PlugMem

A Task-Agnostic Plugin Memory Module for LLM Agents

Ke Yang*, Zixi Chen*, Xuan He*, Jize Jiang*,
Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, ChengXiang Zhai

UIUC · Tsinghua University · Microsoft Research

ICML 2026

The bottleneck has shifted.

2023

Prompt engineering

2026

Memory

Context windows grew from 8K to 10M tokens.
Agents still forget.

Today's memory systems force a bad trade-off

TASK-AGNOSTIC

Store everything, retrieve by similarity

- Verbose, redundant
- Signal buried in noise
- Generic but ineffective

TASK-SPECIFIC

Hand-craft memory for one benchmark

- Wins on its own task
- Breaks when transferred

Neither scales well.

Cognitive science gives us a clue

“Humans don't replay past experience — we distill it.”

Knowing that
semantic / propositional

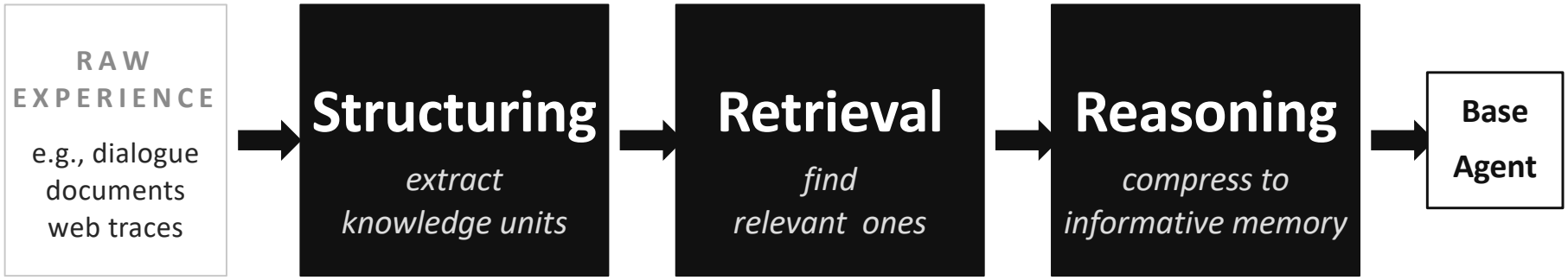
“My friend is allergic to peanuts.”

Knowing how
procedural / prescriptive

“To check out: search → filter → confirm.”

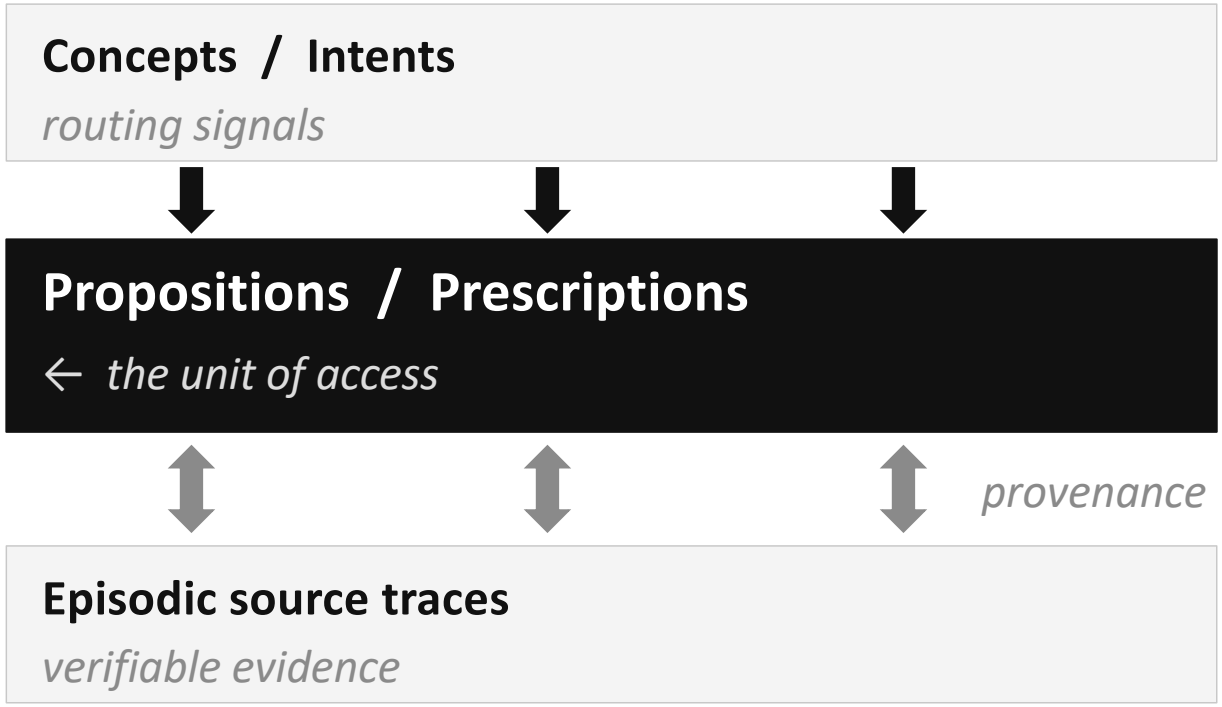
Both must coexist for any non-trivial task.

