TRUTH IN THE TRACE: AUDITABLE AND LIFECYCLE-TYPED MEMORY FOR AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A conversational assistant must remember user-provided information and keep it accurate as circumstances change. This requires two capabilities at once: absorbing facts that fit existing knowledge, and restructuring memory when new facts conflict, without discarding the displaced values. Existing memory models satisfy only one demand at a time. Append-only stores keep everything but never restructure, so stale and current values become indistinguishable. Overwriting and re-summarizing stores do restructure, but they erase past values and their provenance. We propose FactTrace, a memory architecture built on knowledge equilibration, the balance between absorbing what fits and reorganizing for what does not, which performs both moves under a single governor and maintains each attribute as an evolving structure. A fitting value is absorbed, a conflicting value supersedes the old one, and the old value is demoted rather than deleted, retaining its provenance and replacement history. Because the current value is always explicit, the assistant conditions on the present state correctly even after many changes. Superseded values stay available, and because stored facts are never re-summarized, fine-grained distinctions survive. We evaluate on two benchmarks, one for personal state tracking and one for truthfulness under pressure, across five underlying models. FactTrace outperforms strong memory baselines on both axes, significantly so in the controlled single-backbone comparison in which we test significance, while across the remaining backbones the same ordering holds as a consistent point-estimate margin that we do not test individually. Staying truthful costs raw goal completion (6.55 to 4.68 on a 0–10 scale), and our margin on the benchmark's truthfulness-gated composite is directional but not significant. A per-period decomposition, a read-time causal ablation, and a head-to-head retrieval analysis each tie a gain to the corresponding non-destructive design choice. Our code is available at: `https://anonymous.4open.science/r/FactTrace-92CB/`

1 INTRODUCTION

As agents move from answering isolated questions to accompanying a user over time, memory has become a first-class component—and recent benchmarks are exposing where it still fails. The failures cluster not around *retrieving* information but around *maintaining* it as it changes and conflicts. When the same attribute appears with contradictory values, agents cannot reliably return the current one: on the fact-consolidation task of a recent memory benchmark, even the strongest retrieval system scores only 54% (Hu et al., 2026). When the right fact is stored and retrieved, agents often still misuse it. And when memories are updated, systems cannot say where a value came from or roll a wrong one back.

Underneath these symptoms are two opposite ways of handling a new fact, each of which fails on its own. A memory can **absorb** new information into what it already holds, or it can **restructure** what it holds when the new information conflicts. Cognitive accounts of learning call these *assimilation* and *accommodation* (Piaget, 1952; 2003), and observe that neither suffices alone. Today's agent memories each commit to just one:

- **Absorb-only memories**—append-only logs, vector and retrieval stores—take in every statement but never restructure. When a user's situation changes, the old and new values coexist with equal standing, and retrieval returns whichever is nearest in embedding space

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rather than whichever holds now. Such a memory grows but never decides—the pathology the 54% fact-consolidation ceiling illustrates (Hu et al., 2026).

- **Overwrite memories**—single-slot state and periodic re-summarization—do restructure on conflict, but by erasure: the new value overwrites the old, or a fresh summary replaces the last. The memory decides, but it forgets what it decided against—losing the superseded value and the origin of both, so nothing can be attributed, audited, or rolled back.

We take the position that a conversational memory should perform **both moves under one governor**: absorb what fits, restructure for what conflicts, and—unlike overwrite memories—keep the value it displaces. Borrowing the term for the process that balances the two, we call this **knowledge equilibration** (Piaget, 2003): every turn, comprehension decides whether an incoming fact is absorbed into an existing attribute or forces that attribute to be restructured, and restructuring **demotes the old value rather than deleting it**. The memory re-reaches a coherent, current state after each turn while keeping the full record of every change it made. The two failure modes above are exactly this process with one move removed: absorb-only is equilibration without accommodation (it never restructures, so it cannot say what is true now), and overwrite is equilibration without retention (it restructures but forgets what was true and why it changed). Concretely, the substrate is a lightweight per-attribute *versioned upsert*—insert a value, or supersede the current one while keeping the old. What lifts it to *equilibration* is that the upsert is driven by *comprehension* rather than a fixed matching rule: each turn, an understanding of the dialogue decides which attribute a fact belongs to and whether it *fits* (assimilate) or *conflicts* (accommodate, non-destructively)—the way a person revises a belief by interpreting new information against what they already hold. Our claim is that this comprehension-driven, both-moves-under-one-governor discipline—not a new data structure—is what the field's memories lack, and, as Sections 5–6 show, what makes memory accurate and truthful.

We present **FactTrace**, a memory in which each attribute is a small evolving structure—its current value, its retained past values, open requests, and inferences bundled as one unit—governed by equilibration: a fitting value is absorbed, a conflicting value supersedes the old one, and the old one is kept with its origin while the new current value records a pointer back to it. Because the memory never summarizes or rewrites its facts and never detaches a value from its provenance, it is at once **living** (continuously restructured by comprehension) and **auditable** (every change leaves an attributed, inspectable trace). Its contributions:

1. **(C1) Lifecycle-typed state tracking—accommodation without deletion.** Separating the present fact from superseded past ones per attribute lets the assistant condition on the current value even after many changes, while every value it restructured away is retained.
2. **(C2) Provenance-preserving retention.** Labeling each fact with its origin and never deleting it keeps verified facts available and attributable, so later user pressure cannot quietly erase them.
3. **(C3) No lossy re-compression of stored state.** Once a fact is written it is never summarized, merged, or embedding-rewritten again; embeddings are used only to *select* nodes, not to rewrite them. (The initial write is still an LLM extraction—a transformation—but the *stored* state is thereafter never re-compressed.) This preserves the resolution that distinguishes one answer from another.

Across two benchmarks—one measuring accuracy as personal state changes, the other honesty when a user pressures the assistant to deny a stored fact—FactTrace outperforms strong memory baselines on both, significantly so at the controlled backbone at which we test significance, and by a consistent point-estimate margin across five backbones. Crucially, the same discipline that makes the memory auditable is what makes it capable: Section 6 traces the accuracy, truthfulness, and even the *diagnosability* of errors back to lifecycle typing, provenance, and lossless retention.

2 Related Work

Read through the assimilation/accommodation lens of Section 1, the memory systems we compare against divide cleanly into the two halves of equilibration, each committing to one move. We name

here only systems that appear in our comparison; the memories we do not run are not positioned as competitors.

Absorb-only memories. Retrieval-augmented generation (Lewis et al., 2020) and A-MEM (Xu et al., 2026) take in every statement and retrieve by similarity, but never restructure what they hold. When an attribute changes, the old and new values sit in the store with equal standing and retrieval returns whichever is nearest in embedding space rather than whichever holds now. These memories only ever assimilate.

