AgenticLOCI: A Multi-Agent Memory Framework for Conversational AI with Longitudinal Personalization

Anonymous ACL submission

Abstract

Long-term conversational AI requires two distinct capabilities: remembering what was discussed and adapting to how a user's preferences evolve over time. We argue these operate on different timescales and require separate but interoperable architectures. AgenticLOCI is a five-agent framework built on this premise: session retrieval fuses a semantic LLM ranker (Engram, STM) with a keyword-lobe scorer (Hippo, LTM) via Reciprocal Rank Fusion, while personalization is maintained through a versioned deliberation-state memory that accumulates empirically grounded decision-episode snapshots across interactions. We evaluate on three benchmarks using Qwen 2.5-7B-Instruct. On LoCoMo, the AgenticLOCI Backbone achieves F1 = 0.342 versus BM25 = 0.285 while consuming 7.3× fewer input tokens per item. On LongMemEval-S, item-level analysis confirms that Engram and Hippo solve architecturally distinct sub-problems; at 7B scale Hippo's keyword precision leads, a reversal expected to resolve at larger model sizes where LLM ranking is more reliable. On PersonaMem, all configurations perform within a narrow band, establishing that memory benefit is conditional on context-budget pressure. The longitudinal personalization layer is evaluated on a synthetic mobility benchmark with known user trajectories, achieving restraint of 0.976 and move-precision of 0.583 against an ablation baseline. We demonstrate the full framework as a personalized urban mobility assistant for an anonymized German city.

1 Introduction

Conversational AI that is genuinely useful over time must do two things well: remember what was discussed, and adapt to how a user's preferences, constraints, and decision patterns evolve across repeated interactions. Most systems fail at both and, crucially, for different reasons. Systems that forget users when a session ends suffer from an architectural shortcoming: no persistent storage, resulting in models operating purely as stateless responders that repeatedly reprocess context without historical continuity (Borro et al., 2026; Mason, 2026). Systems that remember indiscriminately suffer from a conceptual one: no distinction between a passing remark, a durable fact, a context-bound constraint, and an accumulated preference signal that should shape future interaction (Borro et al., 2026; Pink et al., 2025). Conflating these failures is a primary reason existing systems address neither memory nor personalization well.

Recent work has made substantial progress on the memory side. Hierarchical storage tiers, offline consolidation pipelines, and graph-structured knowledge bases now reliably preserve conversational content across long horizons (Packer et al., 2024; Fang et al., 2025; Wu et al., 2026; Zhang et al., 2026c). We build on these advances. However, memory is not personalization. The unit of memory in these systems is typically the dialogue: topic clusters, episodic events, retrieved utterances, or summarized interaction histories. Such representations preserve what was said, but they do not necessarily explain how a user repeatedly arrives at preferences under competing goals, affective reactions, constraints, and uncertainty.

Where personalization has been addressed, it has been treated as an extension of memory rather than a distinct problem (Sarin et al., 2025; Zhang et al., 2026c; Tan et al., 2025). Systems that maintain a user-profile artifact accumulate facts inside the same pipeline that handles recent dialogue, update them synchronously during the request lifecycle, and provide no audit trail of how the profile evolved (Sarin et al., 2025; Zhang et al., 2026c). Memory and personalization thus share the same timescale, the same storage surface, and the same maintenance discipline, ensuring that neither receives the treatment its problem actually demands.

Our position. We believe memory and person-

alization operate on different timescales and require different architectures. AgenticLOCI is built on this premise. A short consolidation loop runs at session end, deciding which facts are durable enough to enter the system's working context. A long consolidation loop runs over days and weeks, observing the gap between what the system recommends and what the user actually chooses, and uses that signal to update a versioned deliberation-state memory, accumulating empirically grounded decision-episode snapshots rather than overwriting a single static profile. Five specialized agents, each grounded in a specific cognitive function, divide the work so that no single component must be both fast and reflective. We demonstrate the framework on a personal mobility assistant for an anonymized German city, where the right recommendation depends not just on live conditions but on the user's accumulated decision history.

2 Related Work

Agentic memory systems. A growing body of work addresses long-horizon memory for LLM agents, building on earlier research into memory management across dialogue sessions (Bae et al., 2022; Xu et al., 2022). LightMem (Fang et al., 2025), GAM (Wu et al., 2026), and HiMem (Zhang et al., 2026c) organize dialogue into tiered stores and decouple consolidation from online inference via session-end or semantic-boundary triggers, showing strong retrieval on multi-session benchmarks (Maharana et al., 2024; Wu et al., 2024). MemGPT (Packer et al., 2024) frames memory as OS-style paging, letting the model manage which context to fetch on demand. Generative Agents (Park et al., 2023) demonstrate long-horizon behaviour through memory streams, retrieval-based reflection, and planning, establishing a simulation paradigm that informs our longitudinal evaluation design. More recent work either distributes memory operations across specialized components for compute efficiency (Zhang et al., 2026b; Qiao et al., 2026) or trains CRUD-style policies with reinforcement learning (Yan et al., 2026; Yu et al., 2025; Guo et al., 2026). These systems share a common orientation: memory is a retrieval resource serving question answering or task completion, and their retrieval stages largely extend the Retrieval-Augmented Generation paradigm (Lewis et al., 2020) to multi-session settings. They do not explicitly model how repeated decision episodes can be stored, versioned, and aggregated into a longitudinal representation of user preferences.

User-profile components. Early persona-based dialogue work (Zhang et al., 2018) showed that providing agents with a fixed set of personal facts improves consistency; subsequent systems have moved toward dynamic, evolving user models. The systems closest in spirit to AgenticLOCI are Memoria (Sarin et al., 2025), HiMem (Zhang et al., 2026c), and RMM (Tan et al., 2025), all of which pair dialogue memory with a user-model component. RMM uses a prospect-and-retrospect strategy to reflectively update personalized memory, but shares the same timescale as dialogue retrieval and maintains no typed schema or audit history. Memoria accumulates (subject, predicate, object) triplets extracted from each message into a knowledge graph, recency-weighted at retrieval time; HiMem includes a free-form note category for "stable user traits" updated by the same conflict-aware reconsolidation step applied to all other notes. In both cases the user model shares its storage surface and maintenance rhythm with dialogue memory: updates happen synchronously during the request lifecycle, no typed schema separates stable traits from context-bound decision states, and no version history is preserved. The resulting risk is *silent model drift* where the system's belief about the user shifts in ways that are neither inspectable nor reversible. PersonaMem (Jiang et al., 2025) quantifies a related ceiling: across preference-aligned MCQ over multi-session user histories, retrieval-based systems plateau well below the full-context oracle, establishing empirically that ad-hoc session retrieval cannot substitute for a proactive longitudinal personalization layer.

Positioning. AgenticLOCI synthesizes a set of complementary insights from this landscape rather than introducing a wholly new mechanism. It adopts offline sleep-time consolidation from Fang et al. (2025), semantic-boundary triggers for Hippo from Wu et al. (2026), typed conflict classification for profile-update logging from HiMem's reconsolidation step (Zhang et al., 2026c), the fundamental insight that a dedicated user-profile component is necessary from Memoria (Sarin et al., 2025), and typed sub-query routing in Thalamus from Zhang et al. (2026b). Reinforcement-learning memory frameworks (Yan et al., 2026; Yu et al., 2025; Guo et al., 2026) address orthogonal concerns; Engram and Hippo are natural RL fine-tuning targets in

future work. What AgenticLOCI adds over this synthesis is precisely two things: a second, slower consolidation timescale that operates on versioned deliberation-state memory rather than on dialogue events alone, and an append-only version history for every profile update. Appendix B maps the landscape in tabular form.

3 The AgenticLOCI Framework

AgenticLOCI splits memory by timescale rather than by topic. A single turn flows through a fast loop. A pattern across many sessions flows through a slow loop. The two loops never compete for the same prompt slot, and no single agent is asked to be both fast and reflective. Figure 1 shows the resulting structure.

3.1 Architecture

The framework operates on three timescales. *Real time* captures every utterance the moment it occurs, without judgement, into a short-term episodic store. *Sleep time, short loop* runs at session end and decides which facts are durable enough to enter the assistant's working context as *flashbulb events*. The test is operational: would it be helpful, or weird, if the assistant brought this up unprompted three days from now. *Sleep time, long loop* runs across days and weeks as a pipeline: a backend job transfers settled entries from the short-term store into a long-term archive on a configurable cadence, and only then does a long-term consolidator read that archive, reflect across many sessions, and update the versioned deliberation-state memory. The three loops are independent on purpose: a single atypical action does not push the slow loop, a passing remark that contradicts a flashbulb entry does not erase it without thought, and the slow loop is allowed to distinguish noise from repeated preference evidence while the fast loop reacts quickly.

The framework names five agents after structures in the brain. Each mirrors the function its biological inspiration performs.

Broca is the only agent that speaks. It is a ReAct-style language model (Yao et al., 2023) with access to live data tools. On every turn it reads the user's profile, the current flashbulb facts, a natural-language narrative rendered from the Longitudinal Deliberation Profile, and the visible session history, then synthesises one recommendation in human voice. Broca never sees the structured numbers directly, only the narrative. This keeps the reply natural and prevents internal scores from leaking into the conversation.

Thalamus is the orchestrator. It owns the session lifecycle, including the idle timeout that decides when a conversation has effectively ended. It logs turns into Engram, runs Broca, and schedules consolidation. It never generates content.

Engram is the short-term episodic store. Writes are unconditional; reads require the agent to be awake. Engram also owns the Flashbulb mechanism, the short-loop consolidator that asks a reflection model what to keep and writes the decisions back as safe operations on the flashbulb store: the small set of durable facts Broca reloads on every subsequent turn, so a stated constraint such as a stolen bike stays within reach and shapes recommendations until the user resolves it. This curation is deliberately kept off Broca, the online agent that spends each turn on live tool calls and must stay fast. Flashbulb consolidation is reflective work, so it is deferred to session end and assigned to Engram. A separate backend mover transfers settled Engram entries into the long-term archive on a configurable cadence, after which Hippo can act on them.

