ATOD: An Evaluation Framework and Benchmark for Agentic Task-Oriented Dialogue System

Anonymous ACL submission

Abstract

Recent advances in task-oriented dialogue (TOD) systems, driven by large language models (LLMs) with extensive API and tool integration, have enabled conversational agents to coordinate interleaved goals, maintain long-horizon context, and act proactively through asynchronous execution. These capabilities extend beyond traditional TOD systems, yet existing benchmarks lack systematic support for evaluating such agentic behaviors. To address this gap, we introduce **ATOD**, a benchmark and synthetic dialogue generation pipeline that produces richly annotated conversations requiring long-term reasoning. ATOD captures key characteristics of *Advanced TOD*, including multi-goal coordination, dependency management, memory, adaptability, and proactivity. Building on ATOD, we propose **ATOD-Eval**, a holistic evaluation framework that translates these dimensions into fine-grained metrics and supports reproducible offline and online evaluation. We further present a **strong agentic memory-based evaluator** for benchmarking on ATOD. Experiments show that ATOD-Eval enables comprehensive assessment across task completion, agentic capability, and response quality, and that the proposed evaluator offers a better accuracy–efficiency trade-off compared to existing memory- and LLM-based approaches under this evaluation setting.

Table 1: Example dialogue illustrating key agentic capabilities of advanced TOD systems, including multi-goal concurrency, interleaving, long-horizon memory, asynchronous execution, and proactivity.

User Utterance	Agent Response	Key Capabilities
Book a flight to Chicago for next Monday.	I found several options. Would you like me to confirm one now and also arrange a hotel near your destination?	Multi-goal concurrency
Yes, please. Also remind me to pack my passport the night before.	Flight confirmed and hotel reserved. A reminder has been set for Sunday night to pack your passport.	Long-horizon memory, Multi-goal concurrency
What's the weather in Chicago on Monday?	The forecast for Monday in Chicago is rainy, with a high of 12°C.	Interleaved workflows
Also book a dinner at an Italian restaurant for Monday night after my flight.	Your flight lands at 6:30pm. A table at *Luigi's* is reserved for 8:00pm.	Dependency management, Multi-goal concurrency
Arrange a ride from the airport to the hotel.	Ride scheduled for 6:45pm. Do you want the driver's details now or closer to the date?	Asynchronous execution, Proactivity
Closer to the date, please.	Got it. I will share the driver's info the day before your trip.	Long-horizon memory, Asynchronous follow-up

1 Introduction

Task-oriented dialogue (TOD) systems have traditionally been studied as conversational agents that fulfill user requests sequentially through turn-by-turn interactions (Rastogi et al., 2020; Cheng et al., 2022; Wang et al., 2023). With the advent of large language models (LLMs) and agentic AI systems such as ChatGPT (OpenAI, 2024), Claude (Anthropic, 2024), and Gemini (Comanici et al., 2025), expectations for TOD systems have grown substantially. Users now anticipate advanced capabilities, including managing multiple objectives concurrently (*multi-goal concurrency*), progressing while awaiting external API or tool responses (*asynchronous execution*), and flexibly suspending or resuming objectives within a dialogue (*interleaved workflows*). They further expect *proactivity*, where systems offer helpful assistance without digression, while dynamically handling evolving *goal dependencies*. Sustaining *long-horizon memory* is equally critical, as agents must integrate immediate conversational context with persistent knowledge across extended or multi-session interactions. Table 1 illustrates a representative interaction motivating the evaluation challenges considered in this work, where agents coordinate interdependent goals, preserve context, and enable asynchronous progress in non-sequential dialogues. Together, these characteristics define what we refer to as *Advanced TOD* and pose significant challenges for evaluation, requiring assessment not only of response quality and task completion, but also of their interaction in complex dialogue settings. Despite notable progress in automatic evaluation (Liu et al., 2023; Dubois et al., 2024; Zheng et al., 2023; Li et al., 2024; Yao et al., 2024; Jain et al., 2025; Acikgoz et al., 2025) and TOD dataset construction (Budzianowski et al., 2018; Rastogi et al., 2020; Du et al., 2025; Wang et al., 2023; Kulkarni et al., 2024), most benchmarks fail to capture the advanced characteristics outlined above, leaving these capabilities underexplored. In parallel, while recent work has begun to evaluate dialogue systems with memory components (Xu et al., 2025; Chhikara et al., 2025; Maharana et al., 2024; Ong et al., 2024), existing approaches lack standardized protocols for assessing long-horizon retention, adaptive updates, and the management of interleaved goals with complex dependencies. This gap highlights the need for a unified benchmark and holistic evaluation framework to systematically assess advanced TOD behaviors under realistic and complex interaction scenarios.

To fill this gap, we introduce **ATOD**, a *benchmark* and synthetic dialogue generation pipeline that produces richly annotated dialogues requiring long-term recall, interleaved workflows, and explicit goal dependencies. Building on ATOD, we propose **ATOD-Eval**, a holistic evaluation framework that provides standardized benchmarks and fine-grained metrics for systematically capturing advanced TOD capabilities. ATOD-Eval unifies evaluation and benchmarking by jointly assessing goal completion, dependency management, memory consistency, adaptability, proactivity, and multi-goal coordination, translating these dimensions into reproducible metrics for both offline and online settings. We further present a **proposed agentic memory-based evaluator** for benchmarking on ATOD, enabling empirical comparison against strong memory- and LLM-based baselines. Extensive experiments validate the proposed benchmark and evaluation framework, showing that the resulting metrics provide comprehensive and consistent assessment of advanced TOD capabilities, while the proposed evaluator consistently outperforms competitive baselines under this evaluation setting.

2 Related Work

2.1 TOD Systems Evaluation

Automatic evaluation frameworks such as G-Eval (Liu et al., 2023), AlpacaEval (Dubois et al., 2024), and MT-Bench (Zheng et al., 2023) benchmark open-domain dialogue, focusing on fluency and coherence rather than goal- or memory-driven behaviors. For TOD systems, earlier work emphasized turn-level user satisfaction (Walker et al., 2000; Schmitt and Ultes, 2015; Bodigutla et al., 2019), later extending to dialogue-level frameworks such as RoBERTaIQ (Gupta et al., 2021), USDA (Deng et al., 2022), and DQM (Komma et al., 2023). Other studies evaluate task completion using zero-shot LLM judges (Kazi et al., 2024) or interactive protocols with user simulators (Sun et al., 2021; Cheng et al., 2022; Davidson et al., 2023). More recent benchmarks, including AutoTOD (Xu et al., 2024), FNCTOD (Li et al., 2024), τ-Bench (Yao et al., 2024), AutoEval-ToD (Jain et al., 2025), and TD-EVAL (Acikgoz et al., 2025), focus on inform and success rates, without capturing advanced TOD capabilities.