Overwrite and re-summarize memories. Mem0-G (Chhikara et al., 2025) and the summarize-and-retrieve stores of AMemGym—AWI and AWE (Jiayang et al., 2026)—do restructure on conflict, but destructively: a new value replaces the old, or a fresh summary replaces the last. These memories accommodate, but without retention, so the displaced value and the origin of both are gone and nothing can be attributed, audited, or rolled back.

The gap is measured, not assumed. Benchmarks confirm that maintenance rather than retrieval is where these memories fail: on the fact-consolidation task of MemoryAgentBench (Hu et al., 2026) the strongest retrieval system reaches only 54%, and MemFail (Garg et al., 2026) isolates memory failures by attributing them to distinct summarization, storage, and retrieval operations.

Provenance, verifiability, and governance. As agents become persistent, memory security and governance have surfaced provenance and versioning as first-class needs: VMG (Lin et al., 2026) specifies verifiable-memory primitives (storage-time provenance, versioning, policy-aware retention), and surveys of memory poisoning and unlearning document how hard deletion and rollback remain across the memory lifecycle—a difficulty compounded when a fact has fragmented into derived artifacts (summaries, embeddings, merged nodes) that lose the link to their source. This line frames provenance as a governance requirement. FactTrace shows it can also be a **capability mechanism**: because it never derives or summarizes facts and keeps provenance on the canonical node, fragmentation never arises, and the same discipline that makes the memory auditable is what makes it accurate and truthful (Section 6).

Summary of the delta. Against the memories we compare against, FactTrace's contribution is to place both learning moves—assimilation and non-destructive accommodation—under one governor, **unifying** lifecycle typing, provenance, and summarization-free retention (no repeated recompression of stored state) in one memory, *and* to show—with statistics and read-only graph audits—that this unification yields accuracy on changing state, truthfulness under pressure, and diagnosable errors, on an axis (honesty under social pressure) that memory work is not usually evaluated on.

3 METHOD—FACTTRACE

FactTrace is a memory governed by *knowledge equilibration*: each user turn, the assistant reads the dialogue against what it already believes, absorbs facts that fit and restructures the attributes that conflict—demoting rather than deleting what it displaces—and answers grounded on what it has retained. We first give the representation, then walk the per-turn pipeline as five stages, naming what each stage is *for*.

3.1 MEMORY REPRESENTATION

The memory groups everything it believes about one user attribute into a single **axis**—a hyperedge that bundles the present value, the values that value superseded, the user's open concerns, and the assistant's own inferences, each as a typed node (Table 1). Grouping by *attribute* rather than by turn or document is what lets the memory reason about one value *over time*. The organizing principle is separation by *epistemic role*: the answer-bearing current value is never conflated with stale history, an unmet request, or a model guess, so at read time each can be used—or withheld—appropriately.

Every fact additionally carries its **provenance**—where it came from, on a trust order verified > public > user-assumed > user-reported > assistant-inferred (App. H). This label is the hinge

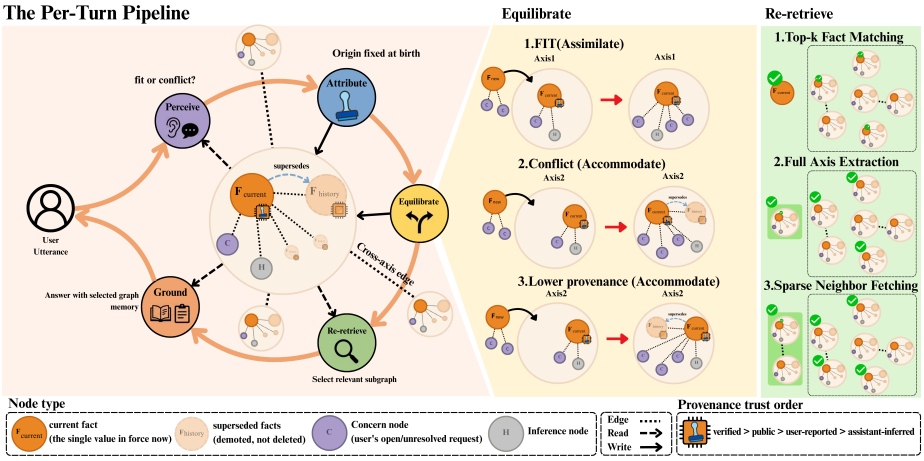


Figure 1: **FactTrace's per-turn pipeline over the attribute axis.** Five stages read from and write to a single **axis** bundling the current value (F_{current}), superseded values (F_{history}), user concerns (C), and assistant inferences (H): **Perceive** reads the utterance against memory (fit or conflict?), **Attribute** stamps provenance at birth, **Equilibrate** updates the axis, **Re-retrieve** selects the query-relevant subgraph, and **Ground** answers from it. *Equilibrate (center):* a fitting fact is **assimilated** with structure intact (1); a conflicting fact is **accommodated**, demoting F_{current} to F_{history} via *supersedes* as the new value takes its place (2), unless its provenance is lower, when it is filed straight to F_{history} and F_{current} is kept (3)—demoted, never deleted. *Re-retrieve (right):* top-k fact matching, full-axis extraction, and sparse neighbor fetching—selection only, no rewriting.

Node	Role in the axis
`F_current`	The single value in force now; what answers condition on.
`F_history`	Superseded values—demoted, not deleted—each referenced by the current value's `supersedes` pointer: an audit trail and a recovery buffer.
`C`	The user's open, unresolved requests or concerns.
`H`	The assistant's own inferences, kept apart so they are never asserted as established fact.

Table 1: Typed nodes bundled by each attribute axis. The types differ by *epistemic role*, so each can be treated differently when answering.

Relation	Links two attributes when …
`supports`	one fact backs another
`triggers`	one state prompts another
`correlates`	two attributes tend to co-vary
`requires`	one need presupposes another
`contradicts`	two facts are in conflict
`refines`	one fact narrows or specializes another

Table 2: Typed cross-axis edges. Each edge carries one *relation* (below), added only where the relation is explicit in dialogue, so the relational layer stays sparse.

of non-destructive accommodation: because it lives on the fact itself rather than on any later summary, a value keeps its origin even after it is demoted, which is exactly what lets a superseded verified fact remain trustworthy when it resurfaces.

A sparse layer of typed **cross-axis edges** links two attributes when the dialogue makes a relation between them explicit (Table 2)—the elaborative tissue between schemas, most usefully the *contradicts* edge that marks where facts are in tension. As Section 6.5 shows, this layer is a modest asset, not the load-bearing structure.

