are annotated with *confirmation-type responses*, in which the final utterance explicitly confirms that the user goal has been achieved. Each session is associated with a gold-standard task goal and corresponding slot-value set, enabling precise evaluation using standard metrics of task success and response quality.

Experimental Setups

Evaluation Metrics. We use four core automatic metrics and human evaluation to evaluate performance: 1) **GPT-4 score**, (1–10) evaluates response quality in terms of fluency, coherence, and informativeness (details in Appendix); 2) **Joint Goal Accuracy** (JGA) measures slot prediction accuracy; 3) **Dialogue Turn Efficiency** (DTE) captures the number of turns required to complete a task, and 4) **Success Rate** (S.R.) indicates whether the user goal is achieved. 5) **Human evaluation** further assesses Accuracy, Informativeness, and Coherence, with **A.I.C**. denoting their average.

Baselines. We evaluate MemGuide against three representative categories of baselines: 1) **General-purpose LLMs.** We assess full-context prompt (FCP)-based dialogue performance using instruction-tuned models including LLaMA3-8B (Touvron et al. 2024), Qwen2.5-7B (Team 2024c), Mistral-7B (Team 2024a), and GPT-4o-mini (Team 2024b). 2)**Traditional Task-Oriented Dialogue Systems.** To evaluate MemGuide under structured DST conditions, we include task-specific baselines such as BERT-DST (Chao and Lane 2019), LDST (Feng et al. 2024), and AutoTOD (Xu et al. 2024a), where the latter incorporates an external memory module for cross-turn goal tracking. While these models were not originally designed for multi-session scenarios, they represent the strongest available TOD pipelines when adapted to this setting. 3) **Long-term Summarization.** We implement a summarization-based baseline inspired by ChatCite (Li et al. 2024c), which condenses session histories into concise summaries used during inference.

Model	Setting	GPT-4	JGA	DTE	S.R.	A.	I.	C.	A.I.C.
LLaMA3-8B	FCP	4.89	**0.64**	5.37	0.82	0.56	1.47	1.74	1.26
	MemGuide	**6.39**	0.63	**3.46**	**0.92**	**0.61**	**1.98**	**2.16**	**1.58**
Qwen-7B	FCP	6.26	0.66	4.93	0.83	0.43	1.24	1.85	1.17
	MemGuide	**6.81**	**0.66**	**4.31**	**0.87**	**0.54**	**1.70**	**2.30**	**1.51**
Mistral-7B	FCP	6.20	0.73	2.52	1.00	0.58	1.63	1.99	1.40
	MemGuide	**6.48**	**0.80**	**1.21**	**1.00**	**0.61**	**2.06**	**2.08**	**1.58**
GPT-4o-mini	FCP	6.93	0.67	6.03	0.88	0.62	1.83	1.90	1.78
	MemGuide	**7.14**	**0.70**	**3.19**	**0.99**	**0.65**	**2.38**	**2.48**	**2.17**

Table 3: Combined results comparing FCP and MemGuide across LLMs. GPT-4, JGA, DTE, and S.R. are automatic metrics; A., I., C., and A.I.C. are human evaluation metrics for accuracy, informativeness, coherence, and their average score.

Model	GPT-4	JGA	DTE	S.R.
Bert-DST*	-	0.07	-	-
LDST*	-	0.23	-	-
AutoTOD†	6.49	0.44	7.80	0.81
ChatCite	6.59	0.660	4.71	0.84
MemGuide	**7.14**	**0.70**	**3.19**	**0.99**

Table 4: Results of traditional TOD models, summary-based methods, and MemGuide. Models marked with * focus on DST only. † indicates a simplified AutoTOD pipeline.

Main Results

Comparision with General-purpose LLMs. Compared to FCP, MemGuide leverages the same underlying LLM as a memory reader to retrieve and utilize relevant long-term memory, yielding substantial improvements across critical metrics. As illustrated in Table 3, our approach consistently outperforms FCP across all tested models, demonstrating robust gains in task accuracy, response quality, and interaction efficiency. For example, when using Mistral-7B as the LLM Reader, MemGuide increases JGA from 0.73 to 0.80 and reduces DTE from 2.52 to 1.21, representing a 52% reduction. LLaMA3-8B achieves the largest improvement in GPT-4 score (from 4.89 to 6.39), while GPT-4o-mini reduces dialogue turns from 6.03 to 3.19, corresponding to a 47.1% decrease.. Similar gains are observed with Qwen-7B and other models. These consistent gains across models confirm that integrating memory-guided reasoning with the same base model enhances not only task accuracy but also response relevance and interaction fluency, validating the effectiveness of intent-aligned memory selection in multi-session task-oriented dialogue.

