ATOD

An Evaluation Framework and Benchmark for Agentic Task-Oriented Dialogue Systems

Yifei Zhang[1,2] Hooshang Nayyeri[1] Rinat Khaziev[1]
Emine Yilmaz[1,3] Gokhan Tur[1,4] Dilek Hakkani-Tür[1,4] Hari P Thadakamalla[1]

[1] Amazon [2] Emory University [3] University College London
[4] University of Illinois Urbana-Champaign

Work performed during Yifei Zhang's internship at Amazon.

1 Task and motivation

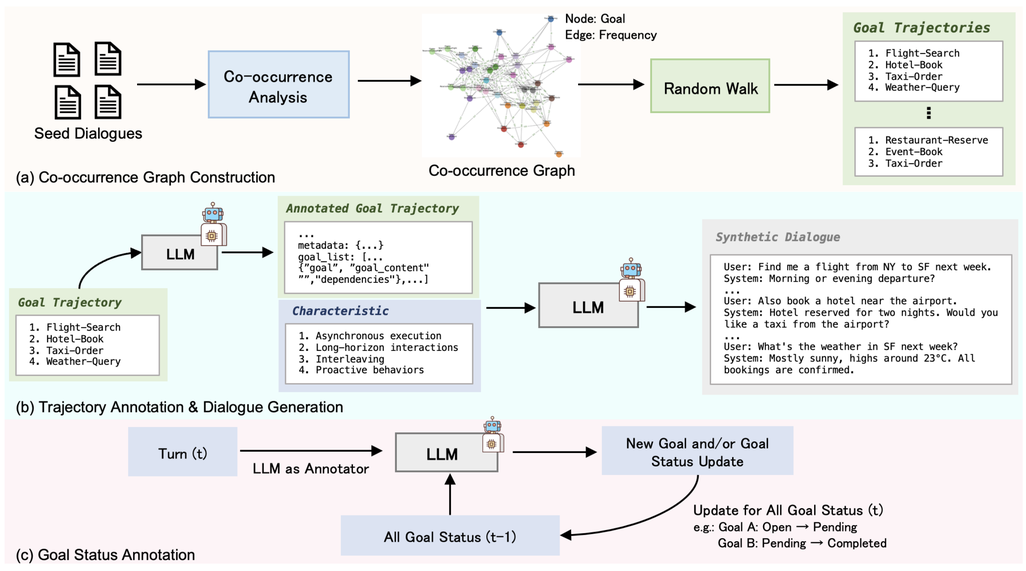
Agentic conversations contain goals that overlap, pause, depend on each other, and resume later. We recover these lifecycles from dialogue evidence.

Given	Recover
A fixed dialogue	Goal identities and descriptions
Each new utterance	States and dependency links
Completed dialogue	Final goal-state reconstruction

States: not mentioned, open, pending, completed, failed, abandoned. Online evaluation means incremental replay of a fixed dialogue.

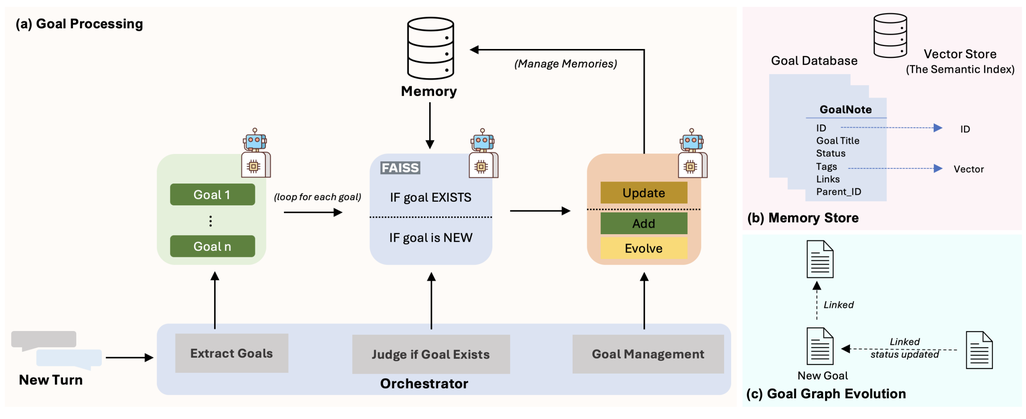
2 ATOD construction

1,000 dialogues: 428 medium / 572 complex



Six properties: concurrency, interleaving, dependencies, long-horizon memory, asynchrony, and proactivity. SGD-based trajectories, LLM generation and turn-level labels; verifier checks and author screening.

3 Agentic memory evaluator



Memory: symbolic goal states + vector retrieval; extract, match, update/insert, and audit. Scope: synthetic replay, point estimates, single-author audit.

4 ATOD-Eval metrics

Metric	Diagnostic target	Scale
dGCR	Completed / decided goals	0–1
NTC	Turns to a decided state	Turns
MRA	Useful, consistent recall	0–1
PE	Grounded AND beneficial	0–1
Relevance	Local response quality	0–1
Coherence	Dialogue consistency	1–5

Decided = completed or failed. Pending and abandoned goals are excluded. Interpret these diagnostics alongside detection F1 and conditional status accuracy.

5 Main results

Memory method	Med. F1 / Acc.	Cplx. F1 / Acc.
Mem0	76.80 / 75.60	61.10 / 71.20
MemoChat	80.38 / 73.88	58.07 / 66.83
MemoryBank	82.56 / 94.23	76.86 / 78.50
RAG	85.59 / 94.85	87.13 / 77.83
A-Mem	87.30 / 89.50	62.00 / 80.60
LLM-Rsum	93.83 / 89.47	89.13 / 69.95
Ours	**91.92 / 92.31**	**86.49 / 84.28**

Percent. All rows share Claude-Sonnet-4 and all-MiniLM-L6-v2. Accuracy is conditional on detected goals. Zero-shot evaluation on the same 1,000 dialogues. [Table 4]

Same-backbone comparison

Method	Med. F1 / Acc.	Cplx. F1 / Acc.
Prompt-only	78.95 / 93.26	75.58 / 84.26
Ours	**91.92 / 92.31**	**86.49 / 84.28**

Detection F1 gains: +12.97 points (medium), +10.91 (complex), with similar conditional accuracy. [Table 5]

6 Reliability and scope

Qwen3-32B judge swap: 90.10 / 90.60 (medium) and 83.10 / 85.30 (complex), F1 / accuracy. This checks our evaluator only; it does not rerank all baselines.

Paper: arxiv.org/abs/2601.11854

Code: github.com/amazon-science/ATOD

Data: huggingface.co/datasets/jimmyzhangyf/ATOD