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.

AACL-IJCNLP 2026 Paper 333-MAIN Virtual presentation

One conversation, several unfinished goals

User request	What the evaluator must retain
Book a flight and hotel.	Concurrent goals
Remind me to pack.	A deferred goal across many turns
What is the weather?	Bookings remain active during interruption
Book dinner after my flight.	Arrival time constrains the reservation

Six properties

Concurrency, interleaving, dependencies, long-horizon memory, asynchronous execution, and proactivity.

The task: reconstruct goal lifecycles

Input

A fixed dialogue revealed one utterance at a time

Recovered state

Goal identities, lifecycle states, and dependency links

Lifecycle states	Interpretation
Not mentioned / Open / Pending	Not yet introduced / active / in progress
Completed / Failed / Abandoned	Achieved / unsuccessful / cancelled

Offline: final state **Online: incremental replay of the same fixed dialogue**

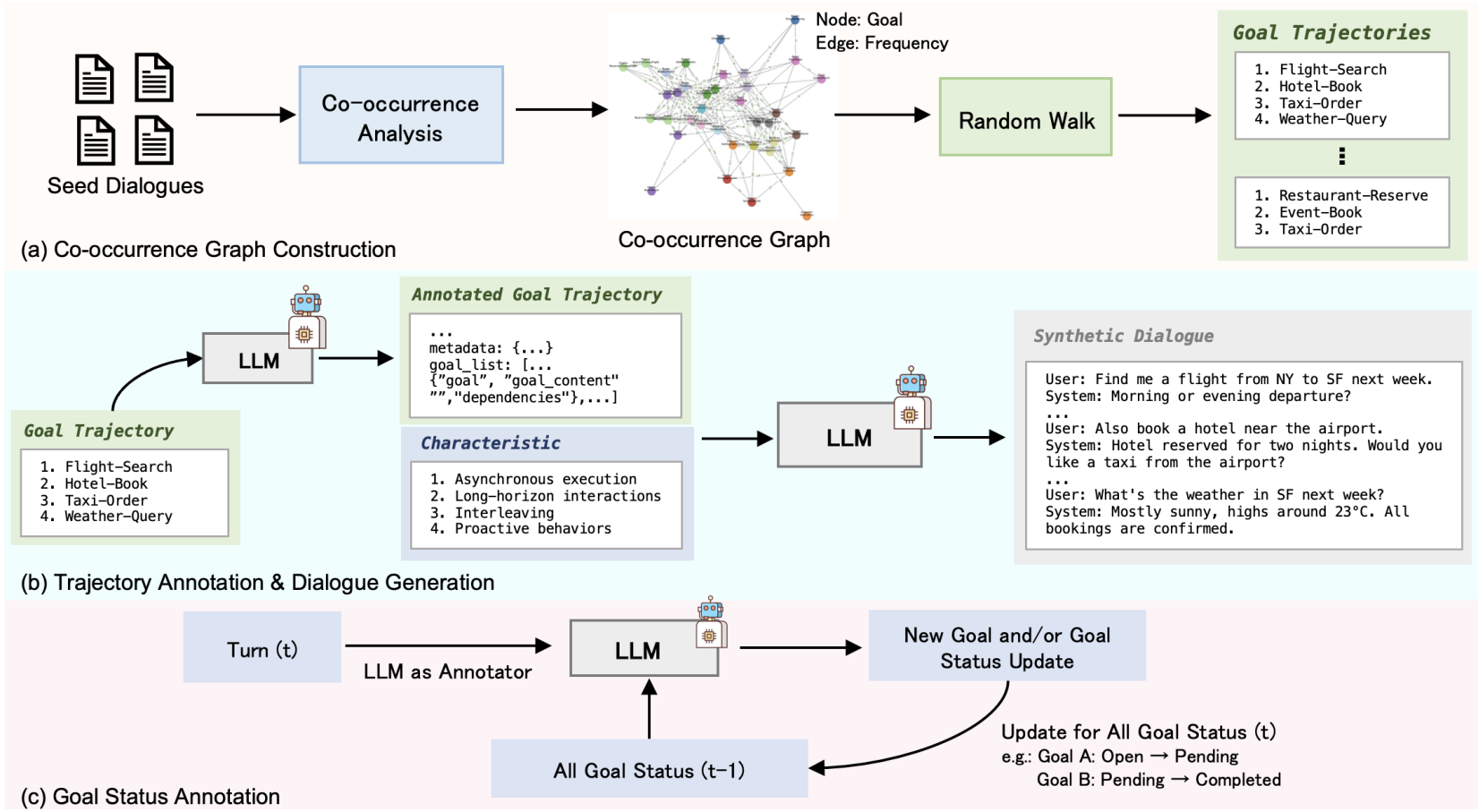
ATOD: construction and annotations

1,000

synthetic dialogues

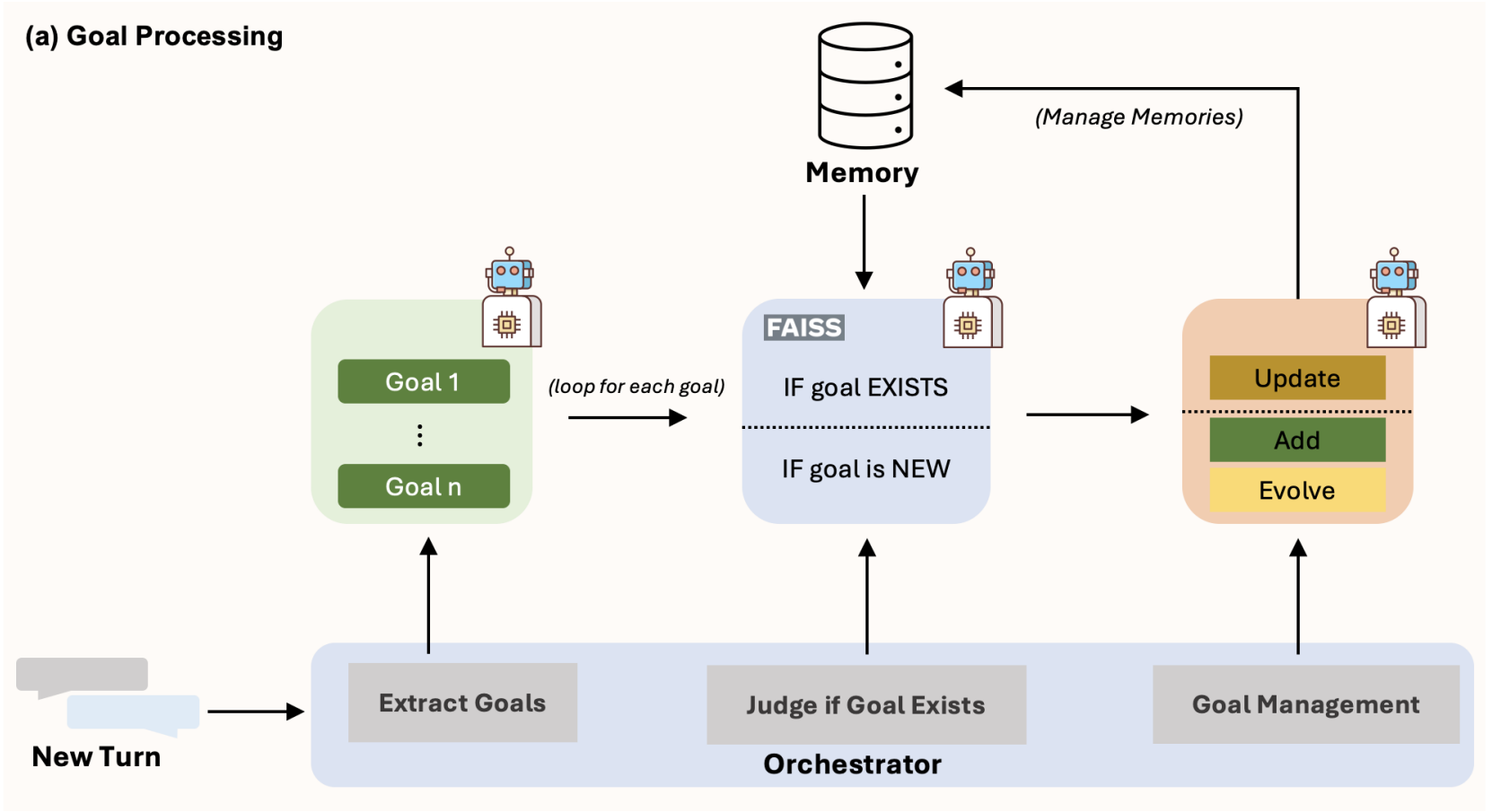
428 medium
572 complex

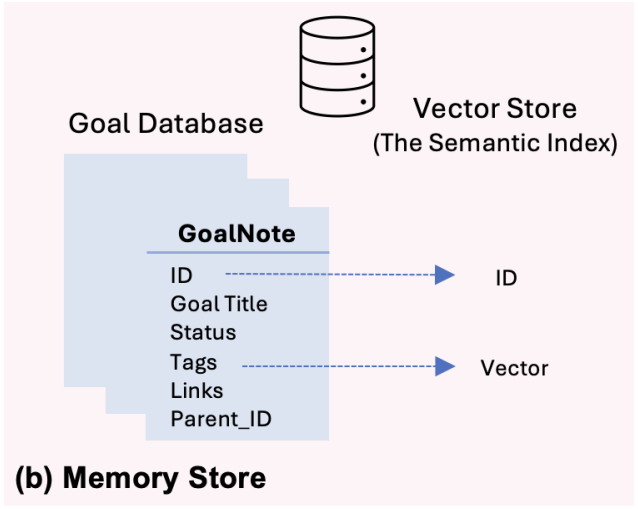
Goal identities
Dependencies
Turn-level states

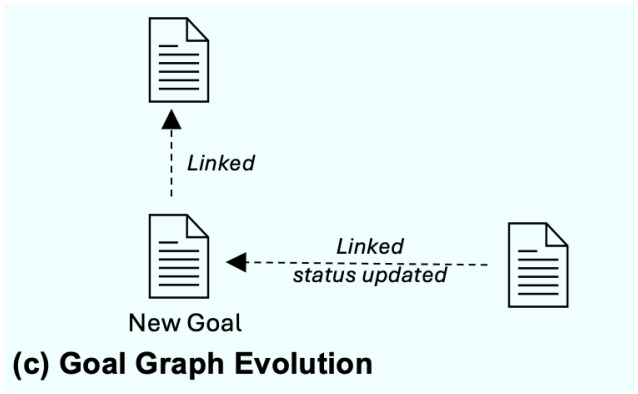


Original paper figure. Generation and annotation use Claude-Sonnet-4.