Hippo is the long-term consolidator. It runs offline, never during a turn, and operates on the long-term archive rather than on Engram directly. For each batch of newly archived sessions it derives lobe snapshots across seven families (mobility, schedule, health, places, social, preferences, transient context) and computes a conservative profile delta. New profile versions are written only when the proposed change clears a confidence threshold. Every change is versioned and forgetting candidates are queued for review rather than executed. The reference implementation drives Hippo's lobe derivation with Recursive Language Models (Zhang et al., 2026a).

Longitudinal Deliberation Profile is the user's evolving decision profile. It stores empirically grounded decision-episode snapshots that summarize needs, values, environmental tolerances, mode weights, contextual constraints, confidence, conflict, commitment stability, and a short explanation of the drivers behind recent decisions. These snapshots originate from external dynamic deliberation architectures: process-oriented decision models that represent deliberation as the temporal evolution of interdependent preference, constraint, and affective-cognitive variables, rather than as a single static preference label (van Gelder, 1998; Busemeyer and Townsend, 1993). In deployment, such

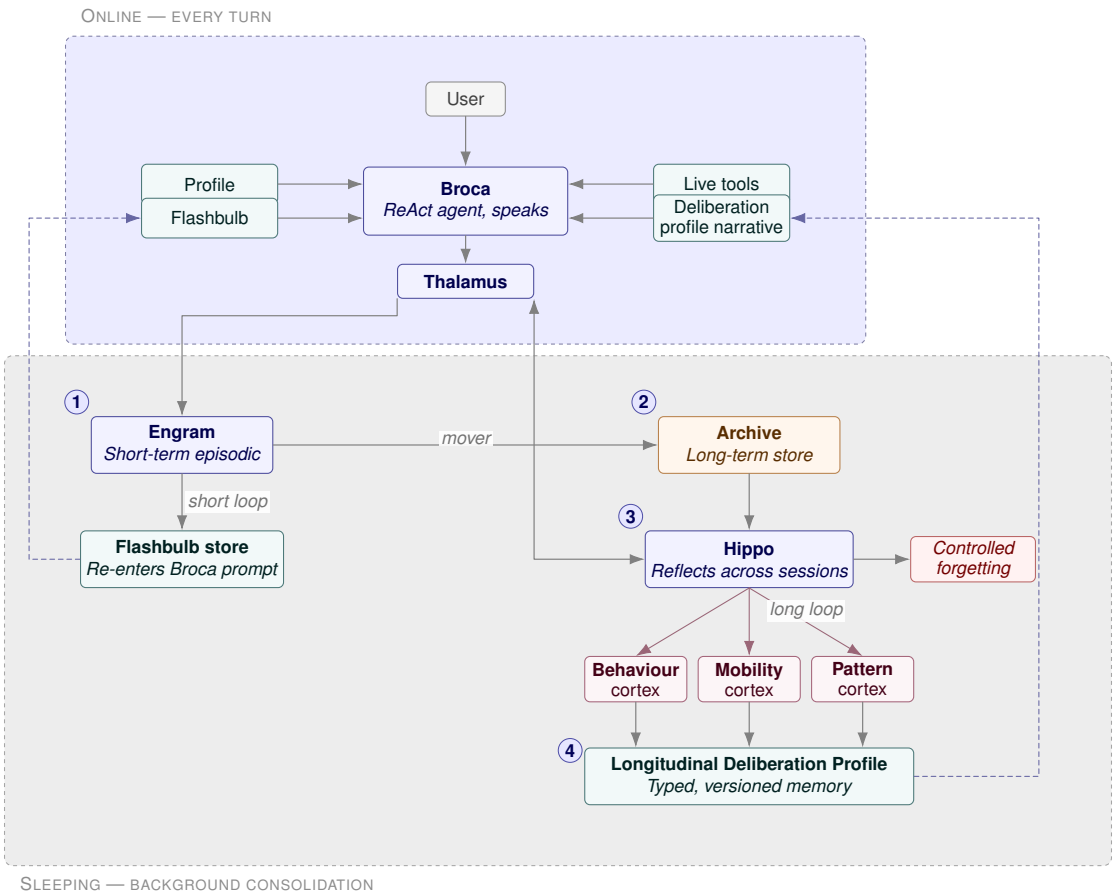

Figure 1: The AgenticLOCI framework. Online (top): Broca rebuilds its prompt on every turn from three always-read surfaces (Profile, Flashbulb, and the deliberation-profile narrative) and calls live tools. Thalamus orchestrates the turn and logs it into Engram. Sleeping (bottom): at session end, Engram's Flashbulb mechanism updates the durable facts that re-enter Broca's prompt (*short loop*). A backend mover transfers settled Engram entries (**1**) into the long-term Archive (**2**), where Hippo (**3**) reflects across sessions, derives behaviour, mobility and pattern cortices, queues forgetting candidates for review, and writes a versioned update to the Longitudinal Deliberation Profile (**4**). Thalamus schedules Hippo's consolidation runs and can query its state. A render step turns the structured profile into the natural-language narrative Broca reads (*long loop*).

architectures can be embedded in human-in-the-loop personalization workflows, where repeated observations, user corrections, and behavioral choices update the profile while preserving earlier versions for audit and revision (Amershi et al., 2014). It is updated only by Hippo. AgenticLOCI does not prescribe the internal equations of the deliberation system; it only requires that the system expose structured, empirically grounded snapshots that can be stored, rendered, versioned, and consolidated. A single snapshot describes one situated decision episode and should not be interpreted as a stable trait by itself. Across repeated interactions, however, the versioned sequence of snapshots can be aggregated into a probabilistic profile of the user's longer-term tendencies. A separate render step produces a natural-language narrative from this structured document; that narrative is what Broca reads on every turn, while the raw numbers stay internal. The audit history of every prior version is preserved, so the trajectory of the system's belief about the user is inspectable and reversible.

Two thin surfaces sit beside the agents and feed Broca's prompt directly. *Profile* holds stable, user-declared facts and preferences. *Flashbulb* holds the curated short list of durable facts Engram has decided are worth keeping always within reach, capped at twenty entries.

3.2 The Two Consolidation Loops

On a chat turn, Thalamus checks the session tracker and, if the prior session has changed or is idle past the configured timeout, fires the short loop. It then records the user turn, asks Broca to respond, and logs the reply. Broca rebuilds its prompt fresh on every turn, so any new deliberation-profile narrative

produced by Hippo between turns becomes visible on the next message.

The short loop runs at session end. The Flashbulb mechanism reads the just-ended transcript, places the current flashbulb list beside it, and asks a reflection model what to add, remove, or update. The plan is returned as strict JSON and applied via method calls on the flashbulb store; the model never writes to disk.

The long loop runs as a backend pipeline on its own cadence. A mover transfers settled Engram entries into the long-term archive. Hippo then reads the archive across many sessions and derives lobes, with particular attention to the gap between what Broca proposed and what the user actually chose. Each proposal and each explicit user confirmation is preserved in the archive with a snapshot of the active deliberation profile at the moment of recommendation, and that gap is the strongest learning signal for the profile delta. The updated profile is rendered into a fresh narrative, which Broca picks up on the next turn.

The architecture treats the long loop as advisory by default. Forgetting candidates are queued for review, and any deletion of archived entries is gated behind an explicit flag, because the system should earn the right to forget.

We instantiate the framework as a personal mobility assistant for an anonymized German city, integrating live IoT sensors, routing services, and a domain knowledge graph through four MCP servers. Implementation details, including the tool surface, urban digital twin, and spatial graph, are deferred to Appendix A.

4 Evaluation

AgenticLOCI's memory subsystems serve distinct cognitive roles on different timescales, and no single benchmark exercises all of them. We therefore structure this section into two parts. Section 4.1 evaluates the episodic retrieval stack, Engram and Hippo, on three public long-term memory benchmarks using a consistent Qwen 2.5-7B-Instruct backbone. Section 4.2 evaluates the two consolidation components, the Flashbulb mechanism and the Longitudinal Deliberation Profile, on a longitudinal simulation: their correctness concerns the timing and accumulation of facts and decision-state evidence across sessions and weeks, when a fact should be captured, dropped, or allowed to shift a durable probabilistic tendency, which single-pass retrieval benchmarks cannot measure.

4.1 Episodic Memory and Retrieval Evaluation

All experiments use **Qwen 2.5-7B-Instruct** (Qwen Team, 2024) as the end-to-end backbone (reader, LLM ranker, and QA judge), so differences across pipelines reflect architecture rather than model selection. Full benchmark descriptions, metrics, and pipeline specifications are in Appendix C.

Systems. **BM25** (Robertson et al., 1994) is the lexical baseline; **Oracle** feeds only gold sessions (upper bound). **Engram** (STM) ranks 3-turn summaries via one LLM call. **Hippo heuristic** (LTM) scores keyword-lobe overlap at zero LLM cost. **AgenticLOCI Retrieval** fuses both via Reciprocal Rank Fusion (Cormack et al., 2009):

$$\text{score}(s) = \frac{1}{r_E(s) + 60} + \frac{1}{r_H(s) + 60} \quad (1)$$

AgenticLOCI Backbone extends retrieval with batched cross-session fact extraction (one additional LLM call). **Deliberation Profile** (direct) compresses the full conversation into a preference-centric structured representation and answers from that. All Hippo and Backbone conditions use LongMemEval-appropriate lobe categories; lobe vocabulary ablation in Appendix D.4.