2.2 TOD Datasets and Benchmarks

Human-curated datasets such as MultiWOZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), RADDLE (Peng et al., 2020), τ-Bench (Yao et al., 2024), and MS-TOD (Du et al., 2025) support dialogue state tracking and task completion, but offer limited long-horizon or multi-session memory. These datasets are largely confined to single sessions with narrowly scoped goals. Synthetic datasets, including TOPDIAL (Wang et al., 2023), TOAD (Liu et al., 2024), LUCID (Stacey et al., 2024), and SynthDST (Kulkarni et al., 2024), introduce personalization and proactivity, yet still fall short in supporting agentic behaviors.

2.3 Memory for Dialogue Systems

Memory mechanisms are critical for retaining context and managing goals over extended interactions. Early approaches such as RAG (Lewis et al., 2020), MemoChat (Lu et al., 2023), and MemoryBank (Zhong et al., 2024) enable session-level recall through retrieval, summarization, or history storage, but lack persistent memory across sessions. More recent agentic memory architectures, including MemGPT (Packer et al., 2023), A-Mem (Xu et al., 2025), mem0 (Chhikara et al., 2025), and MemOS (Li et al., 2025), introduce structured mechanisms for long-term retention. Other works, such as LoCoMo (Maharana et al., 2024), THEANINE (Ong et al., 2024), and MAP (Du et al., 2025), evaluate memory along temporal or efficiency dimensions. However, most studies treat memory in isolation, without standardized protocols linking memory usage to goal management in advanced TOD settings.

3 Problem Formulation

3.1 Characteristics of Advanced TOD

We introduce key characteristics of advanced TOD systems that pose realistic challenges and require long-horizon memory and agentic behaviors, with context carried across extended and interleaved interactions: **Multi-Goal Concurrency.** Users often pursue multiple objectives simultaneously, requiring agents to track and manage parallel goals with distinct states; **Interleaving.** Goals may be suspended, resumed, and alternated across contexts, rather than following strictly sequential workflows; **Long-Horizon Memory.** Goals can span many turns, requiring consistent state tracking and dependency management over extended interactions; **Asynchronous Execution.** Some goals are delayed (e.g., awaiting external confirmation), requiring agents to maintain a PENDING state and resume execution once conditions are met; **Proactivity.** Agents should take initiative by reminding users of pending tasks or suggesting relevant actions with appropriate context.

3.2 Task Formulation

Formally, let $\mathcal{D} = \{(\mathcal{G}_i, \mathcal{C}_i)\}_{i=1}^{N}$ denote a dialogue corpus, where $\mathcal{C}_i = \{c_{i,t}\}_{t=1}^{T_i}$ is the ordered sequence of dialogue turns and $\mathcal{G}_i$ is the associated set of user goals with explicit dependencies. Unlike traditional TOD settings where a goal is confined to a contiguous span, in advanced TOD a single goal $g \in \mathcal{G}_i$ may span disjoint intervals of $\mathcal{C}_i$, being initiated, suspended, and resumed as the dialogue evolves. We represent each goal with a *goal status trajectory* $\{\mathrm{Status}(g, t)\}_{t=1}^{T_i}$ (e.g., OPEN → PENDING → COMPLETED/FAILED), capturing a non-contiguous and interleaved lifecycle over extended interactions. The evaluation objective is to assess how well a system manages interdependent goals, maintains long-horizon trajectories, coordinates asynchronous workflows, and provides proactive support in both *offline* and *online* settings.

4 ATOD: A Synthetic Dialogue Dataset and Generation Pipeline

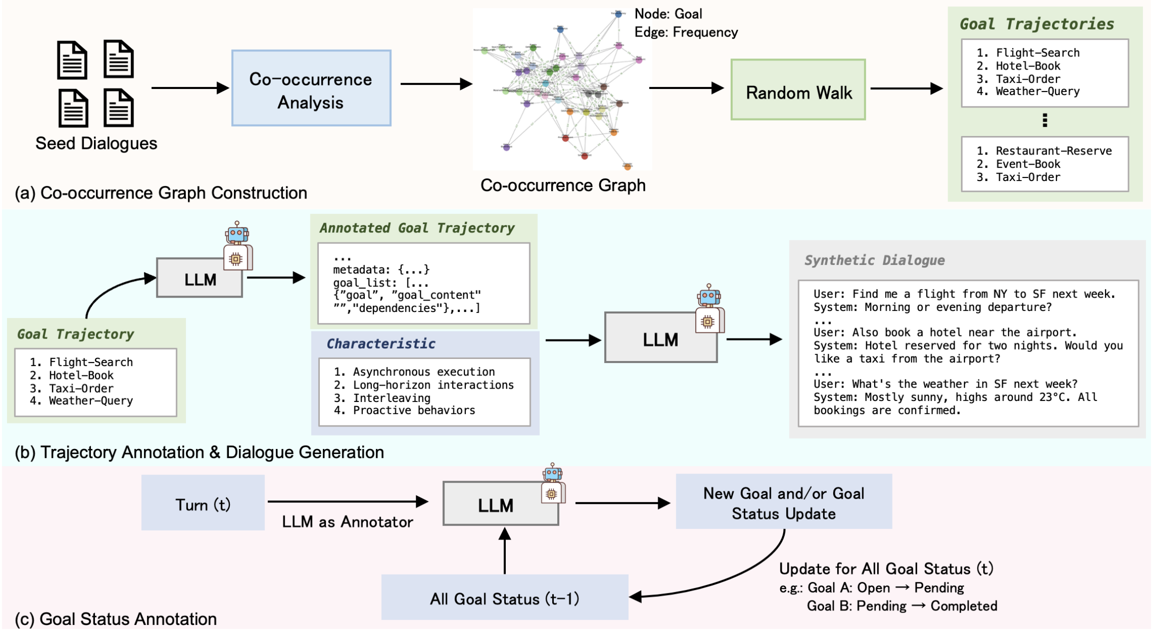
Building on the challenges above, we present **ATOD**, a synthetic dataset and generation pipeline designed to benchmark the advanced TOD characteristics introduced in §3.1. ATOD consists of richly annotated, memory-intensive dialogues that explicitly encode these characteristics, enabling systematic evaluation. As illustrated in Figure 1, the dataset is constructed through a modular LLM-driven pipeline (§4.1–§4.4) with quality control at each stage. We further analyze dataset coverage and compare ATOD with prior benchmarks in §4.5.