Comparison with Traditional TOD and Summarization Baselines. As existing models are not explicitly designed for multi-session TOD, we compare MemGuide with two representative categories: (1) traditional Dialogue State Tracking (DST)-focused models (BERT-DST, LDST, AutoTOD), and (2) a summarization-based approach inspired by ChatCite. As shown in Table 4, MemGuide consistently outperforms all baselines across key metrics. Compared to AutoTOD, it increases JGA from 0.440 to 0.698 and reduces DTE from 7.80 to 3.19. Against the summarization baseline, MemGuide achieves clear improvements across all metrics: the GPT-4 score increases from 6.59 to 7.14, JGA improves from 0.66 to 0.70, DTE decreases from 4.71 to 3.19, and the success rate rises from 0.84 to 0.99. These results demonstrate the effectiveness of MemGuide in improving both task performance and dialogue efficiency.

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**
	MemGuide*	**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**
	MemGuide*	**0.879**	0.976

Table 5: Results of different methods on SGD and MultiWOZ 2.2. MemGuide* is a single-session variant of MemGuide, where the missing slot guided filtering is disabled while retaining the QA memory.

Setting	w/ Raw History	w/ Intent-QA Memory
LLaMA3-8B	5.09	**6.34**
Qwen-7B	6.38	**6.56**
Mistral-7B	5.86	**6.71**
GPT-4o-mini	7.01	**7.14**

Table 6: Comparison of GPT-4 scores using retrieved history vs. intent-QA memory across different LLM settings.

Human Evaluation. We conduct human evaluation to assess the effectiveness of MemGuide in generating confirmation-type responses after memory-guided dialogue planning. Human annotators rate each response along three dimensions: **Accuracy** (binary), **Informativeness** (scored from 0 to 3), and **Coherence** (scored from 0 to 3). The average of these scores, denoted as **A.I.C.**, provides an overall measure of perceived response quality. As shown in Table 4, MemGuide consistently improves human-judged quality across all metrics. All evaluations are conducted under a blind review protocol.

Generalization to Single-Session DST Tasks. To assess the generalization of MemGuide in single-session settings, we focus on dialogue state tracking (DST), a core task that supports downstream components like policy planning and response generation. DST is a widely used and well-defined task that depends on context understanding and supports downstream dialogue components. We evaluate it on two widely-used single-session DST benchmarks, SGD and MultiWOZ2.2. While both benchmarks focus on DST, differences in annotation schemes and domain coverage result in distinct baseline configurations across SGD and MultiWOZ2.2 (Table 5). On SGD, MemGuide achieves a state-of-the-art JGA of 0.846, surpassing strong baselines such as LDST (Feng et al. 2023), GOLOMB (Gulyaev et al. 2020), SGP-DST (Ruan et al. 2020), and TS-DST (Du et al. 2022), and performs comparably to LDST on Average Goal Accuracy (AGA) (Rastogi et al. 2020). On MultiWOZ2.2, MemGuide* attains a JGA of 0.879, significantly outperforming prior models including TRADE (Wu et al. 2019b), TripPy (Heck et al. 2020), and SDP-DST (Lee, Cheng, and Ostendorf 2021). We attribute the superior performance to QA memory's ability to capture slot dependencies more effectively in smaller domain settings, confirming its adaptability and robustness across datasets.

Ablation Study

Effect of intent-aligned Retrieval. Table 6 shows that our intent-aligned retrieval with structured QA memory consistently improves response quality, outperforming unstructured baselines by up to 1.29 points (e.g., LLaMa3-8B: 5.05 $\rightarrow$ 6.34). We attribute this enhancement to the alignment between the retrieved structured context and the model's generative reasoning process. By providing intent-anchored QA pairs, the system can reason over task-relevant content with greater precision and mitigated ambiguity, thereby corroborating the theoretical premise that structured supervision is key to enhancing long-context utility in dialogue systems.

Effect of Missing-Slot Guided Filtering. We assess the impact of the missing-slot guided filtering module by removing it while retaining the same hybrid RAG retrieval. As shown in Figure 4, the absence of filtering results in a significant performance drop: JGA on Qwen2.5-7B drops from 0.74 to 0.41, and DTE on GPT-4o-mini increases from 3.19 to 4.30. This highlights the critical role of fine-grained memory selection in both accuracy and interaction efficiency. This module operates in two stages: the CoT reasoner identifies missing task slots, which guide the memory filter to prioritize QA pairs that fill these gaps. This joint reasoning and filtering strategy significantly improves retrieval quality: compared to semantic-only retrieval, Recall@5 improves by 7.7% on average, raising the performance of text-embedding-3-small from 0.792 to 0.832. By bridging retrieval with task-specific gaps, the filter prioritizes slot-value pairs that are more likely to advance the dialogue.