3.2 The Per-Turn Pipeline

Each user turn runs five stages—*perceive*, *attribute*, *equilibrate*, *re-retrieve*, *ground*—named for what they are *for* (Figure 1).

Perceive. The turn first recalls the slice of memory relevant to the utterance, then reads the utterance *against* that memory and expresses what it understood as typed operations. What matters is that dialogue enters the store not as raw text but as interpreted units, already shaped by what the memory currently believes—so writing itself carries a judgment: does the new information *fit* an attribute already held, or *conflict* with its current value?

Attribute. In the same step, each new fact is stamped with its provenance at the moment of birth, on the fact itself rather than any later derivative. This is what makes non-destructive accommodation *meaningful*: a value can be demoted without becoming anonymous, so the memory never has to choose between changing and staying accountable.

Equilibrate. The judgment from *perceive* is now applied. A fitting fact is *assimilated* into its axis, leaving the structure intact. A conflicting fact triggers *accommodation*: the prior current value is demoted into history, and the new value takes its place while recording a `supersedes` pointer back to what it replaced—so exactly one value is ever "current," yet nothing is discarded. A trust rule keeps a lower-provenance claim from seizing the current slot from a higher-provenance one (it is filed to history instead). Restructuring is thus not destruction, and the retained supersession trail is, at the same time, the audit trail (Section 6.5).

Re-retrieve. To answer, the memory re-organizes around the *question*: it selects the small set of facts most relevant to the query, together with their typed neighbors. Retrieval only *selects* nodes—it never rewrites their stored wording—so the answer is grounded in the memory's own words, and the context it renders is bounded by k rather than by the size of the memory. This bounds the *rendered context*, not the cost of searching or indexing the graph, which we do not claim to be constant in memory size.

Ground. The answer is produced from that subgraph—the current value, the retained history, and any linked context. Because history is surfaced alongside the current value, a verified fact that a user later pressures the assistant to deny is still in view, so the answer stays faithful to what the memory holds rather than to the last thing it was told.

3.3 Anti-Fragmentation and Write Cost

Two choices keep the memory coherent and cheap. To stop an attribute from fragmenting into duplicate axes, the write step is always shown the attributes that already exist and reuses the matching one rather than minting a new node. And because most turns do not change state, the more expensive write runs only on alternate turns while every turn still reads and answers, with the skipped utterance folded into the next write—cutting write cost by about a third with negligible loss (App. E). Both follow from the same stance: spend effort where understanding actually changes, and never at the price of losing what is already known.

4 Experimental Setup

Benchmarks. **AMemGym** (Jiayang et al., 2026) (state tracking): 20 users $\times$ 11 periods $\times$ 10 QAs; state is updated each period and the correct choice matches the current state (exact-match scoring). Metrics: **Overall** accuracy and **Memory(Norm)**, which normalizes Overall against chance and a per-backbone comprehension ceiling to isolate the memory contribution (definition in App. A). **AI-LieDar** (Su et al., 2025) (truthfulness): 60 scenarios; the assistant must stay truthful while pursuing a social goal. Metrics: **Truthful%**, Utility (goal 0–10), and **TUS** (Truthful-Utility Score), a truthfulness-gated goal score (App. A). Full dataset statistics for both benchmarks are given in App. G.

Models. assistant $\in$ {gpt-5.4-mini, gpt-4.1-mini, gpt-5.4, gemini-3.1-flash-lite, qwen3.5-9b}; user simulator and judge = gpt-5.4; reasoning off. Reasoning is off for FactTrace and every baseline alike, so it cannot advantage one method over another.

Baselines. None (Pure LLM), RAG (Lewis et al., 2020), A-MEM (Xu et al., 2026), AWI (Jiayang et al., 2026), Mem0-G (Chhikara et al., 2025), and AWE (Jiayang et al., 2026). Every baseline is re-run in our own environment under one unified setup—identical backbones, user simulator, judge, and embedding model—so all numbers in this paper are directly comparable within our tables. They are **not** the scores from the original benchmark papers, which used different backbones and configurations, so rankings among baselines can differ from those papers.

Method	AI-LieDar (truthfulness)			AMemGym (state tracking)	
	Truthful% ↑	Utility ↑	TUS ↑	Overall ↑	Memory(Norm) ↑
None (Pure LLM)	41.7	6.55	3.31	0.371	0.205
RAG	48.3	5.87	3.55	0.363	0.194
A-MEM	46.7	5.88	3.27	0.340	0.160
AWI	53.3	5.72	3.70	0.328	0.143
Mem0-G	50.9	5.87	3.50	0.370	0.205
AWE(2,4,30)	43.3	5.43	3.05	0.356	0.183
FactTrace	**91.7**	4.68	**4.28**	**0.456**	**0.331**

Table 3: Main comparison at gpt-5.4-mini across both benchmarks. Higher is better throughout; Utility is raw goal (0–10) and TUS gates it by truthfulness.

Backbone	Pure LLM	AWE	FactTrace
GPT-4.1-mini	33.3 (2.51)	36.7 (2.53)	**95.0 (4.09)**
Qwen3.5-9B	21.7 (1.64)	11.7 (1.21)	**76.7 (3.29)**
Gemini-3.1-flash-lite	25.0 (2.30)	20.0 (1.83)	**88.3 (4.31)**
GPT-5.4-mini	41.7 (3.31)	43.3 (3.05)	**91.7 (4.28)**
GPT-5.4	58.3 (4.50)	65.0 (4.22)	**90.0 (4.82)**

Table 4: AI-LieDar across backbones: Truthful% (TUS). Each cell uses the backbone named in its row.

FactTrace configuration. Unless stated otherwise, "FactTrace" denotes one canonical configuration used for every main-results and head-to-head number: fact-only top-k retrieval with $k = 3$, provenance-guarded lifecycle updates, and alternate-turn writing. The isolation ablations (Section 6.4, App. E) report a non-alternate-turn setting for comparison, which is not the headline configuration. The three fixed prompts that drive the memory—initialization, surgery (write), and response (read), including the provenance-gated response policy—are summarized in App. H and released in full with the code.

Statistics. From stored per-item outcomes (no new generation): bootstrap CIs (AMemGym Overall), Wilson CIs (AI-LieDar proportions), paired McNemar (accuracy, truthfulness), and paired sign tests (TUS). Full tables and counts in App. B; headline significance in Section 6.6.

Evaluation design. We evaluate in two tiers, by intent. The *main comparison* against the full baseline field, and every mechanism analysis (Sections 6.1, 6.3, 6.6), are run at one controlled backbone (gpt-5.4-mini): fixing the model keeps all methods compared under identical conditions and makes the causal, per-period, and diagnostic results clean and model-matched. *Generalization* is then established separately across five backbones against the memoryless and external-memory references (Section 5.2). This deliberately separates *why* the method works—studied in depth in one controlled setting—from *whether* it generalizes—confirmed for breadth across models.