4.1 Co-occurrence Graph Construction and Goal Trajectory Sampling

To generate realistic synthetic dialogues, it is important to capture how goals naturally co-occur in user interactions. Independent goal sampling often yields implausible combinations, while fixed templates limit diversity. To address this, we construct a goal co-occurrence graph $G = (V, E)$ from an underlying dialogue dataset, where each node represents a unique goal and each weighted edge reflects empirical co-occurrence frequency. As shown in Figure 1(a), candidate goal sets $S = \{g_1, \dots, g_k\}$ are sampled via stratified random walks of varying lengths over G, preserving realistic correlations while introducing diversity. This procedure is dataset-agnostic; in our experiments, we instantiate it on the Schema-Guided Dialogue (SGD) corpus (Rastogi et al., 2020). The resulting dialogues exhibit richer domain variation and naturally support multi-goal, interleaved, and long-horizon interactions.

4.2 Annotation of Goal Trajectories and Complexity Categorization

As illustrated in Figure 1(b), each sampled goal set S is instantiated into a concrete trajectory via LLM-based annotation, yielding (i) slot values, (ii) inter-goal dependencies D_S capturing prerequisite or blocking relations (e.g., PAYMENT depending on BOOKING), and (iii) natural-language goal descriptions. Each trajectory is assigned a complexity label $c(S)$ reflecting both quantitative attributes (e.g., number of goals, dependency density) and qualitative factors (e.g., interleaving or opportunities for proactivity), with detailed criteria provided in Appendix A.3. This categorization ensures coverage across complexity levels and enables structured evaluation of agentic behaviors. Quality control is applied throughout: the LLM filters duplicate or incompatible goals during sampling, and automatic retries together with LLM-based checks verify slot validity, dependency consistency, and linguistic flu-



```
...
metadata: {...}
goal_list: [...
{"goal", "goal_content"
"","dependencies"},...]
```

```
User: Find me a flight from NY to SF next week.
System: Morning or evening departure?
...
User: Also book a hotel near the airport.
System: Hotel reserved for two nights. Would you
like a taxi from the airport?
...
User: What's the weather in SF next week?
System: Mostly sunny, highs around 23°C. All
bookings are confirmed.
```

Figure 1: ATOD dataset curation pipeline. (a) **Co-occurrence Graph & Trajectory Sampling** (§4.1): Construct a goal co-occurrence graph from seed dialogues and sample diverse multi-goal trajectories via random walks; (b) **Trajectory Annotation & Dialogue Generation** (§4.2–§4.3): An LLM annotates slot values, inter-goal dependencies, and complexity, then generates agentic multi-turn dialogues conditioned on the trajectories; (c) **Goal Status Annotation** (§4.4): At each turn, an LLM labels active goals and updates lifecycle states, enabling fine-grained tracking of dialogue progress.

ency during annotation (Appendix A.5).

4.3 Dialogue Generation

As shown in Figure 1(b), dialogue synthesis is conditioned on the annotated trajectory $\tau = (S, D_S, c(S))$. The goals, dependencies, and targeted complexity profile are combined into structured prompts for LLM-based generation (templates in Appendix A.6). The LLM then produces a natural multi-turn conversation $\mathcal{C} = \{c_t\}_{t=1}^{T}$ that realizes the specified goals while exhibiting interleaving, asynchronous execution, proactive assistance, and dependency-aware coordination.

4.4 Turn-level Goal Status Annotation

Finally, as illustrated in Figure 1(c), an LLM annotator performs iterative turn-level analysis to label the status of each goal at every dialogue turn. Each utterance c_t is annotated with an active goal set $\mathcal{A}_t \subseteq S$ and corresponding statuses $\mathrm{Status}(g, t) \in \{\textsc{Not\_Mentioned}, \textsc{Open}, \textsc{Pending}, \textsc{Completed}, \textsc{Failed}, \textsc{Abandoned}\}$. This design allows goals to be initiated, suspended, resumed, or terminated over time rather than confined to contiguous spans. The resulting turn-aligned `status_history` provides a rich reference for benchmarking multi-goal tracking and asynchronous or interleaved progressions.

4.5 Dataset Coverage

ATOD spans diverse domains and goal complexities, ranging from simple two-goal cases to interdependent, long-horizon workflows. Table 2 compares ATOD with existing benchmarks. While prior datasets capture individual aspects of advanced TOD, ATOD uniquely combines multi-goal concurrency, interleaving with asynchronous execution, explicit dependency modeling, proactive behaviors, and turn-level status annotation. This makes ATOD the first dataset purpose-built to comprehensively support evaluation of advanced TOD systems.

5 Agentic Memory System

Building on ATOD, we introduce ATOD-Eval's *agentic memory system* (Fig. 2), which serves as the evaluation backbone for advanced TOD. While ATOD provides annotated dialogues for benchmarking, the memory system evaluates models directly on dialogue text by assessing whether they can consistently maintain and update goal trajectories throughout interaction. It consists of two key modules: (i) a **dual memory store** (§5.1), and (ii) a **turn-level processing pipeline** (§5.2).

Table 2: Comparison of ATOD with representative TOD benchmarks. "Avg. Turns" denotes per-dialogue averages. Rightmost column (*Goal Status Anno.*) refers to explicit per-turn labeling of each goal's lifecycle state (e.g., PENDING, COMPLETED, FAILED). Other columns indicate support for key agentic features: asynchronous goal management, explicit dependency modeling, interleaving, and proactive behaviors (✓: present, ✗: absent).