LoCoMo ($n = 1{,}986$). Table 1 shows LoCoMo results. AgenticLOCI Backbone (F1 = 0.342) surpasses BM25 (0.285) while consuming 7.3× fewer input tokens (3,072 vs. 22,340), approaching the full-context baseline (0.370). BM25 underperforms because LoCoMo's friend-chat register does not lexically anchor questions to sessions, removing its structural advantage. The direct deliberation-profile condition degenerates on factual QA (F1 = 0.227) while achieving near-perfect adversarial F1 (0.981): its preference-centric schema deliberately discards episodic detail, confirming that no single representation suits all cognitive tasks. BLEU-1 and token costs in Appendix D.1.

LongMemEval-S ($n = 500$). Table 2 shows LongMemEval-S results. BM25 dominates retrieval (R@5 = 0.909) because the benchmark's user-assistant log format provides a structural lexical advantage. At 7B scale, the individual Hippo heuristic component leads its layer

Table 1: LoCoMo results ($n = 1{,}986$, Qwen 2.5-7B). AgenticLOCI Backbone surpasses BM25 on overall F1 while consuming 7.3× fewer input tokens per item. Oracle F1 = 0.536 bounds the 7B reader. Adversarial F1 rewards refusal. BLEU-1 in Appendix D.1.

Pipeline	Overall F1	Multi-Hop	Temporal	Open-Domain	Single-Hop	Adversarial	Avg Input Tok.
Oracle (upper bound)	0.536	0.096	0.172	0.624	0.363	0.836	1,176
Full-context baseline	0.370	0.051	0.080	0.390	0.185	0.729	22,340
BM25 + QA	0.285	0.035	0.042	0.162	0.104	0.861	2,418
AgenticLOCI Retrieval (Engram + Hippo, Top-5)	0.311	0.028	0.043	0.193	0.113	0.913	4,301
AgenticLOCI Backbone (Retrieval + Facts)	**0.342**	**0.049**	**0.059**	**0.268**	**0.167**	0.859	**3,072**
Deliberation Profile (direct extraction)	0.227	0.028	0.001	0.007	0.018	**0.981**	884

Table 2: LongMemEval-S retrieval and QA, Qwen 2.5-7B, $n = 500$. **Bold** = best AgenticLOCI condition. Component-wise breakdown (Engram, Hippo alone) and per-type metrics in Appendices D.3–D.5.

Pipeline	R@1	R@5	MRR	QA
BM25 (session-level)	0.551	0.909	0.909	0.496
AgenticLOCI Retrieval	0.378	0.675	0.727	0.398
AgenticLOCI Backbone	**0.393**	**0.680**	**0.741**	0.386

(R@5 = 0.744, QA = 0.424; Appendix D.3): small-model LLM rankings add noise rather than semantic signal, a scale-contingent finding. The SS-A/SS-U asymmetry confirms the two layers are architecturally complementary, with Engram leading on assistant-stated content and Hippo on user-stated and knowledge-update items; item-level analysis (Appendix D.8) confirms this holds structurally. AgenticLOCI Backbone recovers precision (R@1 = 0.393, MRR = 0.741) through summary-mode ingestion. Full metrics in Appendices D.3–D.5; 32B cross-model results in Appendix D.6.

PersonaMem ($n = 589$, 32K split). All configurations cluster within 5 pp of the full-context baseline (Full-context 0.523; AgenticLOCI Retrieval 0.489; Deliberation Profile 0.477). PersonaMem's 32K context fits within the reader's window, removing the context-budget pressure that makes LoCoMo informative. We treat this as a *control*: flat results confirm that gains on LoCoMo reflect context compression, and that memory systems do not differentially harm preference-MCQ performance. Per-type breakdowns in Appendix D.2.

The retrieval ceiling. Across all three benchmarks, preference-type questions score near-zero for every retrieval system (best: 0.300 on LME-S). This is not a retrieval failure, it reflects a structural limit: no session-retrieval strategy can aggregate implicit preference signals distributed across months of interaction. Static benchmarks cannot exercise this capability; evaluating it requires longitudinal ground truth with known preference trajectories and versioned decision-state evidence, which we address next.

4.2 Memory Consolidation Evaluation

We evaluate consolidation on a personal mobility assistant deployed for an anonymized German city. The deployment grounds Broca in live urban data (weather, traffic, parking, transit) via four MCP servers and a spatial knowledge graph that together realise a work–life urban digital twin (Anonymous, 2026), and exposes routing across walking, cycling, transit, and driving. Mobility is a useful testbed because the right recommendation depends jointly on world state, spatial structure, and durable user tendencies, and the third of these is precisely what longitudinal personalization must learn. Full deployment details are in Appendix A.

The retrieval benchmarks in §4.1 cannot test the two components that embody our central premise, that recent memory and durable personalization are distinct problems on distinct timescales. Flashbulb curates episodic facts at each session's close and the Longitudinal Deliberation Profile is revised only across days and weeks. Its correctness concerns *timing and accumulation*: when a fact should be remembered, when it has gone stale, when a decision-state snapshot should be stored, and when repeated behaviour should shift a probabilistic user tendency. A single pass over a fixed store cannot reveal this, so we evaluate both on a longitudinal simulation with known ground truth.

The simulation comprises ten synthetic users, each with an initial deliberation profile and a script of ninety mobility decisions over thirty days in the city, every decision conditioned on weather, traffic, time of day, and stated life events such as an injury, a stolen bicycle, or a temporary commitment. A deterministic rule engine generates the scripts, so the correct behaviour is known for every event and every profile field; this is the compromise that yields ground truth no unscripted dialogue could supply. Three users warrant little or no change,

Behaviour	Metric	Result
Capture	Recall	0.79 [0.65, 0.92]
	Mean timing error	0.62 days
Expire	Recall	0.74 [0.52, 0.90]
	Mean timing error	1.04 days
Persist	Accuracy	1.00 (6/6)
Pins	Spurious rate	0.046 (5/109)

Table 3: Flashbulb behaviour across ten users (65 events, 109 pins). Intervals are 95% bootstrap CIs over users.

letting us measure restraint as well as sensitivity. We assess Flashbulb in §4.2.1 and the Longitudinal Deliberation Profile in §4.2.2, then in §4.2.3 collapse both loops into a single agent to test what the timescale separation itself contributes, reporting confidence intervals over the ten users throughout. Broca and Thalamus govern dialogue quality rather than consolidation and are left to a planned user study.

4.2.1 Flashbulb Evaluation

Flashbulb decides which newly stated facts are durable enough to keep in view and removes a fact once the user signals it no longer holds. Beyond straightforward cases ("my bike was stolen"), the synthetic events include harder ones: states that look temporary but persist, facts about third parties, and constraints whose resolution is never spoken aloud. Across ten users there are 65 such events, against which the system created 109 pins.

We measure four behaviours: *Capture* (is the fact pinned promptly?), *Expire* (is a temporary pin removed at the right time, neither early nor late?), *Persist* (does a permanent pin survive the full month?), and *Spurious pinning* (how many pins are noise?). Bracketed ranges are 95% bootstrap CIs over users.

Three of the four behaviours are strong (Table 3). The system captures 79% of stated events, and the captures it makes land within about half a day of when the user raised them. It retains every permanent fact it captured, six of six, for the full month, so a durable constraint such as a stolen bicycle is never silently dropped. Spurious pinning is low: a manual review of all 109 pins, with the classifications verified by an independent model from a different family (Cohen's $\kappa = 0.84$; Cohen 1960), found only 5 to be genuine noise, a spurious rate of 4.6%. Expire is the weakest axis at 0.74. The system removes a pin reliably when the user states that a fact is resolved, but it is slower to drop facts whose end is only implied by the passage of time.

Crucially, the errors are not spread evenly. They concentrate on the cases the events were designed to stress: every routine, plainly stated personal constraint, such as an injury, a flat tire, or a stolen bike, is captured correctly. The capture misses are instead ambiguous or indirect events: persistent cold spells and winter storms, which the system treated as transient weather rather than durable constraints; facts about third parties, such as a cluster of muggings reported on local news; and behaviour the user never states but that must be inferred, such as walking more after a rent increase. The expire misses follow the same pattern: the system leaves a pin in place when no resolution is ever spoken, as with a ten-day tram closure that simply ends. Flashbulb therefore handles the routine case reliably, and its residual errors fall on the durability judgments, third-party facts, and unspoken resolutions that make the synthetic benchmark deliberately hard. Per-axis failure breakdown in Appendix E.1.

4.2.2 Longitudinal Deliberation Profile Evaluation

The Longitudinal Deliberation Profile is the long-term personalization layer, and Hippo revises it offline after observing a month of behaviour. The risk for any such model is silent drift: the system quietly changing its belief about the user in ways that are wrong and that no one notices. A good consolidator therefore faces a two-sided test. It must leave a tendency alone when the evidence does not justify a change, and it must change one when the evidence does. Either side is trivial on its own, since a system that never updates never drifts and a system that updates constantly always catches real shifts. The real question is whether Hippo can tell a genuine preference shift from behaviour that was merely forced by circumstance.

From each user's 30-day script we derive a set of expectations for the profile: which fields should stay stable, which may move and in which direction, and how far a move may reasonably go. This yields 251 expectations across the ten users (207 fields that should stay stable, 70 of them boolean flags, and 44 sanctioned directional moves). We score three systems against them: **full Hippo**, the deployed LLM consolidator; **keyword Hippo**, an ablation that replaces the LLM with the heuristic keyword path; and a **no-update** baseline that freezes the profile at its initial state.