Case Study

In our case study, we compared four methods for generating confirmation responses: (1) FCP, (2) Hybrid RAG, (3) ChatCite, and (4) MemGuide. As shown in Table 7, FCP often introduces irrelevant or outdated content due to unfiltered long context, while Hybrid RAG and ChatCite frequently miss critical slot values such as dates or preferences, reflecting retrieval noise or lack of task-specific reasoning. In contrast, MemGuide consistently produces accurate and concise responses by combining intent-aligned retrieval with missing-slot guided filtering. Notably, it achieves higher slot coverage and fluency than other methods, as verified by both GPT-4 scores and manual annotation. These results highlight the value of intent-aware retrieval and task-specific filtering for enhancing response quality in multi-session TOD.

Session 23	**Intention: The user intends to listen to songs.**
Dialogue Context	**User:** Can you play "Drive" again? **Assistant:**
FCP	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)*
Hybrid RAG	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)*
ChatCite	Sure! I've started playing "Drive" again. Is there anything else you'd like? *(Key Slot Omission)*
MemGuide	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)*

Table 7: Case study.

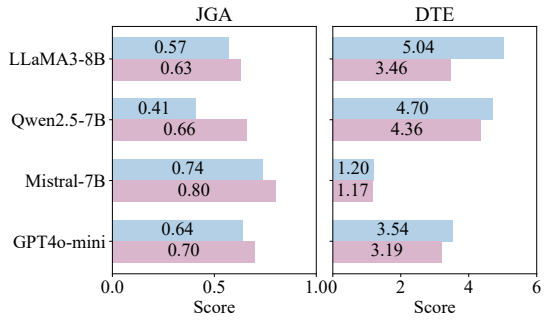

Figure 4: Comparison of Hybrid RAG (blue) vs. MemGuide (purple) across four LLMs on two metrics: JGA and DTE.

Conclusion

We present MemGuide, a two-stage memory-guided framework for multi-session LLM agents. By combining intent-aligned retrieval with missing-slot guided filtering, MemGuide enables task-aware, slot-specific memory selection that surpasses traditional semantic similarity. Evaluated on MS-TOD, our novel benchmark for multi-session TOD, MemGuide significantly improves task success, shortens dialogues, and enhances interaction coherence. These results confirm that structured memory supervision and goal-aware reasoning are critical for developing effective LLM agents.

Acknowledgements

This work is partially supported by Hong Kong RGC GRF No. 14206324, the National Natural Science Foundation of China 62576120, the Major Key Project of PCL2025A09 and Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No.2024ZD020.

References

Bang, N.; Lee, J.; and Koo, M.-W. 2023. Task-Optimized Adapters for an End-to-End Task-Oriented Dialogue System. In *Findings of the Association for Computational Linguistics: ACL 2023*, 7355–7369.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 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*, 5016–5026.

Chao, G.-L.; and Lane, I. 2019. BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer. *arXiv preprint arXiv:1907.03040*.

Chen, W.; Verga, P.; de Jong, M.; Wieting, J.; and Cohen, W. W. 2023. Augmenting Pre-trained Language Models with QA-Memory for Open-Domain Question Answering. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1597–1610. Dubrovnik, Croatia: Association for Computational Linguistics.

Chung, W.; Cahyawijaya, S.; Wilie, B.; Lovenia, H.; and Fung, P. 2023. Instructtods: Large language models for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2310.08885*.

Dai, Y.; He, W.; Li, B.; Wu, Y.; Cao, Z.; An, Z.; Sun, J.; and Li, Y. 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*, 4097–4111.

Du, M.; Cheng, L.; Xu, B.; Wang, Z.; Wang, S.; Yuan, J.; and Pan, C. 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)*, 1–8.

Du, Y.; Wang, H.; Zhao, Z.; Liang, B.; Wang, B.; Zhong, W.; Wang, Z.; and Wong, K.-F. 2024. PerLTQA: A Personal Long-Term Memory Dataset for Memory Classification, Retrieval, and Fusion in Question Answering. In Wong, K.-F.; Zhang, M.; Xu, R.; Li, J.; Wei, Z.; Gui, L.; Liang, B.; and Zhao, R., eds., *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 152–164. Bangkok, Thailand: Association for Computational Linguistics.

Feng, H.; Zhang, W.; Liu, J.; and Sun, M. 2024. LDST: A LLaMA-based Dialogue State Tracking Framework. *arXiv preprint arXiv:2403.12345*.

Feng, Y.; Lu, Z.; Liu, B.; Zhan, L.; and Wu, X.-M. 2023. Towards LLM-driven Dialogue State Tracking. arXiv:2310.14970.