Dataset	Avg. Turns	Async	Dependency	Interleaving	Proactive	Goal Status Anno.
MultiWOZ (Budzianowski et al., 2018)	13	✗	✗	✗	✗	✗
SGD (Rastogi et al., 2020)	20	✗	✗	✗	✗	✗
TOPDIAL (Wang et al., 2023)	12	✗	✗	✗	✓	✗
MS-TOD (Du et al., 2025)	7	✗	✓	✓	✗	✓
TOAD (Liu et al., 2024)	5	✗	✓	✓	✓	✗
LUCID (Stacey et al., 2024)	21	✓	✓	✗	✗	✓
ATOD (Ours)	54	✓	✓	✓	✓	✓

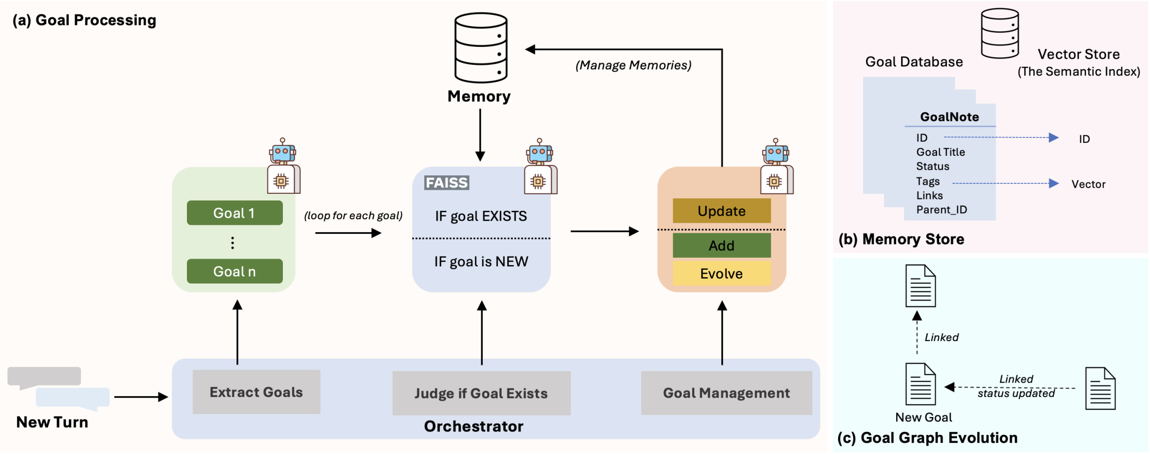


Figure 2: Architecture of the agentic memory system. (a) Turn-level pipeline for goal extraction, existence checking, updating/inserting, and proactive auditing. (b) Dual memory store with symbolic metadata and semantic embeddings. (c) Dependency graph evolution when inserting new goals, with explicit links and status transitions.

5.1 Dual Memory Store

As shown in Fig. 2(b), the memory system maintains a *dual memory store* consisting of: (i) a **structured goal database** $\mathcal{D}_{\text{sym}}$, which persistently records symbolic metadata (e.g., goal content and status), and (ii) a **semantic vector store** $\mathcal{D}_{\text{vec}}$ (e.g., FAISS (Douze et al., 2024)), which indexes embeddings of goal metadata for similarity-based retrieval.

Each goal g is stored as g = {`id`, `status`, `status_history`, `goal_description`, `dependencies`, `parent_id`, `embedding`}, where `status` ∈ {OPEN, PENDING, COMPLETED, FAILED, ABANDONED}; `status_history` logs turn-level transitions; `goal_description` provides a standardized textual form; `dependencies` and `parent_id` encode inter-goal relations; and `embedding` supports semantic retrieval during interaction. This dual design combines accurate symbolic state tracking with flexible semantic matching, supporting robust memory management over long-horizon dialogues.

5.2 Turn-Level Processing Pipeline

At each dialogue turn t, the memory system applies a structured *turn-level processing pipeline* to maintain the lifecycle of all goals. The pipeline consists of four stages: (i) **goal extraction** from the current utterance and context, (ii) **existence checking** against the dual memory store, (iii) **updating or inserting** goals with dependency evolution, and (iv) **proactive auditing** to keep active states consistent. This design enables dynamic tracking and interleaving of multi-goal trajectories over long horizons.

Formally, given user utterance u_t and context c_t, the system extracts candidate goals $\mathcal{G}_t$. Each candidate $g_t^{(i)} \in \mathcal{G}_t$ is matched against the memory store to determine whether to update an existing entry or insert a new one:

$$(u_t, c_t) \xrightarrow{\text{extract}} \mathcal{G}_t$$

$$g_t^{(i)} \xrightarrow{\text{match}} \begin{cases} \text{update}(g^*), & \text{if Match=1}, \\ \text{insert} + \text{evolve}(g_t^{(i)}), & \text{if Match=0}. \end{cases}$$

Stage 1. Existence Checking. For each candidate $g_t^{(i)}$, the system retrieves top-k neighbors $\mathcal{N}_k(g_t^{(i)})$ from $\mathcal{D}_{\text{vec}}$ and applies an LLM-based judge f_{judge} for semantic verification. If confidence $\geq \tau$, Match=1; otherwise, Match=0.

Stage 2. Updating Existing Goals. When Match=1, the *Update* module advances the goal lifecycle (e.g., PENDING → COMPLETED), refreshes slot values and dependencies, and preserves existing inter-goal relations.

Stage 3. Adding and Evolving New Goals. When Match=0, the new goal is inserted into both $\mathcal{D}_{\text{sym}}$ and $\mathcal{D}_{\text{vec}}$. The *Evolve* module links it to related goals $\{g_k \mid \text{rel}(g_t^{(i)}, g_k) \geq \delta\}$, updating the directed dependency graph $G = (\mathcal{V}, \mathcal{E})$ to support interleaved workflows. Capturing such dependencies is essential in advanced TOD, where goals are often logically conditioned on others (e.g., PAYMENT following BOOKING). Maintaining these relations prevents premature completion and enables faithful modeling of complex task dynamics.

Stage 4. Proactive Status Tracking. Beyond event-driven updates, a background auditing process periodically inspects active goals (OPEN, PENDING) against dialogue context and tool outputs. An LLM judge triggers valid transitions (e.g., PENDING → COMPLETED), preventing stale states and ensuring coherence across dependent goals.

Together, these modules maintain consistent, dependency-aware goal states, supporting concurrency and reliable evaluation for advanced TODs.

6 Evaluation Metrics and Framework

Having established ATOD as a benchmark dataset (§4.5) and introduced the agentic memory system that tracks evolving goals (§5), we now present the evaluation framework of **ATOD-Eval**. This framework defines metrics and protocols that assess not only *whether* a system completes tasks, but also *how effectively* it manages complex, interdependent dialogues. ATOD-Eval spans three dimensions: (i) *Task Completion and Efficiency* (§6.1), (ii) *Agentic Capability Metrics* (§6.2), and (iii) *Response Quality Metrics* (§6.3). A unified framework (§6.4) supports both offline and online evaluation.