We report three metrics. *Restraint* is the fraction of should-stay-stable fields the system left unchanged. *Responsiveness* is the fraction of sanctioned moves it made in the correct direction. *Move-*

System	Restraint	Respons.	Move-prec.
Full Hippo (LLM)	0.976	0.159	0.583
Keyword Hippo	0.671	0.568	0.269
No-update	1.000	0.000	–

Table 4: Longitudinal Deliberation Profile consolidation across ten users (251 expectations). Full Hippo is far more restrained and more precise than the keyword ablation; no-update shows restraint alone is trivially achieved.

precision asks, of every change it made to a tracked field, how many were sanctioned. Bracketed ranges in the text are 95% bootstrap CIs over the ten users.

Two of the three results are clean and one needs interpretation (Table 4). Restraint is high: full Hippo leaves 0.976 [0.951, 1.000] of stable fields untouched, thirty points above the keyword ablation at 0.671 [0.598, 0.732], with confidence intervals that do not overlap. The LLM's reasoning, not keyword counting, is what prevents drift. It is also precise when it acts: move-precision is 0.583 against the ablation's 0.269, so it makes far fewer changes overall (12 versus 93 across the ten users) yet each is roughly twice as likely to be warranted. The no-update baseline scores a perfect restraint of 1.000 and zero on everything else, confirming that restraint alone is meaningless without the other axes.

Responsiveness is low at 0.159, but should be read as a lower bound: the metric cannot distinguish a missed preference from a principled decline. In the clearest case (Persona 0), Hippo correctly refused to record "transit up, walking down" because the transit-heavy pattern was forced by a leg injury, flat tyre, and bicycle theft. On constraint-free days the user chose to cycle 18 of 24 times. Responsiveness counts such principled declines as misses, so 0.159 understates the consolidator's quality.

Where Hippo errs, it errs by changing too much (five unwarranted changes across three users; detailed breakdown in Appendix E.2). A system that overwrote the profile in place would hide these; under AgenticLOCI each is a separate, retrievable version.

Write-surface analysis of the four protected snapshot blocks (deliberation dynamics, mode activations, valences, stressor tolerances) shows the LLM path writes into a protected block on five users (33 of 40 preserved) against 40 of 40 for the keyword path; the practical implication for deployment is detailed in Appendix E.2.

Metric	AgenticLOCI (split)	Monolith
Fast loop (Flashbulb)		
Capture recall	0.79	0.69
Expire recall	0.74	0.87
Persist accuracy	1.00	1.00
Slow loop (Profile)		
Restraint	**0.976**	0.874
Move-precision	**0.583**	0.235

Table 5: Splitting the loops across agents versus collapsing them into one agent, on the ten-user simulation. Fast-loop differences are within noise (95% bootstrap CIs over users span zero); the slow-loop advantages of the split are significant (restraint and move-precision CIs exclude zero).

4.2.3 Collapsing the Loops into One Agent

We test whether the timescale separation itself is necessary by collapsing both loops into a single *monolithic* agent: one model that speaks, pins facts, and consolidates the profile inside one continuously growing context, using the same backbone and rules as above.

The result is one-sided (Table 5). The fast loop is a tie, pinning facts once per turn is light work the online agent absorbs. The slow loop is where separation earns its place: restraint falls from 0.976 to 0.874 and move-precision from 0.583 to 0.235, both significant. An agent busy routing overwrites the profile on thin evidence; it is the slow loop specifically that needs its own timescale to weigh a month of behaviour at rest.

5 Discussion and Conclusion

The LoCoMo results establish the layered design's value: AgenticLOCI Backbone (F1 = 0.342) outperforms both BM25 (0.285) and plain retrieval (0.311) with 7.3× fewer tokens. On LongMemEval, RRF fusion underperforms Hippo alone at 7B because the small model adds ranker noise, a scale-contingent finding addressed in Limitations. Across all three benchmarks, preference-type questions score near-zero; the Longitudinal Deliberation Profile addresses this failure mode with restraint = 0.976 and move-precision = 0.583, all updates versioned and reversible.

We presented AgenticLOCI, a five-agent framework separating memory into two consolidation timescales: a fast loop for episodic flashbulb facts and a slow loop for typed, versioned deliberation-state memory accumulated across weeks, so that no component must be both fast and reflective.

Limitations

Model scale and LLM-ranker reliability. All experiments use Qwen 2.5-7B-Instruct. At this scale, Engram's LLM-based ranker introduces noise that causes RRF fusion to underperform Hippo heuristic alone on LongMemEval-S. This is a scale-contingent finding, not an architectural one: larger models are better calibrated as rankers, and the framework is designed with component-level replaceability in mind. The immediate fix, replacing Engram's LLM ranker with a dense embedding model (Karpukhin et al., 2020), also resolves the context-length ceiling that prevents independent RRF fusion at LongMemEval-M scale (≈500 sessions/item), where the 7B model's context limit forces a BM25 pre-filter that collapses signal independence.

Evaluation scope. The retrieval evaluation covers the episodic memory stack on three static benchmarks; the longitudinal evaluation covers Flashbulb and the Longitudinal Deliberation Profile on a synthetic 30-day simulation. **Thalamus and Broca** are conversational orchestration components whose quality is best measured through user studies on coherence and preference adaptation over extended sessions, which we defer to a planned deployment study.

Benchmark fit. LongMemEval's user-assistant log format provides a structural advantage to BM25 that makes it difficult for semantic retrieval to compete on this benchmark. LoCoMo better reflects conversational memory in realistic friend-chat settings. Neither benchmark directly exercises the longitudinal personalization pathway, which is purpose-built for the implicit, distributed preference signals that PersonaMem shows are within 5 pp of the full-context baseline when context fits. The longitudinal synthetic benchmark in §4.2 exists specifically because no public benchmark exercises consolidation timing or versioned preference accumulation.

Gap to reported baselines. The LongMemEval paper's primary QA result of 0.700 (Wu et al., 2024) uses a GPT-4o judge, Top-10 retrieved sessions, and the _M split. Our Qwen 2.5-7B Hippo heuristic (QA = 0.424, Top-5, _S) is separated from this number by model scale, retrieval budget, and benchmark split simultaneously. The closest model-matched baseline in the paper (Llama 3.1 70B (Grattafiori et al., 2024), Top-5, _M) is 0.588. We frame the contribution as a systems and architecture result, demonstrating that structured two-timescale memory beats lexical retrieval on conversational QA (LoCoMo) and characterizes the architectural boundary between retrieval and personalization, rather than as a state-of-the-art QA claim.

Synthetic longitudinal evaluation. The Longitudinal Deliberation Profile evaluation uses a deterministic rule engine to generate ground-truth user trajectories. This is the compromise that makes restraint and responsiveness measurable at all, but it introduces two limitations. First, the rule engine encodes our own assumptions about what constitutes a preference shift versus circumstance-forced behaviour; a real user's reasoning may differ. Second, ten synthetic users is small for tight confidence intervals on rare-event metrics like move-precision. Real-deployment evaluation on the anonymized system is the natural next step.

Ethics Statement

This work complies with the ACL Code of Ethics. We discuss three areas of ethical relevance.

Use of generative AI assistance. In accordance with the ACL Policy on AI Writing Assistance, we disclose the following use of generative AI tools during the preparation of this paper. **Claude Code** (Anthropic, 2025a) was used for coding assistance during framework implementation, evaluation scripting, and data processing. **Claude Sonnet 4.6** (Anthropic, 2025b) was used for writing assistance, including literature search, prose drafting, and editing of individual sections. All AI-generated text and code was critically reviewed, verified, and substantially revised by the authors. All research ideas, experimental design decisions, result interpretations, and novelty claims originate with the human authors. The final paper and its findings are the authors' sole responsibility.

Data and privacy. AgenticLOCI is evaluated on three publicly released benchmarks: LongMemEval (Wu et al., 2024) and LoCoMo (Maharana et al., 2024), which use synthetically generated or scripted dialogues, and PersonaMem (Jiang et al., 2025), which uses synthetically constructed user personas. All benchmarks contain no real user data. No real user data was collected or processed as part of this research. The anonymized deployment described in Appendix A operates on a

single-user, opt-in basis with no data shared beyond the local instance. The Longitudinal Deliberation Profile's append-only audit history is designed to give users full visibility and recoverability over their stored profile, which we view as a minimal requirement for responsible deployment of long-term personalized agents.

Societal considerations. Long-term user modeling raises concerns around profiling, data retention, and the potential for systems to entrench rather than surface user preferences. The Longitudinal Deliberation Profile's typed schema and version history are explicit design responses to these concerns: every update is inspectable, and no information is deleted without a recoverable trace. Future deployments should pair the architecture with clear user-facing controls for profile review, correction, and deletion in compliance with applicable data protection regulations.

References

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Anonymous. 2026. [citation anonymized for review]. Anonymized for blind review.

Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol. Accessed: 2026-05-17.

Anthropic. 2025a. Claude code: Agentic coding in the terminal. https://docs.anthropic.com/en/docs/claude-code. Accessed: May 2026.

Anthropic. 2025b. Claude sonnet 4.5 model card. https://www.anthropic.com/claude/sonnet. Model release, claude-sonnet-4-6.

Richard C. Atkinson and Richard M. Shiffrin. 1968. Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence, editors, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 2, pages 89–195. Academic Press.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, UAE. Association for Computational Linguistics.

Luiz C. Borro, Luiz A. B. Macarini, Gordon Tindall, Michael Montero, and Adam B. Struck. 2026. Memori: A persistent memory layer for efficient, context-aware llm agents. *Preprint*, arXiv:2603.19935.

Jerome R. Busemeyer and James T. Townsend. 1993. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459.

Flavio Cirillo, Gürkan Solmaz, Everton Luís Berz, Martin Peter Bauer, Bin Cheng, and Ernő Kovacs. 2019. A standard-based open source IoT platform: FIWARE. *IEEE Internet of Things Magazine*, 2:12–18.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. LightMem: Lightweight and Efficient Memory-Augmented Generation. *arXiv e-prints*, arXiv:2510.18866.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ziliang Guo, Ziheng Li, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2026. MemFactory: Unified Inference & Training Framework for Agent Memory. *arXiv preprint*. ArXiv:2603.29493 [cs].