A symbolic–vector memory evaluator







Extract candidates / Match / Update or insert / Audit active goals

ATOD-Eval: complementary diagnostic metrics

Metric	What it measures	Scale
dGCR	Completed / (completed + failed) goals	0–1
NTC	Turns from initiation to a decided state	Turns
MRA	Sufficient, consistent retrieved memory	0–1
PE	Grounded AND beneficial proactivity	0–1
Relevance / Coherence	Local relevance / global consistency	0–1 / 1–5

Pending and abandoned goals do not enter the decided-goal denominator.

Experimental protocol

Memory comparison

6 baselines + our evaluator
Shared Claude-Sonnet-4 backbone
Shared all-MiniLM-L6-v2 embeddings
Zero-shot on the same test dialogues

Two primary outcomes

Goal detection F1
Which goals did the system recover?

Conditional status accuracy
Are detected goals in the right state?

High conditional accuracy can coexist with missed goals.

Temperature 0 and seed 42; point estimates, without significance tests.

Results: detection and status tracking together

Method	Medium F1	Medium Acc.	Complex F1	Complex 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**

Medium: above 90% on both axes.
Complex: highest conditional status accuracy in the memory block.

Controlling for the backbone model

Same backbone	Medium F1	Medium Acc.	Complex F1	Complex Acc.