6.1 Task Completion and Efficiency

We evaluate whether goals are accomplished and how efficiently they progress through the dialogue.

Dependency-Aware Goal Completion Rate (dGCR). Conventional goal completion metrics treat all goals equally, unfairly penalizing systems when goals remain blocked by unmet prerequisites. We define a dependency-aware variant that considers only goals whose prerequisites are satisfied. Formally, let $S(g)$ denote the status of goal g in $\mathcal{D}_{\text{sym}}$, and let $\mathcal{U}_{\text{dec}} = \{g \in \mathcal{U} \mid S(g) \in \{\text{COMPLETED}, \text{FAILED}\}\}$. Then, $\text{dGCR} = \frac{|\{g \in \mathcal{U} : S(g) = \text{COMPLETED}\}|}{|\mathcal{U}_{\text{dec}}|}$. This formulation avoids bias from dependency-locked goals and provides a faithful measure of system performance in multi-goal workflows.

Turns to Completion (NTC). For each completed goal, NTC computes the average number of turns between initiation and completion, capturing execution efficiency and complementing dGCR.

6.2 Agentic Capability Metrics

Beyond task success, we assess whether systems exhibit agentic behaviors such as memory recall and proactive action.

Memory Recall Accuracy. This metric measures the proportion of retrieval queries whose outputs match the ground-truth memory state, including slot values, goal statuses, and historical context.

Proactivity Effectiveness. We evaluate proactive behaviors by identifying goal or state changes initiated without explicit user prompts and assessing whether these actions are contextually appropriate and beneficial.

6.3 Response Quality Metrics

In addition to task outcomes and agentic behaviors, we assess conversational quality, focusing on *turn-level relevance* and *dialogue-level coherence*, following prior work (Liu et al., 2023; Dubois et al., 2024; Zheng et al., 2023). These metrics ensure that systems maintain natural and consistent interactions alongside effective goal management.

6.4 Evaluation Framework

Together, these metrics form a unified framework that jointly evaluates task outcomes, agentic behaviors, and conversational quality. ATOD-Eval supports both **offline** benchmark analysis and **online** tracking, enabling consistent assessment across static datasets and real-time deployments.

Table 3: Comparison of goal detection accuracy and status tracking accuracy for each method, broken down by dialogue complexity. All results are reported as percentages (%) and averaged over the test set.

Category	**Method**	**Medium**		**Complex**	
		Goal Detection F1	Status Tracking Acc.	Goal Detection F1	Status Tracking Acc.
LLM-based	`DeepSeek-R1`	52.84	96.36	36.63	74.42
	`Claude-3.5-Sonnet`	74.08	92.94	82.92	76.10
	`Claude-3.7-Sonnet`	76.67	92.47	72.97	78.64
	`Claude-4-Sonnet`	78.95	93.26	75.58	84.26
Memory-based	RAG (Lewis et al., 2020)	85.59	94.85	87.13	77.83
	MemoChat (Lu et al., 2023)	80.38	73.88	58.07	66.83
	MemoryBank (Zhong et al., 2024)	82.56	94.23	76.86	78.50
	LLM-Rsum (Wang et al., 2025)	93.83	89.47	89.13	69.95
	Ours	**91.92**	**92.31**	**86.49**	**84.28**

7 Experimental Setup

We evaluate the framework's *capability*, *validity*, and *efficiency* using task and cost metrics: (1) **Module Capability.** We assess whether the agentic memory system supports both final and online evaluation by reporting *Goal Detection Accuracy* (coverage of correctly identified active goals) and *Status Tracking Accuracy* (state classification accuracy among detected goals), measured at the final dialogue state and across normalized dialogue progress. Implementation details are provided in Appendix B.1; (2) **Metric Validity.** We examine whether the proposed metrics reflect task success by analyzing their correlations with *Dependency-Aware Goal Completion Rate (dGCR)*, reporting Pearson's r and Spearman's ρ. The evaluated metrics include *Turns to Completion (NTC)*, *Memory Recall Accuracy*, *Proactivity Effectiveness*, and subjective response quality at both turn and dialogue levels; (3) **Efficiency.** We measure computational cost via per-turn update latency and average token usage to assess scalability under increasingly complex dialogue conditions.

We compare against two classes of baselines: (i) **LLM-based judges**: following (Kazi et al., 2024), we prompt LLMs (`Claude-3.5-Sonnet`, `Claude-3.7-Sonnet`, `Claude-4-Sonnet`, `DeepSeek-R1`) in a zero-shot manner to infer goal status and task completion; (ii) **Memory-based evaluators**: we adapt representative memory-augmented frameworks, including RAG (Lewis et al., 2020), MemoChat (Lu et al., 2023), MemoryBank (Zhong et al., 2024), and LLM-Rsum (Wang et al., 2025). Since these architectures primarily target open-domain retention, we adapt their prompting strategies to align with our specific goal status schema, enabling fair comparison.

8 Results

8.1 Evaluation of the Memory System

Table 3 reports goal detection and status tracking results under medium and complex dialogues. Memory-based approaches substantially outperform LLM judges, highlighting the importance of explicit memory structures for advanced TOD evaluation. Among them, our method achieves competitive accuracy in medium settings and exhibits stronger robustness in complex ones, where most baselines experience notable degradation. These results indicate that our memory system maintains reliable goal tracking under challenging conditions and offers more stable performance than prior approaches as dialogue complexity increases.

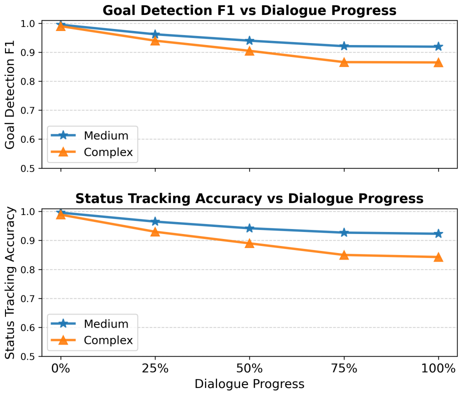


Figure 3: Goal detection F1 (top) and status tracking accuracy (bottom) vs. normalized dialogue progress (0–100%) under Medium and Complex settings.

We further analyze performance as a function of dialogue progress, as shown in Figure 3. Both goal detection and status tracking exhibit near-perfect accuracy at early stages and remain stable as the dialogue unfolds, with only mild degradation even in complex cases. This stability supports reliable online evaluation and aligns with the design goal of ATOD-Eval to emphasize dependency-aware tracking under complex and long-horizon dialogues.