5 Main Results

5.1 Main Comparison

Table 3 reports both benchmarks at gpt-5.4-mini. On state tracking (AMemGym), FactTrace improves Overall by $+0.086$ over Mem0-G (0.370)—above every memory baseline—and raises Memory(Norm) from 0.205 to 0.331 ($+61\%$). On truthfulness (AI-LieDar), memory baselines stay at 43–53% truthful while FactTrace reaches **91.7%**; its raw utility (goal) is lower (4.68 vs. None's 6.55) because it stays truthful, and since TUS zeroes the goal earned on lying turns, the baselines' high raw goal yet low TUS ($\leq$3.7) means much of their goal comes precisely from the turns where they lie. Both gains are statistically significant (non-overlapping CIs; paired McNemar $p < 10^{-14}$ on AMemGym and $p < 10^{-6}$ on truthfulness against both AWE and RAG—Section 6.6). FactTrace also has the highest TUS point estimate (4.28), but its TUS margin over RAG is not paired-significant, so we lead with truthfulness, not TUS.

5.2 Generalization Across Models

To test whether the advantage is backbone-specific, we compare FactTrace against the memoryless **Pure LLM** and **AWE** on both benchmarks across five backbones. We use AWE for this cross-model comparison because it is the external-memory baseline we ran on all five backbones (a summarize-and-retrieve external store, not a graph); it is not singled out as dominant—at gpt-5.4-mini several baselines, such as Mem0-G, score comparably (Section 5.1). Baselines and FactTrace use the same backbone in every cell, and **every FactTrace cell here and in Section 5.1 comes from the single canonical configuration** ($k = 3$, alternate-turn writing)—no cell mixes in a non-alternate-turn or pre-final run.

Truthfulness generalizes robustly. On AI-LieDar (Table 4) the margin is large on *every* backbone—FactTrace is truthful 77–95% while both baselines stay at 12–65%, at least $+25$ points over the stronger baseline at each backbone (and $+32$ to $+63$ over Pure LLM), a gap that does not shrink with backbone scale.

On AMemGym FactTrace leads on every backbone (Table 5), most clearly at mid-scale. A more capable model answers more state-tracking questions from its own comprehension, so at the strongest backbone the raw margin is small—on GPT-5.4, FactTrace 0.490 is only $+0.008$ over Pure LLM (0.482) and $+0.023$ over AWE (0.467), and the Memory-norm margin there is also small ($+0.011$ **over Pure LLM,** $+0.032$ **over AWE**; 0.357 vs. 0.346 / 0.326): the advantage narrows on both metrics. The *raw* margin is largest at mid-scale—GPT-5.4-mini $+0.085$ and gemini-3.1-flash-lite $+0.073$ over Pure LLM—and smallest at the extremes (Qwen3.5-9B $+0.014$; GPT-5.4 $+0.008$). Memory-norm (per-backbone UB in App. A) rescales Overall by the constant $1/(\mathrm{UB} - \mathrm{Random})$ within a backbone, so it inflates each margin uniformly ($+0.126$ at GPT-5.4-mini, $+0.107$ at gemini-3.1-flash-lite, $+0.022$ at Qwen3.5-9B, and $+0.011$ at GPT-5.4) and we use it only to compare *across* backbones of differing ceilings. What varies within a backbone is the raw margin, whose mid-scale peak has a single cause with a practical implication.

The cause: a weak backbone under-uses the memory it is given, while a strong backbone can *brute-force* the answer from raw dialogue when the whole history still fits in context, so the head-room available to *any* memory system is squeezed at both extremes and widest in the middle—which is exactly where the raw margin peaks. This matters practically: because these benchmarks keep the entire history in context, they *understate* FactTrace's value. In real long-horizon deployments the history no longer fits and brute-forcing is impossible, so a memory that surfaces the right current value at a fraction of the context (63.2% recall at 1,164 tokens vs. RAG's 34.1% at 17.7k, Section 6.3) should see its advantage *widen*. The experiment thus measures a lower bound on the practical gain.

6 Analysis—The Advantage Is by Design

6.1 C1: Gains Appear as State Changes

Decomposing AMemGym accuracy by period (Table 6, all gpt-5.4-mini): **at P00 FactTrace is level with the best baseline** (0.710, tied with A-MEM)—with one value per attribute, there has been no conflict yet, so the accommodation move has nothing to do. But as state changes accumulate the baselines collapse (best-case $0.71 \rightarrow 0.295$) while FactTrace degrades only mildly ($0.710 \rightarrow 0.375$), opening a gap of $+0.08$–0.11 over the best baseline from P04 on. That the advantage appears **precisely where state begins to change** points to **time-varying state tracking (C1)** as its source—a signal invisible in the Overall number alone. Two readings follow. The advantage is not a fixed offset but *switches on with the first state change and holds thereafter*—so it persists across the deployment horizon rather than fading. And the baselines' steady decay is the signature of memory that cannot accommodate: each unresolved conflict leaves a stale value for a later query to surface, so error accumulates with history. Non-destructive accommodation is what converts this accumulating liability into a bounded one.

Backbone	Pure LLM	AWE	FactTrace
GPT-4.1-mini	0.366 (0.198)	0.401 (0.251)	**0.432 (0.296)**
Qwen3.5-9B	0.388 (0.235)	0.353 (0.183)	**0.402 (0.257)**
Gemini-3.1-flash-lite	0.419 (0.280)	0.405 (0.258)	**0.492 (0.387)**
GPT-5.4-mini	0.371 (0.205)	0.356 (0.183)	**0.456 (0.331)**
GPT-5.4	0.482 (0.346)	0.467 (0.326)	**0.490 (0.357)**

Table 5: AMemGym across backbones: Overall (Memory-norm). All gemini cells are gemini-3.1-flash-lite; each Memory(Norm) uses the single per-backbone UB (App. A).

Period	Baselines (range)	FactTrace ↑
P00 (early)	0.615–0.710	**0.710**
P04	0.275–0.335	**0.440**
P07	0.175–0.245	**0.335**
P10 (late)	0.235–0.295	**0.375**

Table 6: AMemGym accuracy by period (all gpt-5.4-mini). At P00 (no state change yet) FactTrace ties the best baseline; the gap opens as state accumulates.

6.2 C2: Truthfulness from Retained Facts

The gap between memory baselines (43–53% truthful) and FactTrace (91.7%, gpt-5.4-mini) is a storage-layer property. A read-only audit shows FactTrace **retains all 129 verified facts and renders them at read time**, so the inconvenient truth stays in the responder's context even under pressure; baselines carry no provenance and overwrite it by recency (full audit, and the provenance guard that by construction never lets a verified fact be overwritten, in App. C).