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870. Association for Computational Linguistics.

Tony Mason. 2026. The missing memory hierarchy: Demand paging for llm context windows.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint*. ArXiv:2310.08560 [cs].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. ACM.

Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. 2025. Position: Episodic memory is the missing piece for long-term llm agents. *Preprint*, arXiv:2502.06975.

Jingyang Qiao, Weicheng Meng, Yu Cheng, Zhihang Lin, Zhizhong Zhang, Xin Tan, Jingyu Gong, Kun Shao, and Yuan Xie. 2026. Memory Intelligence Agent. *arXiv preprint*. ArXiv:2604.04503 [cs].

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Samarth Sarin, Lovepreet Singh, Bhaskarjit Sarmah, and Dhagash Mehta. 2025. Memoria: A Scalable Agentic Memory Framework for Personalized Conversational AI. *arXiv preprint*. ArXiv:2512.12686 [cs].

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439, Vienna, Austria. Association for Computational Linguistics.

Tim van Gelder. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21(5):615–628.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *Preprint*, arXiv:2410.10813.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, and Hongwei Wang. 2026. GAM: Hierarchical Graph-based Agentic Memory for LLM Agents. *arXiv preprint*. ArXiv:2604.12285 [cs].

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197. Association for Computational Linguistics.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2026. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. *arXiv preprint*. ArXiv:2508.19828 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2025. MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent. *arXiv preprint*. ArXiv:2507.02259 [cs].

Alex L. Zhang, Tim Kraska, and Omar Khattab. 2026a. Recursive Language Models. *arXiv preprint*. ArXiv:2512.24601 [cs.AI].

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. 2026b. Lightweight LLM Agent Memory with Small Language Models. *arXiv preprint*. ArXiv:2604.07798 [cs].

Ningning Zhang, Xingxing Yang, Zhizhong Tan, Weiping Deng, and Wenyong Wang. 2026c. HiMem: Hierarchical Long-Term Memory for LLM Long-Horizon Agents. *arXiv preprint*. ArXiv:2601.06377 [cs].

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a bike. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

A Reference Implementation: A Personal Mobility Assistant for an Anonymized German City

We realise the framework as a personal mobility assistant for an anonymized German city. Mobility in a real city is one of the few everyday decisions where the right answer depends on three things at once: the conditions of the world right now, the spatial structure of the city, and who the user is. The first two on their own get you intelligent routing; the memory layer is what turns intelligent into personal.

Substrate. Broca runs as a LangGraph ReAct agent (Yao et al., 2023) over an OpenAI-compatible chat endpoint. The reference deployment uses an institutional HPC service with a Qwen 2.5-7B-Instruct model (Qwen Team, 2024) for the conversational role; Hippo's offline reflection uses Recursive Language Models in a sandboxed Python REPL. The agent has access to eleven tools. Eight live in four stdio Model Context Protocol (MCP) subprocesses (Anthropic, 2024): a Neo4j server for the campus knowledge graph, a FIWARE server (Cirillo et al., 2019) for the urban digital twin, a routing server wrapping ORS and TomTom, and a context bridge that fuses the three behind one call. Three more are in-process tools that connect Broca to the framework itself: `propose_route`, `record_route_selection`, and `recall_memory`.

Urban digital twin and spatial graph. The FIWARE server exposes the urban work–life digital twin of the anonymized city (Anonymous, 2026). It serves eight live signal types: weather, parking availability, air quality, traffic flow, room sensors, vehicles, water levels, and a generic digital twin type. Broca can ask the city, in one line, for the temperature at a particular building, the free parking spots at the science harbour, or the congestion along a street, and the answer is either fresh or honestly absent. Every coordinate is validated against a city-level bounding box at the entry point, which catches latitude and longitude swap bugs before they leave the process.

The Neo4j server holds the static spatial structure beneath the digital twin: buildings, transit stops, lines, streets, landmarks, points of interest, and the relationships that matter for routing. The graph is read-only; a schema validator rejects writes and any reference to a label or relationship type outside the schema. A dedicated transit tool resolves natural origin and destination names through a full-text search and runs a gated path query that prefers shared lines and falls back to one-transfer routes. Walking, cycling, and driving come from ORS and TomTom through the routing server, with a compound tool that collapses five sequential calls (resolve, resolve, walk, cycle, drive) into one.

These three layers together give Broca what one might call *intelligent* routing: the world as it is right now, the world as it is mapped, and every alternative route between any two named places.

From intelligent to personal. What turns this combination into something that feels personal is the memory layer, wiring the flashbulb facts and the deliberation-profile narrative into the same prompt as the live data, on every turn. The same question, *from the opera house to the central square*, changes meaning entirely when a flashbulb entry says *bike was stolen, walking and transit only*. Broca filters cycling out of the candidate set before it calls the routing tools. The reply acknowledges the constraint in human terms (*since the bike is still gone, cycling is out*) and recommends transit. The constraint persists across sessions and is removed at the next consolidation if the user mentions a new bike. Across many sessions, persistent divergence between what Broca recommends and what the user actually chooses feeds Hippo's offline lobe derivation, which eventually nudges the longitudinal profile's mode weights in a versioned update. The next session, Broca rebuilds its prompt with a fresh narrative in view. The baseline has moved closer to the user, and the move is preserved in the audit history.

B Related Work: Detailed Comparison

Table 6 maps the agentic-memory and personalized-conversational-AI landscape against Agentic-LOCI's architectural contributions.

C Evaluation Setup

C.1 Benchmarks and Metrics

LoCoMo (Maharana et al., 2024) presents 1,986 free-form QA items across 10 synthetic multi-session dialogues spanning 5+ months. Categories include single-hop, temporal, multi-hop, open-domain, and adversarial (unanswerable) questions. Standard metrics are token-overlap F1 and unigram

Table 6: Memory systems in context: their architectural contributions, how they approach user modeling, and their relationship to AgenticLOCI. †LightMem-Z = Zhang et al. (2026b), distinct from Fang et al. (2025).

System	Use Case	Key Architectural Contribution	User Modeling	Adopted / Influence in AgenticLOCI
MemGPT (Packer et al., 2024)	General agents	OS-style context paging; self-directed memory invocation	None	Self-directed invocation pattern (Broca → Thalamus)
LightMem (Fang et al., 2025)	Long-context QA	Sleep-time deduplication; Atkinson–Shiffrin STM/LTM tiers (Atkinson and Shiffrin, 1968)	None	Offline sleep-time consolidation principle
GAM (Wu et al., 2026)	Multi-turn QA	Topic+event graph (TAN+EPG); semantic FSM consolidation	None	Semantic-event triggers for Hippo consolidation boundaries
HiMem (Zhang et al., 2026c)	Multi-turn QA	Episode+note hierarchy; typed conflict-aware reconsolidation	Free-form "stable traits" note category	Typed conflict logging (extend/revise/contradict) for profile updates
Memoria (Sarin et al., 2025)	Personalized chat	Recency-weighted KG persona engine; triplet extraction	Flat (subj., pred., obj.) persona graph	Core insight: a dedicated user-profile component is necessary
RMM (Tan et al., 2025)	Personalized dialogue	Prospective Reflection (topic-based session-end memory extraction + merge); Retrospective Reflection (RL-based retrieval reranker refinement via attribution signals)	None (topic summaries only; no typed user persona)	Reflective session-end consolidation pattern (Flashbulb analogue); RL retrieval refinement as future Engram fine-tuning target
MIA (Qiao et al., 2026)	Research agents	Manager-Planner-Executor; process-oriented memory reuse	None	Cognitive multi-agent decomposition pattern
LightMem-Z† (Zhang et al., 2026b)	Long-context QA	Online SLM + offline LLM; controller/selector/writer roles	None	Typed sub-query routing in Thalamus
MemFactory (Guo et al., 2026)	Memory RL training	Unified GRPO framework for CRUD memory policies	None	Future RL fine-tuning target for Engram and Hippo policies
AgenticLOCI (ours)	**Personalized interaction (Domain-Agnostic)**	**Two-tier episodic/archive; RRF session fusion (Engram+Hippo)**	**Typed, versioned deliberation-state memory**	**Novel: dual consolidation timescales + audit-versioned longitudinal profile history**

BLEU-1; adversarial items are scored against *I don't know* so explicit refusal earns credit.

LongMemEval-S (Wu et al., 2024) presents 500 questions over a haystack of ≈48 sessions per item. Stage 1 reports Recall@K and MRR over retrieved sessions; Stage 2 QA accuracy uses the paper's category-specific LLM-judge prompts. Question types: knowledge-update (KU), multi-session (MR), single-session-user (SS-U), single-session-assistant (SS-A), preference (Pref), temporal-reasoning (TR).

PersonaMem (Jiang et al., 2025) provides 589 multiple-choice questions testing preference recall, preference evolution tracking, and preference-aligned recommendations over long shared conversation histories (32K-token split). Accuracy is strict-protocol MCQ scoring.

C.2 Pipeline Descriptions

BM25 (Robertson et al., 1994) provides session-level lexical retrieval. **Oracle** feeds only gold evidence sessions to the reader (upper bound). **Full-context** provides the entire unfiltered conversation history. **Engram** (STM) ranks 3-turn session summaries via a single LLM call at query time. **Hippo heuristic** (LTM) scores keyword-lobe overlap at zero LLM cost; two lobe taxonomies are evaluated (domain-specific mobility and LongMemEval general-domain; see Appendix F.4). **AgenticLOCI Retrieval** fuses Engram and Hippo rankings via RRF (Eq. 1). **AgenticLOCI Backbone** extends AgenticLOCI Retrieval with batched Flashbulb-style fact extraction: summary-mode ingestion followed by one LLM call to extract cross-session facts injected into the reader prompt. **Deliberation Profile** (direct extraction) compresses the entire conversation into a preference-centric structured representation and answers from that. All Hippo and Backbone conditions in the primary 7B evaluation use LongMemEval-appropriate lobe categories.