PlugMem: three modules, one memory backbone



Same module. Three very different tasks. No redesign.

Knowledge as the unit of memory



Not entities.

Not text chunks.

Knowledge.

One module, three benchmarks, every category beaten

Benchmark	Task type	Strong baseline	PlugMem
LongMemEval	Long-horizon dialogue	LiCoMemory 74.1%	**82.8%**
HotpotQA	Multi-hop knowledge QA	HippoRAG2 73.3 F1	**74.1 F1**
WebArena	Web navigation	AgentOccam 42.1% SR	**52.6% SR**

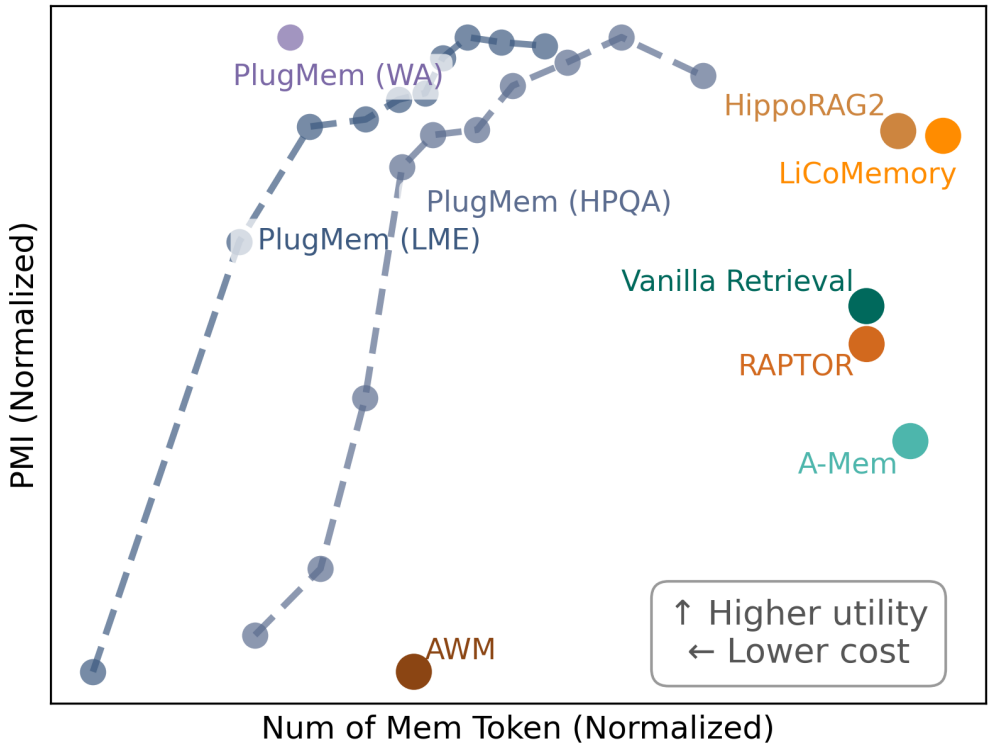
** SR on WebArena is reported on shopping tasks. More to find in paper.*

25×
fewer memory tokens
LongMemEval

7×
fewer memory tokens
HotpotQA

29×
fewer memory tokens
WebArena

Higher utility, lower agent-side cost — across the board



READ THIS PLOT

X memory cost
less is better

Y decision utility
more is better

PlugMem sits

upper-left.

Across all three benchmarks.

Information density = bits of decision-relevant info per memory token — a task-agnostic metric we introduce.

Knowledge is the right unit of memory.

Episodes are the evidence beneath it.

Same module — long dialogue, multi-hop QA, web agents.

PLUGMEM PAPER

EXTENDING IN THE PAPER

- Information-theoretic measure of memory utility and efficiency
- Memory update & decay operations
- Knowledge transfer mitigating cold-start problem
- Layering task-specific heuristics on top

PLUGMEM REPO

CODE · DATA · UPDATES

`github.com/`

`TIMAN-group/PlugMem`

Support Claude Code and OpenClaw.

Thanks! Happy to chat about extending PlugMem to your agent setting.