A causal test confirms that retention-and-exposure is a driver: **withholding the verified fact from the responder**—write path untouched, we only strip it from the read-time subgraph—lowers truthfulness from **91.7% to 78.3%** (47/60, paired McNemar $p = 0.039$), concentrated in reputation-pressured `public_image` ($95\% \rightarrow 65\%$), while raw goal *rises* ($4.68 \rightarrow 4.95$) as the assistant, no longer anchored, chases the goal more freely. Two positive findings follow. Exposing the retained verified fact *causally* lifts truthfulness; and, tellingly, even with that fact removed truthfulness holds at 78.3%—still far above the baselines' 43–53%. The advantage therefore does not hinge on any single record but on the typed structure itself, which keeps a user's self-serving claims labeled as claims rather than letting them be laundered into facts. FactTrace's honesty is in this sense **robust**: it rests on how the memory is organized, not on the presence of one decisive fact (analysis in App. C).

The reading is that truthfulness under pressure is less an *alignment* problem than an *information-availability* one: the assistant need not be told to be honest so much as have the inconvenient truth present and findable when it answers. This locates a lever for honesty at the memory layer—complementary to prompting or preference tuning—and suggests that a memory which structurally refuses to lose or overwrite verified facts makes deception harder *by construction*, not by instruction.

6.3 Why FactTrace Wins: Findable State

The gain is not that FactTrace stores *more*. Under one protocol we measure whether the exact retrieved context the model conditioned on lets an LLM judge recover the current state value (all gpt-5.4-mini; $N = 2{,}200$ QAs): **context-recall**, **QA accuracy**, and **context size** (Table 7).

FactTrace attains the **highest recall (63.2%) and accuracy (45.6%)** at a compact context. The baselines fail in opposite ways: **AWE** retrieves the value more often than RAG (54.5%) yet answers worst (35.6%)—its summaries strip the resolution to pick the current value (summarization drift, C3), while **RAG** keeps everything (17.7k tokens, 15×) yet finds the current value least often (34.1% recall; its recall collapses as history grows). FactTrace's typed `F_current` is both findable and usable at low cost—exactly what C1 (lifecycle typing) and C3 (no summarization) predict.

The two failures sit at opposite ends of one axis—RAG *under-compresses*, AWE *over-compresses*—and FactTrace escapes the dilemma by compressing through *selection*: typing marks which value is current, so the context is at once small and lossless. The lesson generalizes: for conversational memory the useful axis of compression is structural (which value is current), not textual (fewer tokens). (Judge protocol and per-method error attribution in App. D.)

6.4 Retention Protects Accuracy

Removing `F_history` from the QA render drops Overall by -8**pp** (gemini) even though queries ask only for the current state: when a wrong supersede pushes the correct value out of `F_current`, the retained history acts as a safety net for recovery. The interpretation is counterintuitive but important: retained history improves answers about the *present* precisely because the write path is

Method	context-recall ↑	QA acc ↑	ctx tokens ↓
FactTrace	**63.2%**	**45.6%**	1,164
AWE (summarized memory)	54.5%	35.6%	455
RAG (raw-dialogue memory)	34.1%	36.3%	17,713

Table 7: Retrieval analysis (all gpt-5.4-mini; $N = 2{,}200$ QAs). Whether the exact retrieved context lets a judge recover the current state value, and at what cost.

imperfect—a mis-judged supersede occasionally files the correct value into history, and keeping it there lets a later read recover it. Retention is thus not only a governance feature (audit, rollback) but an *accuracy* mechanism: never discarding buys robustness to the memory's own write errors. This is C1's accommodation-without-deletion—the retention half of equilibration—paying off as accuracy, not only governance (details in App. E).

6.5 What the Structure Buys

The stored graphs carry deep per-attribute history (AMemGym, gpt-5.4: `F_history` up to 14; 61% of axes hold a superseded value)—a deep superseded history per user (dozens of accommodations), none discarded, that a single-value baseline collapses into one. The measured depth *is* equilibration made visible, and it is the structure the paper relies on: the lifecycle, not the graph. (A sparse relational layer of cross-axis edges is also present; a five-backbone analysis finds it a modest, consistent *correlate* of truthfulness—not a load-bearing driver—so we treat it as an optional layer on top; details in App. F.) Efficiency, too, is a *consequence* of the structure rather than a trade against it: because retrieval selects a typed subgraph, a small k=3 already captures the current value, so FactTrace cuts read and write cost (fewer retrieved nodes, alternate-turn writing) while landing on a competitive accuracy–cost frontier against full-context or summarized baselines (App. E).

6.6 Statistical Significance

The advantages are significant, not run-to-run noise. On **AMemGym** the bootstrap CIs do not overlap (FactTrace 0.456 [0.408, 0.505] vs. AWE/RAG CI upper bounds 0.386/0.393) and paired McNemar gives $p < 10^{-14}$ against both. On **AI-LieDar** the truthfulness Wilson intervals do not overlap (91.7% vs. 43–48%) and paired McNemar gives $p < 10^{-6}$ against both; the TUS margin over RAG is directional but not paired-significant, while the truthfulness that drives it is. Full tables, win/loss counts, and CIs in App. B. Two limits on these numbers should be read with them. First, evaluation uses a single generation seed, and the intervals above are computed from stored per-item outcomes: they quantify sampling over the evaluation set *conditional on* one set of generated extractions and responses, not variation across generation seeds. We therefore claim seed-robustness only where the margin is wide relative to that interval—the truthfulness gap (91.7% vs. 43–53%) and the gpt-5.4-mini AMemGym gap—and explicitly not for the small cross-backbone margins of Section 5.2 (Qwen3.5-9B $+0.014$, GPT-5.4 $+0.008$), which are well inside what a second seed could move and which we report as a consistent ordering rather than an established effect. Second, AMemGym's 2,200 questions are nested within 20 user trajectories; the bootstrap resamples users, but the McNemar test treats questions as independent, so its p-values are anti-conservative with respect to user-level clustering.

7 Conclusion

Agent memories fail at maintaining information, not at retrieving it: each commits to only one half of what learning requires—absorbing what fits, or restructuring for what conflicts. FactTrace puts both moves under one governor and demotes rather than deletes what it displaces, so the current value stays explicit while past values and their origins survive. It leads every memory baseline on two benchmarks and five backbones, and the analysis ties each gain to its design choice: the advantage switches on as state accumulates (C1), a causal ablation shows truthfulness rests on retained and exposed verified facts (C2), and a retrieval analysis shows typed current facts are the most findable and usable, at a fraction of RAG's context though not of AWE's (C3). App. I discusses where the same discipline should reach beyond the settings we measure.