D Extended Evaluation Tables

This appendix provides extended results complementing the main evaluation. Sections D.1–D.5 use Qwen 2.5-7B-Instruct; Sections D.4 and D.6–D.8 use Qwen 2.5-32B-Instruct.

D.1 LoCoMo Extended Results

Table 7 reports per-category BLEU-1 alongside F1, confirming the same ranking pattern. Table 8 reports exact LLM token costs per item. The AgenticLOCI Backbone's 7.3× token reduction over the full-context baseline (3,072 vs. 22,340 average input tokens) holds across all 1,986 items.

D.2 PersonaMem Extended Results

Table 9 reports accuracy by question type. Recall and fact-retrieval types are most variable across pipelines; preference evolution and reasoning types are more uniformly high. Across all 13 topics, two dataset-level floor/ceiling effects dominate: `foodRecommendation` scores 1.000 and `therapy` scores 0.000 for every pipeline, reflect-

Table 7: LoCoMo F1 and BLEU-1 by category ($n = 1{,}986$, Qwen 2.5-7B). **Bold** = best AgenticLOCI condition per column.

	Overall		F1 by category				
Pipeline	**F1**	**BLEU-1**	**Multi-Hop**	**Temporal**	**Open-Domain**	**Single-Hop**	**Adversarial**
Oracle (upper bound)	0.536	0.556	0.096	0.172	0.624	0.363	0.836
Full-context baseline	0.370	0.378	0.051	0.080	0.390	0.185	0.729
BM25 + QA	0.285	0.268	0.035	0.042	0.162	0.104	0.861
AgenticLOCI Retrieval (Engram+Hippo, Top-5)	0.311	0.319	0.028	0.043	0.193	0.113	0.913
AgenticLOCI Backbone	**0.342**	**0.351**	**0.049**	**0.059**	**0.268**	**0.167**	0.859
Deliberation Profile (direct extraction)	0.227	0.226	0.028	0.001	0.007	0.018	**0.981**

Table 8: LoCoMo token cost per item (exact, from OpenAI usage field). Framework-internal LLM calls (Engram ranker, Hippo RLM) are not surfaced in usage and show —.

Pipeline	**LLM/item**	**Avg input tokens**	**Avg output tokens**	**Avg total tokens**
BM25 + QA	1.0	2,418	10	2,429
Oracle context	1.0	1,176	6	1,183
Full-context baseline	1.0	22,340	7	22,347
AgenticLOCI Retrieval	2.0	4,301	6	4,308
AgenticLOCI Backbone	2.0	3,072	6	3,079
Deliberation Profile (direct)	2.0	884	11	895

ing benchmark-level saturation and unsolvability respectively rather than any difference between memory architectures.

D.3 Full Stage 1 Retrieval Metrics (LongMemEval-S, $n = 500$)

Table 10 extends the main retrieval table with R@3, R@10, Rany@5, Rany@10, P@1, P@5, and includes all evaluated pipeline conditions. BM25 session-level remains the reference at R@5 = 0.909. The `Per-session Flashbulb` ablation ($n = 30$) is a production-fidelity ablation: turn-level ingestion fragments Engram's ranker output, causing R@1 = 0.000 (gold sessions are present within the top-10 but not at rank 1). The `AgenticLOCI Backbone` achieves the best AgenticLOCI R@1 (0.393) and MRR (0.741) by using summary-mode ingestion rather than turn-level.

Table 11 provides per-type R@5 breakdowns. The SS-A/SS-U asymmetry between Engram and Hippo (Engram stronger on SS-A; Hippo stronger on SS-U and KU) persists at 7B. The AgenticLOCI Backbone achieves balanced coverage across SS-U, SS-A, and KU by combining both layers' strengths through summary-mode ingestion and batched fact extraction.

D.4 Lobe Vocabulary Ablation: Per-Type R@5

This ablation was conducted with Qwen2.5-32B-Instruct ($n = 470$, abstention excluded); full 32B LongMemEval-S retrieval and QA results appear in Appendix D.6. The lobe vocabulary comparison is model-agnostic: keyword overlap scoring does not depend on the reader model, so the relative gains reported here transfer directly to the 7B evaluation setting.

Table 13 presents the per-type R@5 comparison between domain-specific mobility lobe categories and the LongMemEval general-domain categories for both Hippo and Combined. The largest gain from LME lobes is on SS-A for Hippo (+12.5 pp): mobility keywords cover predominantly user-stated travel and schedule content, making assistant-stated facts nearly invisible to keyword scoring, while the general-domain categories include topics covering assistant-initiated statements (`media_entertainment`, `work_career`, `health_wellness`). The TR type gains +6.8 pp in Combined, reflecting better temporal keyword coverage in the general-domain lobe set. The only category that decreases for Combined is KU (-2.1 pp): mobility keywords happen to overlap with some knowledge-update sessions by coincidence, scheduling and travel-change events are common in LongMemEval's KU items and replacing them with broader general-domain categories removes this incidental match. Across every R@K and NDCG metric, LME lobes yield higher overall performance, making this the correct primary reported setting.

Table 9: PersonaMem accuracy by question type ($n = 589$, 32K split, Qwen 2.5-7B). Column abbreviations follow PersonaMem naming: Recall = recall of user-shared facts; Facts = recalling facts mentioned by user; Evolution = track full preference evolution; Reasons = recalling reasons behind updates; RecAlign = preference-aligned recommendations; NewIdeas = suggest new ideas; Generalize = generalizing to new scenarios.

Pipeline	Overall	Recall	Facts	Evolution	Reasons	RecAlign	NewIdeas	Generalize
Full-context baseline	0.523	0.310	0.588	0.755	0.818	0.509	0.183	0.474
AgenticLOCI Retrieval (Engram+Hippo)	0.489	0.333	0.529	0.676	0.808	0.473	0.151	0.386
Deliberation Profile (direct)	0.477	0.395	0.294	0.604	0.778	0.436	0.183	0.404

Table 10: Full Stage 1 retrieval metrics on LongMemEval-S ($n = 500$, Qwen 2.5-7B). **Bold** = best AgenticLOCI (non-BM25) result per column. LLM = LLM calls per query at retrieval time.

Pipeline	R@1	R@3	R@5	R@10	Rany@5	Rany@10	MRR	P@1	P@5	LLM
BM25 (session-level)	0.551	0.858	0.909	0.946	0.962	0.982	0.909	0.866	0.333	0
Engram (STM)	0.335	0.480	0.545	0.631	0.756	0.812	0.636	0.540	0.193	1
Hippo heuristic (LME lobes)	0.336	0.649	**0.744**	0.850	0.844	0.912	0.689	0.566	0.278	0
Hippo + RLM REPL	0.333	0.608	0.699	0.798	0.812	0.876	0.671	0.558	0.258	1
AgenticLOCI Retrieval (Eng+Hippo)	0.378	0.616	0.676	0.726	0.840	0.878	0.727	0.630	0.246	1
AgenticLOCI Retrieval + RLM	0.377	0.612	0.669	0.726	0.834	0.878	0.726	0.630	0.245	2
AgenticLOCI Backbone	**0.393**	**0.621**	0.680	**0.732**	**0.844**	**0.878**	**0.741**	**0.650**	**0.250**	1
Per-session Flashbulb ($n = 30$)	0.000	0.300	0.633	0.800	0.633	0.800	0.212	0.000	0.127	—

D.5 Full Stage 2 QA Accuracy (LongMemEval-S, $n = 500$)

Table 14 reports QA accuracy for all pipeline conditions. Hippo heuristic (QA = 0.424) is the best single AgenticLOCI component at 7B, matching the pattern in retrieval. AgenticLOCI Retrieval (0.398) and AgenticLOCI Backbone (0.386) perform comparably; the Backbone's slight QA disadvantage relative to plain retrieval reflects that batched fact extraction adds tokens that occasionally dilute the reader's attention, while MRR and R@1 advantages show it selects better sessions overall. AgenticLOCI Retrieval + RLM ($n = 499$, 0.387) does not improve over the heuristic, consistent with the LongMemEval finding that RLM-generated lobes do not benefit the reader at this scale. Preference QA is near-zero for all systems including BM25, establishing a benchmark-level floor.

D.6 Qwen2.5-32B-Instruct Results (LongMemEval-S)

Tables 15 and 16 reproduce the full LongMemEval-S evaluation run with Qwen2.5-32B-Instruct ($n = 470$ retrieval, $n = 500$ QA), the model used for the lobe ablation (Appendix D.4), error analysis (Appendix D.8), and the LongMemEval-M experiment (Appendix D.7). These results predate the primary 7B evaluation and are included for cross-model comparison. "Combined heuristic (LME lobes)" is equivalent to AgenticLOCI Retrieval under the 7B nomenclature. The 32B run achieves Combined R@5 = 0.855 vs. 0.676 at 7B, confirming that RRF complementarity strengthens with model capacity.

D.7 LongMemEval-M Results (Qwen2.5-32B)

LongMemEval-M ($\approx$500 haystack sessions per item vs. $\approx$48 in _S) enables direct comparison with Table 9 of Wu et al. (2024) and tests whether the STM+LTM complementarity observed on _S holds at larger scale. For _M, the Combined pipeline uses two-stage retrieval: BM25 pre-selection of the 50 most relevant sessions (necessary because all session summaries combined exceed Qwen2.5-32B's 32K token context limit), followed by Engram LLM ranking over the pre-selected candidates. Hippo scores the full 500-session haystack unaffected.