8.2 Efficiency Analysis

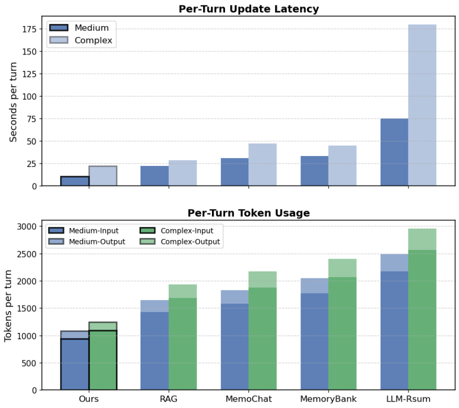


Figure 4: Per-turn update latency (top) and token usage (bottom) across methods. Latency is reported as mean per-turn time with range bars; token usage reports mean input and output tokens per turn under Medium and Complex settings.

As shown in Figure 4, our method achieves the lowest per-turn update latency, remaining below 25 seconds even for complex dialogues, while baseline systems incur much higher costs (e.g., LLM-Rsum exceeds 180 seconds per turn). Latency results are reported as mean values with range bars computed from log segments. For token usage, our method also consistently consumes fewer input and output tokens in both Medium and Complex dialogues, achieving substantial savings over all baselines. This efficiency stems from selective goal matching and lightweight updates, which reduce redundant LLM calls and allow the system to scale effectively under increasing dialogue complexity.

8.3 Metric Validity Analysis

Table 4: Average results of proposed evaluation metrics across medium- and complex-complexity dialogues.

Metric	**Medium**	**Complex**
dGCR	0.967	0.930
Turns to Completion	7.04	10.50
Memory Recall Accuracy	0.913	0.743
Proactivity Effectiveness	0.619	0.586
Turn-level Quality	0.752	0.766
Dialogue-level Quality	4.40	4.45

We first summarize the average values of our proposed metrics (Table 4), which reflect different dimensions of system behavior: efficiency, memory, proactivity, and interaction quality. As expected, medium dialogues are shorter and yield higher memory recall, whereas complex dialogues require more turns and show reduced recall, consistent with their higher difficulty.

Table 5: Correlation of proposed evaluation metrics with dGCR, reported as Pearson's r and Spearman's ρ under Medium and Complex settings.

Metric	**Medium**		**Complex**	
	r	ρ	r	ρ
Turns to Completion	+0.08	+0.16	+0.20	+0.05
Memory Recall Accuracy	+0.75	+0.60	+0.44	+0.43
Proactivity Effectiveness	−0.05	−0.03	+0.16	+0.12
Turn-level Quality	+0.22	+0.29	+0.08	+0.09
Dialogue-level Quality	−0.11	−0.08	+0.13	+0.25

To examine validity, we then analyze correlations with *dGCR* (Table 5). Among all metrics, *Memory Recall Accuracy* correlates most strongly with dGCR in both settings, highlighting the role of accurate memory in dependency-aware success. *Turns to Completion* and *Turn-level Quality* show weaker but complementary alignment, capturing efficiency and local interaction quality. *Proactivity Effectiveness* correlates only marginally, suggesting that richer proactive scenarios would be needed to reveal its value. Overall, the metrics provide complementary perspectives: some align closely with dependency-sensitive success, while others contribute efficiency- and quality-oriented signals.

9 Conclusions

We introduced **ATOD**, a benchmark that captures key characteristics of advanced task-oriented dialogue, including multi-goal concurrency, dependency management, long-horizon memory, asynchrony, and proactivity, together with turn-level goal status annotations for fine-grained evaluation. Building on this benchmark, we proposed **ATOD-Eval**, a holistic evaluation framework that translates these capabilities into reproducible metrics for offline and online settings. We further presented a **proposed agentic memory-based evaluator** for benchmarking on ATOD. Experimental results show that, under the proposed evaluation setting, this evaluator consistently outperforms LLM- and memory-based baselines on goal detection and status tracking, while incurring lower update latency and token usage. Overall, ATOD and ATOD-Eval provide a unified and scalable foundation for evaluating next-generation TOD systems.

Limitations

This work evaluates advanced task-oriented dialogues under a fixed set of dialogue attributes and does not incorporate user-specific contextual signals into either response generation or evaluation. In real-world deployments, contextual factors like user demographics, long-term preferences, and interaction history may significantly influence dialogue dynamics and task outcomes. While persona-augmented multi-turn dialogue settings have been explored in prior work, they differ from the ATOD scenarios considered here and are therefore outside the scope of the current benchmark. Additionally, the proposed framework is restricted to text-based dialogue attributes and agent responses. Although this setting aligns with many existing conversational and voice-based systems, it does not capture richer multimodal interactions involving visual or other non-textual signals. Consequently, modality-specific challenges and interactions are not reflected in the current evaluation.

Ethical Considerations

This work introduces a synthetic benchmark and evaluation framework for agentic task-oriented dialogue systems. Since the dataset is constructed via an LLM-driven pipeline using the public Schema-Guided Dialogue (SGD) dataset as a seed, it does not contain real user data or Personally Identifiable Information (PII). However, we acknowledge that synthetic dialogues generated by Large Language Models may inherently reflect the biases present in the underlying models. While we applied multi-stage quality control and filtering to ensure the relevance and safety of the content, users of this benchmark should be aware of these potential limitations. This dataset is intended solely for research purposes to advance the evaluation of complex dialogue capabilities.

References

Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Td-eval: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. *arXiv preprint arXiv:2504.19982*.

Anthropic. 2024. Building effective agents. `https://www.anthropic.com/engineering/building-effective-agents`. Accessed: July 2025.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1908.07064*.

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

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is multiwoz a solved task? an interactive tod evaluation framework with user simulator. *arXiv preprint arXiv:2210.14529*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User simulation with large language models for evaluating task-oriented dialogue. *arXiv preprint arXiv:2309.13233*.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *Proceedings of the ACM Web Conference 2022*, pages 2998–3008.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z Pan, Ruifeng Xu, and Kam-Fai Wong. 2025. Bridging the long-term gap: A memory-active policy for multi-session task-oriented dialogue. *arXiv preprint arXiv:2505.20231*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Chenlei Edward Guo. 2021. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems.

Arihant Jain, Purav Aggarwal, Rishav Sahay, Chaosheng Dong, and Anoop Saladi. 2025. Autoeval-tod: Automated evaluation of task-oriented dialog systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10133–10148.

Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tür, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 913–920. IEEE.

Abishek Komma, Nagesh Panyam Chandrasekarasastry, Timothy Leffel, Anuj Goyal, Angeliki Metallinou, Spyros Matsoukas, and Aram Galstyan. 2023. Toward more accurate and generalizable evaluation metrics for task-oriented dialogs. *arXiv preprint arXiv:2306.03984*.

Atharva Kulkarni, Bo-Hsiang Tseng, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Hong Yu, and Shruti Bhargava. 2024. Synthdst: Synthetic data is all you need for few-shot dialog state tracking. *arXiv preprint arXiv:2402.02285*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, et al. 2025. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. 2024. Towards lifelong dialogue agents via timeline-based memory management. *arXiv preprint arXiv:2406.10996*.

OpenAI. 2024. Introducing chatgpt agent. https://openai.com/index/introducing-chatgpt-agent/. Accessed: July 2025.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2020. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. *arXiv preprint arXiv:2012.14666*.

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

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36.

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*.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3-4):363–377.

Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. *arXiv preprint arXiv:2310.07397*.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193.

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

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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

A Appendix A

A.1 Synthetic Dataset Quality Analysis

We assess the quality of synthetic dialogues along five dimensions: coherence, fluency, consistency, relevance, and naturalness (Liu et al., 2023). As shown in Table 6, both medium- and complex-level dialogues achieve strong results across all criteria, particularly in fluency and relevance. These results demonstrate that our generation pipeline produces realistic and high-quality conversations suitable for downstream evaluation.

Table 6: LLMs evaluation of synthetic dialogues along five dimensions. Scores are averaged over medium- and complex-level dialogues, reported on a 1–5 Likert scale (higher is better).

Dimension	Medium	Complex
Coherence	4.04	3.92
Fluency	5.00	5.00
Consistency	4.42	4.04
Relevance	4.58	4.62
Naturalness	4.04	4.04

In addition, our pipeline employs a separate LLM-based judge at multiple stages (§4.2, §4.3, and §4.4), including trajectory sampling, goal annotation & classification, and status annotation, to ensure the quality of outputs at each step of the generation process. This layered evaluation helps maintain both faithfulness and consistency throughout the synthetic dialogue construction.

A.2 Goal Extraction, Co-occurrence Graph Statistics, and Sampling Strategy

We first extract goal sequences from the SGD dataset, where each sequence is an ordered list of user goals (domain–intent pairs) within a dialogue. Table 7 summarizes the extraction results. All 10,739 sequences are multi-domain, with an average length of 3.90 goals and a range of 2–8 goals. These sequences span 16 unique domains and 37 unique intents, providing a rich basis for building the co-occurrence graph.

Statistic	Value
Total Goal Sequences	10,739
Avg. Sequence Length	3.90
Length Range	2–8
Unique Domains	16
Unique Intents	37

Table 7: Summary statistics for extracted goal sequences.

We then construct a goal co-occurrence graph where each node is a unique goal and edges represent co-occurrence within the same sequence. Table 8 shows its statistics. The graph contains 52 nodes and 396 edges, forming a single connected component with relatively high density (0.2986) and average degree (15.23), indicating frequent goal co-occurrence across dialogues. This structure supports diverse sampling of multi-goal trajectories, including high-degree hubs (up to 29) and rare goals (degree as low as 2).

Statistic	Value
Total Nodes (Unique Goals)	52
Total Edges (Co-occurrences)	396
Graph Density	0.2986
Average Degree	15.23
Max Degree	29
Min Degree	2

Table 8: Summary statistics for the co-occurrence graph.

We sample goal trajectories from this graph by selecting connected subgraphs that satisfy the desired complexity criteria (§A.3), ensuring diversity in goal count, domain coverage, and dependency patterns.

A.3 Complexity Criteria

Our pipeline (§ 4) uses a two-category complexity system (*medium* vs. *complex*), combining quantitative thresholds with qualitative LLM analysis for balanced distribution. Table 9 shows the criteria based on goals, turns, domains, and advanced agentic behaviors.

Compl.	Goals	Turns	Async.	Inter.	Dep.	Proac.	Def.
Medium	2–8	8–35	✓	✓	≤ 2	✗	✗
Complex	7+	30+	✓	✓	≥ 2	✓	✓

Table 9: Criteria for medium vs. complex dialogues. Columns: Goals, Turns, Async. (asynchronous), Inter. (interleaving), Dep. (dependencies), Proac. (proactivity), Def. (defectiveness). ✓ = present, ✗ = absent. Ambiguous cases are resolved using domain diversity, dependency depth, and behaviors.

For categorization process, we follow a three-step procedure. First, *goal sampling* draws trajectories under a two-category distribution (default: 65% medium, 35% complex). Second, *annotation* enriches sampled goals with slots, dependencies, and realistic characteristics. Third, *hybrid classification* assigns complexity using pre-defined rules combined with LLM analysis, considering quantitative factors (goal count, domain diversity, dependency structures), qualitative factors (goal interdependence and coordination complexity), and realistic dialogue requirements such as interleaving and proactivity needs.

A.4 Annotated Trajectories and Metadata Specification

Table **??** shows the template used in § 4.4 for turn-level goal status annotation. The prompt is instantiated with the current turn, the list of goals with their current statuses, and the expected JSON schema. A complete example of the resulting annotated dialogue is provided in Listing 2.

```
{
  "dialogue_id": "string",
  "complexity_class": "medium | complex"
    ,
  "metadata": {
    "num_goals": "integer",
    "estimated_turns": "integer",
    "async_execution": "boolean",
    "interleaving": "boolean",
    "proactivity": "boolean"
  },
  "goal_list": [
    {
      "id": "string",
      "domain": "string",
      "intent": "string",
      "slots": ["string", ...],
      "slot_values": {
        "slot_name_1": "value1",
        "slot_name_2": "value2"
      },
      "dependencies": ["goal_id", ...],
      "content": "string",
      "core_content": "string",
      "classification_method": "
          pre_defined | model_based",
      "dependency_label": "boolean",
      "defectiveness_label": "boolean"
    }
    // ...more goals
  ]
}
```

Listing 1: Annotation schema for dialogue trajectories and goal-level metadata.

A.5 ATOD: Quality Control

We use the following LLM-based quality control prompt to verify goal clarity, slot validity, and annotation consistency of annotated goals (§4.2) prior to dialogue generation.