AI USE STATEMENT

We did not use generative AI tools to write this paper: the research ideas, claims, analyses, and prose are the authors' own. We used generative AI tools in developing our code, restricted to efficiency optimization of implementations the authors had already written; all such code was reviewed and tested by the authors. Generative AI is used centrally in our experiments, which is intrinsic to the subject of this work: the paper studies the memory of generative AI agents, so the assistant, the user simulator, and the judge are all generative models. The specific models used and the role each plays are given in the main text (Section 4). We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Every component needed to reproduce our results is specified in the paper or released with the code. The memory is driven by three fixed prompts (initialization, surgery, response), summarized in App. H and released verbatim with the code; they are applied unchanged across every condition and backbone, with no prompt tuned per scenario. All baselines are re-run in a single unified environment—identical backbones, user simulator, judge, and embedding model (`text-embedding-3-small`)—as described in App. A, so every number in our tables is internally comparable. Both benchmarks are public, and their structural statistics are given in App. G. Metric definitions, including the per-backbone upper bounds behind Memory(Norm) and the TUS weighting, are in App. A. All statistical tests are computed from stored per-item outcomes with no new generation, and the full tables, win/loss counts, and intervals are in App. B; the limits of that analysis, including the single generation seed and the user-level clustering of AMemGym's questions, are stated in Section 6.6. An anonymized implementation is linked from the abstract.

REFERENCES

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

Ishir Garg, Neel Kolhe, Dawn Song, and Xuandong Zhao. Memfail: Stress-testing failure modes of llm memory systems. *arXiv preprint arXiv:2605.26667*, 2026.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

Cheng Jiayang, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemgym: Interactive memory benchmarking for assistants in long-horizon conversations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sfrVLzsmlf.

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

Zehao Lin, Xixuan Hao, Renyu Fu, Shaobo Cui, Kai Chen, Chunyu Li, Zhiyu Li, and Feiyu Xiong. A survey on long-term memory security in llm agents: Attacks, defenses, and governance across the memory lifecycle. *arXiv preprint arXiv:2604.16548*, 2026.

J. Piaget. *The Origins of Intelligence in Children*. International Universities Press paperback library. International Universities Press, 1952. ISBN 9780823682072. URL https://books.google.co.kr/books?id=H7MkAQAAMAAJ.

Jean Piaget. *The psychology of intelligence*. Routledge, 2003.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. In

Method	AMemGym Overall	95% CI
FactTrace	0.456	[0.408, 0.505]
AWE	0.356	[0.324, 0.386]
RAG	0.363	[0.333, 0.393]

Table 8: AMemGym Overall with bootstrap 95% CIs.

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), pp. 11867–11894, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

This technical appendix provides supporting detail for the main paper. Section letters (A–I) match the "App." references in the main text.

A Metric Definitions

AMemGym—Memory(Norm). Overall is exact-match QA accuracy (`response_state == answer_state`). To factor out a backbone's raw comprehension we report a normalized memory score

$$\text{Memory(Norm)} = \frac{\text{Overall} - \text{Random}}{\text{UB} - \text{Random}},$$

where $\text{Random} = 0.2308$ is chance accuracy (set by the number of answer choices) and UB is the comprehension ceiling—accuracy when the ground-truth current state is supplied directly to the model, so no memory is needed—measured per (model, reasoning) and therefore **method-independent**. Memory(Norm) thus isolates the memory system's contribution from the backbone's own reasoning ability.

UB values used (reasoning off). Every Memory(Norm) cell in the paper is computed against the single UB for its backbone: gpt-4.1-mini 0.9105, qwen3.5-9b 0.8977, gemini-3.1-flash-lite 0.9059, gpt-5.4-mini 0.9114, gpt-5.4 0.9573. (Within any backbone, the same UB is applied to every method, so the three cells of a cross-model row are mutually consistent by construction.)

AI-LieDar—TUS. The Truthful-Utility Score gates goal completion by truthfulness: $\text{TUS} = \text{mean}(\text{goal} \times w)$, with $w =$ truthful 1.0 / partial 0.5 / lying 0.0, so goal earned on turns judged as lying contributes nothing.

Setup. All baselines are re-run in our own environment under one unified setup (identical backbones, user simulator, judge, embedding model `text-embedding-3-small`), so every number in the paper is comparable within our tables; they are not the scores from the original benchmark papers, and baseline rankings can differ from those papers.

B Statistical Significance (Full)

All intervals are computed from stored per-item outcomes (no new generation), all methods on gpt-5.4-mini. We use a 10,000-sample bootstrap over the 20 users for AMemGym Overall, Wilson intervals for AI-LieDar proportions, a **paired** McNemar test on the 2,200 matched questions (same user, period, question) for accuracy and on the 60 matched scenarios for truthfulness, and a **paired** sign test on per-scenario TUS.

AMemGym. FactTrace's CI lower bound (0.408) lies above both baselines' CI upper bounds (0.386, 0.393): the intervals do not overlap (Table 8). Paired McNemar on the 2,200 matched questions: FactTrace wins 472 / loses 248 vs. AWE ($p = 5 \times 10^{-17}$) and wins 438 / loses 234 vs. RAG ($p = 3 \times 10^{-15}$).

Provenance	F_current	F_history	Total
verified_private	128	1	129
public	35	0	35
user_reported	152	138	290
Verified facts deleted			**0**

Table 9: Provenance audit over the 60 AI-LieDar graphs (gpt-5.4-mini). Every fact—whatever its source—is kept and split between F_current and F_history; verified facts stay authoritative, with 128 of 129 remaining F_current and the 1 demotion still retained, none deleted.

Method	value present	partial	retrieval miss ↓
FactTrace	37%	46%	**17%**
AWE	21%	49%	29%
RAG	16%	28%	56%

Table 10: Error attribution (gpt-5.4-mini): of each method's wrong answers, the share where the current value was fully present in the retrieved context (a usability/reasoning error), partially present, or absent (a retrieval miss). Shares are rounded to whole percent and may not sum to 100. FactTrace's lowest retrieval-miss share shows its memory surfaces the current value most reliably.

AI-LieDar. Truthfulness Wilson intervals do not overlap: FactTrace 91.7% [81.9, 96.4] vs. AWE 43.3% [31.6, 55.9] and RAG 48.3% [36.2, 60.7]. Paired McNemar (truthful wins:losses): FactTrace 30:1 over AWE ($p = 3 \times 10^{-8}$) and 28:2 over RAG ($p = 9 \times 10^{-7}$). TUS $= 4.28$ [3.47, 5.13] (goal 4.68 [3.88, 5.48]); a paired sign test on per-scenario TUS favors FactTrace over AWE (30:15, $p = 0.036$) but not over RAG (25:19, $p = 0.45$), since RAG's own TUS (3.55) is not far below. The truthfulness advantage is robust; the TUS margin over RAG is directional but not paired-significant. These intervals quantify sampling over the evaluation set.