Prompt-only	78.95	93.26	75.58	84.26
Memory evaluator	**91.92**	**92.31**	**86.49**	**84.28**
Difference (points)	+12.97	−0.95	+10.91	+0.02

The largest gains are in goal detection.

This comparison evaluates the complete memory system, not its individual components.

Why completion alone is insufficient

Goal	Final state	Decided?
g_1	Completed	Yes
g_2	Failed	Yes
g_3 depends on g_2	Pending	No

dGCR = 1 / 2

Read with goal detection and conditional status accuracy.

Equal completion can hide differences in memory recall, turn efficiency, and beneficial proactivity.

Reliability checks and remaining limits

Judge for our evaluator	Medium F1 / Acc.	Complex F1 / Acc.
Default	91.92 / 92.31	86.49 / 84.28
Qwen3-32B	90.10 / 90.60	83.10 / 85.30

Judge swap checks our evaluator only; it does not rerank all baselines.

Synthetic data and shared model-family biases
Single-author audit, without independent multi-annotator validation
Fixed-dialogue replay; interactive agent success remains untested

Goal lifecycles make agent evaluation more explicit

ATOD	Turn-level goals and dependencies in 1,000 synthetic dialogues
ATOD-Eval	Diagnostic metrics and an explicit memory evaluator

Paper: arxiv.org/abs/2601.11854

Code: github.com/amazon-science/ATOD

Data: huggingface.co/datasets/jimmyzhangyf/ATOD