Quality Assessment Prompt

`You are a quality judge for annotated goal trajectories.`

Input:
TRAJECTORY ({num_goals} goals, {complexity} complexity):
{goals_text}

Task:
Assess whether this trajectory is ready for dialogue generation. Check:

- Goal descriptions are clear and specific
- Slot values are realistic (no placeholders)
- All required fields are present
- Annotations are logically consistent

Output format:
Respond with exactly one word: `PASS` or `FAIL`

A.6 Dialogue Generation Prompt

Below, we present the exact prompt template used in §4.3 to instantiate LLM-based dialogue generation. Placeholders (e.g., {complexity}, {estimated_turns}, {goal_descriptions}, {agentic_attrs}) are filled programmatically from the annotated trajectory metadata, as described in §A.4.

Dialogue Generation Prompt Template

`Generate a realistic task-oriented dialogue between USER and SYSTEM.`

Requirements:

- **Complexity:** {complexity}
- **Length:** {estimated_turns} turns
- **Goals:** {goal_descriptions}
- **Attributes:** {agentic_attrs}

{combined_guidance}
{outcome_guidance}

Dialogue Structure:

1. User introduces goals naturally throughout the conversation
2. System works on goals under realistic constraints and limitations
3. Natural obstacles, delays, and preference changes may occur
4. The dialogue ends at a natural stopping point
5. Goal completion may be partial or conditional, reflecting real-world scenarios

Natural Conversation Patterns:

- Users express needs and preferences as they arise
- System responds helpfully while handling practical constraints
- Users may add, revise, or abandon goals based on new information
- Availability, pricing, or technical limitations may surface
- Conversations conclude when users are satisfied or defer decisions

Format: Alternating USER/SYSTEM turns, starting with USER.

A.7 Goal Status Annotation Prompt

Below, we present the exact prompt template used in §4.4 for turn-level goal status annotation. The prompt is instantiated with the current dialogue turn, the list of goals with their current statuses, and the expected JSON schema. Listing 2 further provides a sample annotated dialogue instance, illustrating how goal status transitions and full goal states (`all_goals`) are tracked across turns.

Goal Status Annotation Prompt Template

`You are tracking goal status in a task-oriented dialogue. Analyze only the current turn and update statuses based on what actually happens.`

Current Turn to Analyze:
{last_turn}

Goals and Current Statuses:
{goal_descriptions}

Status Meanings:

- `NOT_MENTIONED`: goal exists but has not appeared in the dialogue
- `OPEN`: mentioned by the user, no action started
- `PENDING`: system is actively working on the goal
- `COMPLETED`: goal successfully finished
- `FAILED`: goal failed due to system or availability issues
- `ABANDONED`: user cancelled or changed their mind

Critical Rules:

1. Change a goal's status *only* if something definitive occurs in the current turn
2. `PENDING` goals may transition only to `COMPLETED`, `FAILED`, or `ABANDONED`
3. If no clear change occurs, preserve the existing status

Terminal States (Do Not Change):
{goal_id: current_status, . . . }

Current Statuses (JSON Template):
`{json_template}`

Instruction: Respond with *only* the JSON above, updating *only* goals whose status clearly changes in the current turn.

```
{
  "dialogue_id": "...",
  "complexity_class": "complex",
  "metadata": {
    "num_goals": ...,
    "num_turns": ...,
    "async_execution": true,
    "interleaving": true,
    "proactivity": true
  },
  "goal_list": [...],
  "turns": [
    {
      "turn_id": 1,
      "speaker": "USER",
      "utterance": "I need to book a
          hotel in Chicago.",
      "goal_status_changes": [
        {"goal_id": "g1", "new_status":
            "open"}
      ],
```

```
      "all_goals": {
        "g1": "open",
        "g2": "not_mentioned",
        "g3": "not_mentioned"
      }
    }
    // ... remaining turns omitted
  ]
}
```

Listing 2: Sample annotated ATOD dialogue with turn-level status tracking.

B Appendix B

B.1 Implementation Details

Our memory system is instantiated with `Claude-3.7-Sonnet` (accessed via the Amazon Bedrock API) as the primary LLM judge. For embedding-based retrieval, we use `MiniLM-L6-v2` embeddings indexed with FAISS for efficient nearest-neighbor search.

B.2 Agentic Memory System Templates

As detailed in §5, the agentic memory system is implemented through a set of modular LLM prompt templates. Below, we present the templates used for (i) goal extraction from individual conversation turns, (ii) turn-level goal status classification, and (iii) goal graph evolution for establishing links and dependencies. Together, these templates enable structured, consistent, and interpretable memory management across multi-turn dialogues.

Goal Extraction Prompt

`Extract user goals from this conversation turn. Use standardized core_content patterns.`

Conversation Turn:
User: {user_utterance}
System: {system_response}

Core Content Patterns (examples):

- `"book hotel"` — hotels, rentals
- `"book flight"` — flights
- `"book ticket"` — bus, concert, train
- `"check account"` — balance, account information
- `"search restaurant"` — restaurant discovery
- `"book restaurant"` — reservations

Status Labels: `OPEN`, `PENDING`, `COMPLETED`, `FAILED`

Output format (JSON array):
`[ {"goal_content": "...", "core_content": "...", "status": "OPEN"}, ... ]`

Goal Status Classification Prompt

`Analyze this conversation turn and classify the status of the specific goal below.`

Goal to classify:
`"{goal_content}"`

Conversation Turn:
User: {user_utterance}
System: {system_response}

Status Definitions:

- `OPEN`: mentioned, no action taken
- `PENDING`: system processing or requesting information
- `COMPLETED`: successfully achieved
- `FAILED`: explicitly failed
- `ABANDONED`: cancelled by the user

Transition Examples:

- `"book a flight"` → OPEN
- `"which dates?"` → PENDING
- `"flight booked"` → COMPLETED
- `"no flights available"` → FAILED
- `"never mind"` → ABANDONED

Output format (JSON):
`{"status": "STATUS"}`

Goal Evolution Prompt

`Analyze relationships between a new goal and existing related goals.`

New Goal:
Content: {new_goal.content}
Core Content: {new_goal.core_content}

Related Goals (top-k by semantic similarity):
{related_goals_context}

Task: For each related goal, determine the relationship type:

- `link`: semantically related but independent
- dependency: new goal depends on the related goal
- none: no significant relationship

Output format (JSON):

```
{
  "goal_id_1": "relationship_type",
  "goal_id_2": "relationship_type"
}
```