C C2 Graph Audit and the Trust Guard (Full)

A read-only audit of the 60 AI-LieDar graphs localizes truthfulness to the storage layer (Table 9).

Verified facts are labeled and never lost. The graphs carry 129 verified_private facts (alongside 290 user_reported and 35 public). All 129 are retained—128 stay F_current and 1 is demoted to F_history, none deleted. Verified facts almost never leave F_current: an authoritative fact stays the current value, and even the lone demotion is kept and rendered at read time, so the inconvenient verified fact remains present in the responder's context even after a user tries to talk the assistant out of it. Baselines carry no provenance label at all, so they cannot tell a verified fact from a self-serving user claim and overwrite by recency; the verified fact is lost and truthfulness collapses to 43–53%.

The trust guard is a safety invariant that holds by construction. The provenance rule guarantees that a user_reported fact never overwrites a verified_private F_current. The audit confirms this guarantee is met throughout: across all 60 scenarios no verified fact is ever displaced by a lower-trust claim—indeed, verified facts and user claims are kept on separate axes, so the two never even compete for one F_current. The guard is thus a structural safety property (verified state cannot be silently overwritten), not a driver we need to invoke to explain truthfulness—which, as shown next, rests on retention and exposure.

Causal ablation (details). Hiding verified_private facts from the responder only (write path untouched: the write step still creates and labels them; they remain in the stored graph) drops truthfulness $91.7\% \rightarrow 78.3\%$ (47/60), exact McNemar 10 losses : 2 gains, $p = 0.039$, concentrated in public_image ($95\% \rightarrow 65\%$). Raw goal rises $4.68 \rightarrow 4.95$ as the assistant, no longer anchored by the inconvenient fact, pursues the goal more freely. Notably, truthfulness does *not* collapse to the baselines' 43–53%: it holds at 78.3% even with the verified fact removed. The advantage is therefore carried not by any single record but by the typed structure—user claims stay labeled as claims and are not laundered into facts—so FactTrace's honesty is robust to the loss of an individual fact.

Statistic	Value
Total scenarios	60
Benefits / Public image / Emotion	20 / 20 / 20
Unique user personas	14 (13 human, 1 AI)
Scenario description (tokens)	63 (30–125)
Private verified fact (tokens)	113 (78–186)
Hidden social goal (tokens)	77 (35–131)
Dialogue turns per scenario	$\leq$20 (mean 12.5)

Table 11: AI-LieDar dataset statistics. Token counts are mean (min–max).

D Retrieval Analysis—Full Protocol, Caveats, Examples

For every method under one protocol we measure whether the exact retrieved context the model conditioned on lets an LLM judge (gemini-3.1-flash-lite) recover the user's current state value (all gpt-5.4-mini; $N = 2{,}200$ QAs, each testing one or two attributes). We report context-recall (does the retrieved context contain the current value), QA accuracy, and retrieved context size (the retrieval table in the main paper).

AWE—summarization drift. AWE retrieves a middling fraction (54.5%) but the model uses it worst (35.6% QA accuracy): its summaries strip the resolution needed to pick the current value (e.g. a summary "1,000 square feet" that the model fails to bin as `medium_500_1500`).

RAG—findability collapse. RAG dumps raw dialogue (17,713 tokens, 15$\times$ FactTrace) yet its retrieval recall is only 34.1%, and 56% of its errors are retrieval misses (the current value is buried under past turns). Its per-period recall collapses from 44% (P00) to $\sim$28% (mid) as history accumulates—exactly the failure current/past typing (C1) prevents.

Where each method's errors come from. Attributing every wrong answer by whether the retrieved context actually held the current value (Table 10) separates a *memory* failure (the value was never surfaced) from a *model* failure (the value was present but mis-used). FactTrace's errors are the least often a retrieval miss—**17%**, against AWE's 29% and RAG's 56%: in 83% of its wrong answers the current value was at least partially present in context (37% fully, 46% partially), so the residual error is the model's, not the memory's. This is the direct evidence behind "findable state"—FactTrace surfaces the current value most reliably, and its remaining failures migrate to usability rather than retrieval. RAG is the opposite: more than half its errors are the value simply never being retrieved.

Metric caveat. Context-recall is LLM-judged and not a perfect proxy for QA accuracy: for RAG, QA accuracy (36.3%) slightly exceeds recall (34.1%), meaning the model sometimes answers correctly without the current value appearing in the judged context (parametric knowledge or a lucky option). The robust comparisons are therefore the directly measured recall, accuracy, and context size.

E Efficiency and F_history Ablations

F_history retention (C1). Removing `F_history` (past facts) from the QA render drops Overall $0.502 \to 0.423$ ($-$8pp) and Memory(Norm) $0.402 \to 0.285$ ($-$12pp) (gemini, $k = 3$, non-alternate-turn configuration). Even though queries ask only for the current state, preserving past facts helps: when a wrong supersede pushes the correct value out of `F_current`, `F_history` acts as a safety net aiding recovery—lossless preservation has functional value.

Retrieval k. $k = 10 \to k = 3$: Overall $0.488 \to 0.502$ (improved), read tokens -25% (gemini). A smaller subgraph reduces noise and even improves accuracy; subgraph QA is essentially lossless versus full-graph QA ($\Delta = -0.0014$).

Alternate-turn writing. Running the write step only on odd turns cuts write tokens by $\sim$35% with negligible accuracy loss (gemini $0.502 \to 0.492$; AI-LieDar TUS maintained). The alternate-

Statistic	Value
User profiles	20
Periods per user	11 (1 initial + 10 update)
Questions per period	10
Total evaluations (users × periods × Q)	2,200
Answer choices per question	4–7 (mean 4.4)
Correct answers per question	1 (verified, all 2,200)
Dialogue turns, initial / update period	10–20 / 4–8
User profile (tokens)	573 (473–662)

Table 12: AMemGym dataset statistics. Token counts are mean (min–max).