Table 17 shows that BM25 achieves R@5 = 0.774 on _M, substantially higher than the paper's BM25 K=V result of 0.634 (Table 9 of Wu et al. 2024); this gap is attributable to implementation differences in tokenization rather than a methodological discrepancy. Combined (LME lobes) reaches R@5 = 0.683, below _M BM25 (0.774) and below _S Combined (0.855). The drop is consistent with the design constraint: the BM25 pre-filter means Engram ranks only BM25-selected candidates, so its semantic signal is no longer independent of BM25, and Hippo's keyword scores over 500 sessions contribute more noise than they do at _S scale. QA on _M shows Combined accuracy of 0.442, below the _M BM25 upper bound of 0.520.

Table 11: R@5 by question type on LongMemEval-S ($n = 500$, Qwen 2.5-7B). **Bold** = best AgenticLOCI (non-BM25) result per column.

Pipeline	KU	MR	SS-U	SS-A	Pref	TR
BM25 (session-level)	0.987	0.842	0.986	0.982	0.800	0.882
Engram (STM)	0.526	0.504	0.557	**0.714**	0.433	0.546
Hippo heuristic (LME lobes)	**0.853**	0.724	**0.871**	0.554	0.467	**0.777**
Hippo + RLM REPL	0.840	0.665	0.871	0.429	0.467	0.726
AgenticLOCI Retrieval	0.750	**0.645**	0.714	0.714	**0.533**	0.658
AgenticLOCI Retrieval + RLM	0.750	0.643	0.686	0.696	0.533	0.658
AgenticLOCI Backbone	0.756	0.650	0.714	0.732	0.467	0.672

Table 12: MRR by question type on LongMemEval-S ($n = 500$, Qwen 2.5-7B). **Bold** = best AgenticLOCI (non-BM25) result per column.

Pipeline	KU	MR	SS-U	SS-A	Pref	TR
BM25 (session-level)	0.977	0.923	0.937	0.982	0.584	0.882
Engram (STM)	0.751	0.660	0.479	**0.657**	0.335	0.686
Hippo heuristic (LME lobes)	0.825	0.757	0.706	0.465	0.358	0.701
Hippo + RLM REPL	0.814	0.736	0.732	0.346	0.350	0.698
AgenticLOCI Retrieval	0.835	**0.788**	0.615	0.599	0.351	**0.802**
AgenticLOCI Retrieval + RLM	**0.835**	0.792	0.601	0.592	**0.364**	0.800
AgenticLOCI Backbone	0.851	0.786	**0.638**	0.634	0.351	0.820

The reversal from _S is entirely attributable to the pre-filter degrading complementarity.

D.8 Item-Level Error Analysis

This analysis uses Qwen2.5-32B-Instruct retrieval runs ($n = 470$, abstention excluded). The complementarity patterns described are architectural and replicate under the 7B evaluation.

Table 18 breaks down, per question type, the 470 non-abstention items by which layer retrieves the answer session first (R@1 = 1) while the other fails (R@1 = 0). All figures use Hippo heuristic with LongMemEval lobe categories and Engram only (STM), evaluated independently.

Three patterns stand out. First, on *knowledge-update* (KU), there are **zero** items where one system wins R@1 and the other fails. Both layers succeed and fail on exactly the same KU questions, confirming that Hippo's +18 pp R@5 advantage over Engram on this type is purely a *coverage* effect rather than evidence that Hippo systematically identifies updated sessions better at rank 1. Second, Engram's dominance on *single-session-assistant* (27 of 65 Engram-only wins) is structurally explained: Hippo ingests only user turns during consolidation, making assistant-stated facts largely invisible to keyword scoring. Engram's LLM ranker comprehends the full session exchange and maps follow-up questions to the correct prior session. Third, Hippo's wins on *preference* (6 items) and *single-session-user* (10 items) share a common pattern: the correct session contains a specific, rare token, a proper noun, a precise figure, a brand name, that appears in exactly one session in the haystack.

The 23 items where both layers fail simultaneously decompose as follows: implicit preferences aggregated across sessions without explicit statement (7), assistant content with sparse keyword signal (6), temporal questions requiring date arithmetic rather than session identification (4), sparse biographical facts mentioned only once in passing (3), and multi-session questions whose evidence spans disjoint sessions that no top-5 window covers (3). The preference and biographical failures directly motivate the Longitudinal Deliberation Profile as a proactive store for facts the retrieval stack cannot reliably surface.

E Consolidation Error Breakdowns

E.1 Flashbulb Failure Analysis

Table 19 decomposes the Flashbulb misses by axis and type, complementing the aggregate results in §4.2.1.

E.2 Longitudinal Profile Violation Breakdown

Table 20 lists the five unwarranted profile changes made by full Hippo across the three users with errors, complementing the aggregate results in §4.2.2. We also examined the four *protected* snapshot blocks (deliberation dynamics, mode activa-

Table 13: Lobe vocabulary ablation: R@5 by question type on LongMemEval-S ($n = 470$, abstention excluded). Run with Qwen2.5-32B-Instruct (see Appendix D.6 for full 32B results). Δ = LME lobes $-$ mobility lobes (signed). Largest per-type gains in **bold**.

Pipeline / Δ	**Overall**	**KU**	**MR**	**SS-U**	**SS-A**	**Pref**	**TR**
Hippo (mobility lobes)	0.705	0.868	0.676	0.891	0.429	0.467	0.725
Hippo (LME lobes)	0.747	0.868	0.720	0.906	0.554	0.467	0.774
Δ	+0.042	0.000	+0.044	+0.016	**+0.125**	0.000	+0.049
Combined (mobility lobes)	0.834	0.910	0.817	0.906	0.839	0.733	0.790
Combined (LME lobes)	0.855	0.889	0.832	0.922	0.839	0.733	0.858
Δ	+0.021	-0.021	+0.015	+0.016	0.000	0.000	**+0.068**

Table 14: Full Stage 2 QA accuracy on LongMemEval-S ($n = 500$, Qwen 2.5-7B reader and judge). AgenticLOCI Retrieval + RLM $n = 499$ (one item no valid response). **Bold** = best AgenticLOCI (non-BM25) condition per column.

Pipeline	**Overall**	**KU**	**MR**	**SS-U**	**SS-A**	**Pref**	**TR**
BM25 (session-level)	0.496	0.603	0.338	0.900	0.911	0.200	0.271
Engram (STM)	0.344	0.385	0.180	0.614	0.732	0.233	0.203
Hippo heuristic (LME lobes)	**0.424**	0.526	**0.293**	**0.843**	0.554	0.200	**0.271**
Hippo + RLM REPL	0.399	**0.590**	0.212	0.857	0.446	0.267	0.241
AgenticLOCI Retrieval	0.398	0.538	0.211	0.729	0.679	**0.300**	0.233
AgenticLOCI Retrieval + RLM	0.387	0.526	0.211	0.710	**0.643**	0.267	0.233
AgenticLOCI Backbone	0.386	0.513	0.158	0.757	0.732	0.133	0.256
Per-session Flashbulb ($n = 30$)	0.633	—	—	0.633	—	—	—

tions, valences, stressor tolerances) that Hippo is instructed not to overwrite. The keyword path preserves all four blocks for all ten users (40 of 40), because its write surface is restricted in code. The LLM path preserves 33 of 40, with five users receiving a small write into a protected block. The lesson for deployment: code-level write-surface restrictions are more reliable than prompt-level instructions to leave a region untouched; a production system should enforce both.

F Prompts and Implementation Details

This section documents the prompt designs and implementation choices for AgenticLOCI agents, complementing the architecture description in Section 3. All prompts are domain-supplied via the DomainPack contract; the anonymized mobility domain is used as the reference instantiation here.

F.1 Broca: Conversational Agent Prompt Design

Broca's system prompt encodes five design principles. (1) **Synthesize, don't list:** Broca commits to a single recommendation rather than presenting options; bullet lists of transport modes are explicitly forbidden. (2) **Translate internal signals to natural language:** deliberation-profile fields are never shown to the user; the prompt maps each field to a natural phrase (*"speed_need high"* $\rightarrow$ *"you don't like wasting time"*; *"avoid_traffic flag"* $\rightarrow$ *"you hate sitting in traffic"*). (3) **Flashbulb-first filtering:** before recommending a mode, Broca hard-filters on active Flashbulb constraints ("bike stolen $\rightarrow$ drop cycling"). (4) **Machine-readable route tag:** every routing decision appends `<route mode="..." origin="..." destination="..."/>` parsed server-side to render a map card; users never see the tag. (5) **Confidence-calibrated closers:** the final line varies with deliberation confidence (confidence > 0.4: *"Sound good?"*; 0.2–0.4: *"Sound good, or want a different angle?"*; < 0.2: *"That work for you, or is there something I'm missing?"*).

The routing pipeline encoded in the prompt runs five steps in order: (i) filter hard by Flashbulb constraints, (ii) fetch live MCP data (weather, traffic, parking, transit) in parallel, (iii) synthesize one mode from the deliberation profile, live data, and temporal context, (iv) write a single reasoning-trail paragraph in natural language only, (v) append the `<route/>` tag.

F.2 Flashbulb: Sleep-Time Consolidation Prompt

The Flashbulb consolidation prompt runs at session end, triggered by Thalamus. Its core design heuristic is a *"weird test"*: *would it be helpful or weird if*

Table 15: Stage 1 retrieval on LongMemEval-S ($n = 470$, abstention excluded; Qwen2.5-32B-Instruct). **Bold** = best non-BM25 result per column.