Gulyaev, P.; Elistratova, E.; Konovalov, V.; Kuratov, Y.; Pugachev, L.; and Burtsev, M. 2020. Goal-Oriented Multi-Task BERT-Based Dialogue State Tracker. arXiv:2002.02450.

Heck, M.; van Niekerk, C.; Lubis, N.; Geishauser, C.; Lin, H.-C.; Moresi, M.; and Gašić, M. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. arXiv:2005.02877.

Hu, S.; Zhou, H.; Hergul, M.; Gritta, M.; Zhang, G.; Iacobacci, I.; Vulić, I.; and Korhonen, A. 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.

Hudeček, V.; and Dusek, O. 2023. Are Large Language Models All You Need for Task-Oriented Dialogue? In Stoyanchev, S.; Joty, S.; Schlangen, D.; Dusek, O.; Kennington, C.; and Alikhani, M., eds., *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 216–228. Prague, Czechia: Association for Computational Linguistics.

Lee, C.-H.; Cheng, H.; and Ostendorf, M. 2021. Dialogue State Tracking with a Language Model using Schema-Driven Prompting. arXiv:2109.07506.

Li, H.; Yang, C.; Zhang, A.; Deng, Y.; Wang, X.; and Chua, T.-S. 2024a. Hello Again! LLM-powered Personalized Agent for Long-term Dialogue. *arXiv preprint arXiv:2406.05925*.

Li, M.; Peng, B.; Gao, J.; and Zhang, Z. 2024b. Opera: Harmonizing task-oriented dialogs and information seeking experience. *ACM Transactions on the Web*, 18(4): 1–27.

Li, Y.; Chen, L.; Liu, A.; Yu, K.; and Wen, L. 2024c. ChatCite: LLM agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, Y.; Fang, Y.; Vandyke, D.; and Collier, N. 2024. TOAD: Task-Oriented Automatic Dialogs with Diverse Response Styles. *arXiv preprint arXiv:2402.10137*.

Lu, J.; An, S.; Lin, M.; Pergola, G.; He, Y.; Yin, D.; Sun, X.; and Wu, Y. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Madotto, A.; Lin, Z.; Winata, G. I.; and Fung, P. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Meta AI. 2024. Meta LLaMA 3.1-8B-Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Nguyen, Q.-V.; Nguyen, Q.-C.; Pham, H.; and Bui, K.-H. N. 2025. Spec-TOD: A Specialized Instruction-Tuned LLM Framework for Efficient Task-Oriented Dialogue Systems. *arXiv preprint arXiv:2507.04841*.

Peng, B.; Li, X.; Gao, J.; Liu, J.; and Wong, K.-F. 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)*, 2182–2192.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8689–8696.

Robertson, S.; and Zaragoza, H. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Ruan, Y.-P.; Ling, Z.-H.; Gu, J.-C.; and Liu, Q. 2020. Fine-Tuning BERT for Schema-Guided Zero-Shot Dialogue State Tracking. arXiv:2002.00181.

Shah, P.; Hakkani-Tür, D.; Tür, G.; Rastogi, A.; Bapna, A.; Nayak, N.; and Heck, L. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Stacey, J.; Cheng, J.; Torr, J.; Guigue, T.; Driesen, J.; Coca, A.; Gaynor, M.; and Johannsen, A. 2024. LUCID: LLM-Generated Utterances for Complex and Interesting Dialogues. *arXiv preprint arXiv:2403.00462*.

Team, M. A. 2024a. Mistral 7B: A High-Performance Language Model. *arXiv preprint arXiv:2402.12345*.

Team, O. 2024b. GPT-4: OpenAI's Advanced Language Model. *OpenAI Research*.

Team, Q. 2024c. Qwen-2.5: Advanced Large Language Model with Enhanced Capabilities. *arXiv preprint arXiv:2409.12345*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Batra, A.; Randriamihaja, S.; et al. 2024. LLaMA 3: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2401.00778*.

Wang, K.; Tian, J.; Wang, R.; Quan, X.; and Yu, J. 2020. Multi-Domain Dialogue Acts and Response Co-Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7125–7134.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gasic, M.; Barahona, L. M. R.; Su, P.-H.; Ultes, S.; and Young, S. 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*, 438–449.

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019a. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 808–819.

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019b. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. arXiv:1905.08743.

Xu, H.-D.; Mao, X.-L.; Yang, P.; Sun, F.; and Huang, H.-Y. 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)*, 2748–2763.

Xu, W.; Huang, Z.; Hu, W.; Fang, X.; Cherukuri, R.; Nayyar, N.; Malandri, L.; and Sengamedu, S. 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)*, 59–72.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 19724–19731.