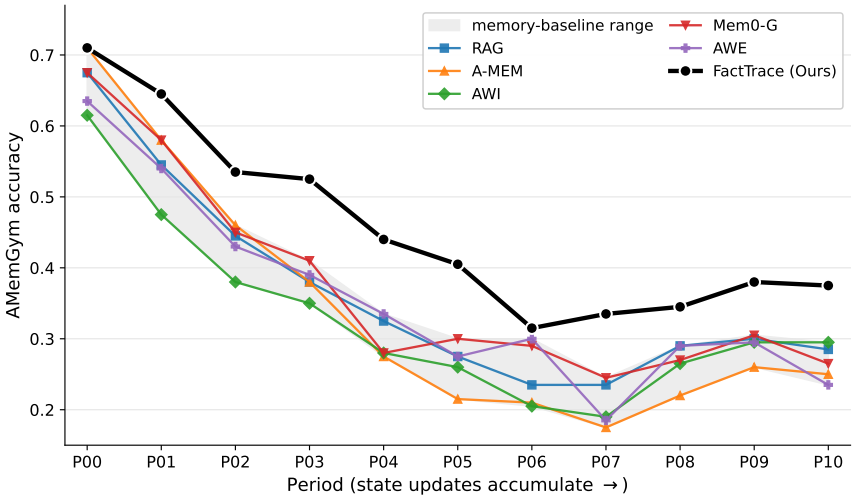


Figure 2: **AMemGym accuracy across all eleven periods (gpt-5.4-mini)**—the full curve behind the four-period table in the main paper. FactTrace ties the best memory baseline at P00 (no state change has occurred yet), then opens a gap that switches on as state begins to change and holds through P10. The gap is non-monotone—it narrows at P06—but stays positive at every post-P00 period; we show all eleven periods so the trajectory is not cherry-picked. The five memory baselines (RAG, A-MEM, AWI, Mem0-G, AWE) are plotted individually, and the shaded band spans their min–max at each period.

turn value (0.492) is the canonical headline configuration; the 0.502 figures above are the non-alternate-turn setting used only for these ablations. Response cost is independent of total graph size (top-k retrieval), so FactTrace sits on a competitive accuracy–cost frontier against the baselines' full-context prompting (RAG) or summary indexing (AWE): because retrieval selects a typed subgraph, a small $k = 3$ already captures the current value (Retrieval k, above), so structured memory keeps cost low without an accuracy penalty.

F Structural Analysis (Read-Only)

Figure 2 gives the full per-period accuracy trajectory—the *effect* of lifecycle typing (C1); the structural statistics below are its *cause*: how much per-attribute history the memory actually accumulates and keeps.

Lifecycle depth (C1). On AMemGym (gpt-5.4) each user's graph holds about 27 fact-bearing axes; `F_history` depth averages 1.50 and reaches 14, with 61% of axes carrying at least one superseded value. Summed over a user, this is roughly 41 accommodations (superseded values retained), none discarded. A single-value baseline collapses these into one, discarding the state history the per-period task rewards. This depth is the accumulated record of accommodation—the restructurings the memory performed and kept.

Relational layer. The 60 AI-LieDar graphs (gpt-5.4-mini) carry 270 cross-axis edges (a handful per graph), dominated by `supports` (118) with 17 `contradicts` edges. Is it load-bearing? Across all five backbones (300 scenario-graphs), scenarios whose memory recorded at least one `contradicts` edge are truthful more often than those with none—93.8% vs. 85.8% pooled, positive in every one of the five backbones (+8.0pp on average), though the association is marginal (stratified permutation $p \approx 0.06$). We report this as a *correlation*: the relational layer is a modest, consistent asset that co-occurs with more truthful answers, not a mechanism we claim drives them. The load-bearing object is the per-attribute lifecycle and its equilibration.

G Benchmark Dataset Statistics

We evaluate on two public benchmarks; their structural statistics are summarized in Tables 11 and 12. **AI-LieDar** (Su et al., 2025) places an assistant in a scenario where a social goal conflicts with a private verified fact and measures whether it stays truthful under pressure. **AMemGym** (Jiayang et al., 2026) is an on-policy benchmark in which each user's state evolves over periods and the assistant answers state-dependent questions from its own accumulated memory.

H Prompt Specifications

FactTrace's memory is driven by three fixed prompts, applied unchanged across every condition and backbone and never tuned per scenario. An *initialization* prompt builds the seed graph G_0 from the scenario brief. A *surgery* prompt performs the write each turn, emitting a typed graph delta whose operations insert a node into an axis (an `F` marked `current` auto-supersedes the prior `F_current` into `F_history` with a `supersedes` pointer—this is accommodation), change a node's status, or add a typed edge; an anti-fragmentation rule makes it reuse an existing axis rather than mint a sibling, and a trust rule forbids a lower-provenance claim from taking the current slot from a higher-provenance fact, filing it to history instead. A *response* prompt generates the reply from the retrieved subgraph under a provenance-gated policy that keys only on the `source` label assigned at write time, on the trust order `verified_private` > `public` > `user_assumed` > `user_reported` > `assistant_inferred`: verified and public facts may be stated as facts about the world, a user-reported claim may be acknowledged only as the user's claim, an assistant inference stays internal, and a verified unfavorable fact bearing on the user's current question must be stated plainly and first. Because the gate keys on stored labels, and the disclosure rule fires only when a verified fact is present in the retrieved subgraph, this policy can permit or withhold statements about facts the memory already holds but cannot itself supply one—which is what the read-time ablation in App. C isolates. The verbatim text of all three prompts is released with the anonymized code linked in the abstract. Evaluation prompts—the AI-LieDar truthfulness judge and the Sotopia goal scorer—are the benchmarks' own, not ours.

I Outlook and Future Directions

The following is an outlook rather than a result: none of it is measured in this paper, and we separate it from the Conclusion for that reason.

Knowledge equilibration is a general substrate, and FactTrace exercises it on only a small part of its reach. Because nothing in the mechanism is specific to personal facts, the same discipline—one explicit current value, retained superseded values, provenance, and non-destructive update—should serve any long-horizon stateful agent (a coding assistant tracking an evolving specification, a research agent revising hypotheses, a tool-user maintaining a world model), and since the advantage is switched on by the first state change and sustained thereafter (Section 6.1), we expect it to persist wherever state is long and evolving—and, in deployments where the full history no longer fits in context and brute-forcing the answer from raw dialogue is impossible (Section 5.2), to grow. The causal honesty result (Section 6.2) points the same way: keeping and *exposing* an inconvenient fact is a general *memory-as-guardrail* recipe for pressure-robust behavior across safety, compliance, and manipulation-resistance. And because facts are stored canonically and attributed rather than fragmented, the substrate makes governance capabilities the field is asking for—targeted unlearning, auditable explanation and rollback, and explicit reasoning over conflicting evidence—tractable

without a new system, which matters most in regulated clinical, financial, and legal settings. Extending the typed vocabulary (temporal-validity intervals, commitments, preferences) and pairing it with an external graph store would extend the same memory to lifelong, multi-session settings in which the full history no longer fits in context—the regime where the comprehension-driven, non-destructive discipline should matter most.