Pipeline	R@1	R@5	R@10	Rany@5	NDCG@5	MRR	P@1	LLM
BM25 (session-level)	0.558	0.913	0.948	0.962	0.883	0.909	0.870	0
Engram (STM)	0.473	0.679	0.703	0.843	0.688	0.794	0.751	1
Hippo heuristic (mobility lobes)	0.332	0.705	0.802	0.815	0.626	0.662	0.553	0
Hippo heuristic (LME lobes)	0.337	0.747	0.854	0.843	0.660	0.684	0.566	0
Hippo + RLM REPL	0.337	0.704	0.808	0.819	0.627	0.667	0.562	1
AgenticLOCI Retrieval (mobility lobes)	0.504	0.834	0.888	0.926	0.801	0.857	0.811	1
AgenticLOCI Retrieval (LME lobes)	**0.509**	**0.855**	**0.902**	**0.936**	**0.817**	**0.866**	**0.823**	1
AgenticLOCI Retrieval + RLM REPL	0.468	0.773	0.827	0.870	0.747	0.803	0.747	2

Table 16: Stage 2 QA accuracy on LongMemEval-S ($n = 500$; Qwen2.5-32B-Instruct reader and judge). **Bold** = best non-BM25 condition per column. AgenticLOCI Retrieval + RLM $n = 499$.

Pipeline	Overall	KU	MR	SS-U	SS-A	Pref	TR
BM25 (session-level)	0.528	0.564	0.466	0.929	0.946	0.067	0.286
Engram (STM)	0.384	0.321	0.271	0.729	0.768	0.133	0.248
Hippo heuristic (LME lobes)	0.460	0.641	0.353	0.871	0.571	0.000	0.301
AgenticLOCI Retrieval (LME lobes)	**0.514**	**0.564**	**0.459**	**0.871**	**0.821**	0.067	**0.323**
AgenticLOCI Retrieval (mobility lobes)	0.528	0.628	0.451	0.914	0.821	0.100	0.316
AgenticLOCI Retrieval + RLM REPL	0.461	0.597	0.376	0.857	0.839	0.033	0.196

Broca mentioned this fact unprompted three days from now? If 'weird,' do not add it. This conservative criterion qualifies persistent transport constraints ("bike was stolen", "uses a wheelchair"), temporary health constraints with week-scale effects ("sprained ankle, two-week recovery"), recurring schedules ("class at Building 03 every Monday 9am"), and explicit user requests ("please remember ...") for Flashbulb storage. One-off route requests route to the `events` store; casual preference mentions route to `preferences`; current trip destinations and weather snapshots are discarded. The prompt returns strict JSON covering `add_events`, `remove_event_ids`, `update_events`, `events`, `preferences`, `active_intents`, and `notes`; no prose output is permitted.

F.3 Hippo: RLM Batch Consolidation Prompt

When RLM mode is enabled, Hippo makes one batch LLM call per item via a Python REPL environment (`rlms` package). The LLM is given three tools: `get_sessions()`, `get_lobe_names()`, and `get_lobe_keywords(name)`, and is instructed to iterate over sessions, match keyword occurrences per lobe, build a result dictionary `{session_id: [{name, signals:[{keyword,count}], confidence, summary}]}`, and return it via `FINAL_VAR('result_json')`. A critical implementation note: `FINAL_VAR` requires a variable name string, not an expression; calling `FINAL_VAR(json.dumps(result))` passes a JSON string as a variable name, which silently fails and falls back to heuristic mode. This prompt engineering constraint explains why RLM $\approx$ heuristic on retrieval recall (R@5 0.704 vs. 0.705): any REPL call that encounters an error falls back to the heuristic, leaving the majority of items scored identically.

F.4 Hippo Lobe Taxonomies

Hippo uses two lobe taxonomies selectable via `custom_lobe_rules`: the default domain-specific mobility taxonomy (7 lobes; Table 21) and the LongMemEval general-domain taxonomy (10 lobes; Table 22).

Both taxonomies are passed via the `custom_lobe_rules` parameter of the `Hippo` constructor; no code changes are required per domain. The SS-A retrieval gain from LME lobes (+12.5 pp; Table 13) is the clearest illustration of why domain-appropriate keywords matter.

Table 17: Stage 1 retrieval on LongMemEval-M ($n = 470$, abstention excluded; Qwen2.5-32B-Instruct). "Combined heuristic (LME lobes)" = AgenticLOCI Retrieval with LME lobe categories. Combined uses two-stage retrieval: BM25 pre-filter (top-50) then Engram LLM ranking; Hippo scores all sessions.

Pipeline	**R@1**	**R@5**	**R@10**	**NDCG@5**	**NDCG@10**	**MRR**	**P@1**	**Rany@5**
BM25 (session-level)	0.439	0.774	0.834	0.722	0.746	0.764	0.685	0.872
Combined heuristic (LME lobes)	0.435	0.683	0.729	0.662	0.680	0.741	0.681	0.821

Table 18: Item-level error analysis: where each layer wins at R@1 while the other fails, by question type ($n = 470$, abstention excluded). Hippo = Hippo heuristic (LME lobes); Engram = Engram only (STM).

Outcome	**KU**	**MR**	**SS-U**	**SS-A**	**Pref**	**TR**	**Total**
Hippo wins, Engram fails	0	0	10	3	6	2	21
Engram wins, Hippo fails	0	0	18	27	12	8	65
Both fail (R@1 and R@5 = 0)	0	3	3	6	7	4	23
Both correct (R@1 = 1)	0	0	28	17	1	4	50

Axis	Failure type	Count
Capture	Ambiguous durability (persistent weather)	4
	Third-party fact (not the user's own state)	2
	One-off or inferred event	7
Expire	Implicit resolution (pin left in place)	8
	Removed one day early	2

Table 19: Where Flashbulb's errors fall. Of 65 events, the 13 capture misses and 10 expire misses concentrate on the deliberately hard cases; routine stated facts are handled reliably.

User	Incorrect change	Magnitude
2	environmental need collapsed	-0.22
3	environmental need inflated	$+0.09$
3	physical-activity need inflated	$+0.10$
9	`avoid_traffic` flipped on	flag
9	`prefer_covered_paths` on	flag

Table 20: The five profile violations by full Hippo; the other seven users had none. Each is an inspectable, reversible entry in the version history.

F.5 Engram: LLM-Based Session Retrieval

Engram's `process_query` constructs a single prompt at query time containing all stored session summaries and asks the LLM to rank sessions by relevance. No separate prompt file is used; the prompt is assembled inline from stored entries and the query. In evaluation, sessions are ingested in `summary` mode: the first three user turns of each session are concatenated into a single entry via `record_turn`. On LongMemEval-S (≈48 sessions/item), this yields approximately $48 \times 50 = 2{,}400$ prompt tokens, comfortably within Qwen2.5-32B's 32K limit.

On LongMemEval-M (≈500 sessions/item), the same approach would exceed the 32K limit. A BM25 pre-filter selecting the top-50 candidate sessions is therefore applied before the Engram LLM call. This design decision is the root cause of the performance reversal on _M (Appendix D.7): the pre-filter makes Engram's signal correlated with BM25, breaking the independence assumption behind RRF. Replacing Engram's LLM ranker with a dense embedding model (no context-length constraint) would restore the independent-signal property at _M scale and is an explicit target for future work.

F.6 LongMemEval QA Judge Configuration

QA accuracy is evaluated using the exact per-type judge prompts from (Wu et al., 2024), sourced without modification. Each item receives the question, the ground-truth answer rubric, and the system's generated answer; the judge returns binary correct/incorrect with type-specific criteria. The knowledge-update type requires identifying the *updated* fact, not the original; single-session types (SS-U, SS-A) require evidence from the correct session.

The primary model difference from the paper's results is judge model: we use **Qwen2.5-32B-Instruct** (local deployment on an institutional GPU server via vLLM), while the paper uses **GPT-4o** (`gpt-4o-2024-08-06`). We use Top-5 retrieved sessions as context; the paper's 0.700 figure uses Top-10. The most model-matched comparison in the paper is Llama 3.1 70B reader, Top-5, LongMemEval-M = 0.588, placing our 0.514 (LME-S) approximately 7 pp below that reference. The paper further reports Chain-of-Note prompting gaining up to 10 pp on oracle retrieval; we use standard prompts throughout.

Table 21: Domain-specific mobility lobe taxonomy (default, anonymized deployment). Each lobe maps to a longitudinal deliberation-profile dimension.

Lobe	Purpose / Sample Keywords
`mobility`	Transport mode signals: bike, tram, bus, route, commute, parking
`schedule_routines`	Regular temporal patterns: every, weekly, schedule, monday–friday
`health_constraints`	Physical capability limits: hurt, injury, wheelchair, fatigue, sprain
`places_familiarity`	Location preferences: home, office, campus, avoid, prefer
`social_contacts`	People and appointments: family, partner, colleague, doctor
`preferences_values`	Decision-making values: prefer, hate, cost, fast, safe, eco
`transient_changes`	Time-limited states: today, this week, temporary, recently, just

Table 22: LongMemEval general-domain lobe taxonomy (`-lobe-set longmemeval`). Used only in benchmark evaluation; covers the 164-attribute user ontology of Wu et al. (2024).

Lobe	Sample Keywords
`life_events`	graduated, married, promoted, hired, moved, surgery, retired
`possessions`	bought, car, bike, phone, laptop, pet, dog, house, instrument
`hobbies_interests`	cooking, gardening, photography, gaming, hiking, yoga, swimming
`media_entertainment`	book, movie, show, podcast, music, album, stream, netflix
`food_dining`	restaurant, recipe, vegetarian, coffee, barbecue, diet
`work_career`	job, boss, meeting, deadline, salary, promotion, company
`health_wellness`	doctor, hospital, medication, exercise, injury, therapy, stress
`preferences_tastes`	favorite, love, hate, prefer, enjoy, recommend, avoid
`social_relationships`	friend, family, partner, spouse, colleague, daughter, neighbor
`travel_places`	travel, vacation, visit, hotel, flight, beach, abroad, camping