Soul: Constructive Memory for Persona-Centric AI

Soul Research
research@soul.app

Abstract

We present SOUL, a memory architecture for persona-centric AI that achieves a new state-of-the-art (SOTA) result on the PersonaMem benchmark with 67.2% accuracy, surpassing the best published result (52%, full-context prompting) by 15 percentage points while consuming less than 1% of the context window. SOUL treats memory as constructive rather than retentive: it extracts interpreted impressions instead of atomic facts, decays low-saliency details via saliency-weighted forgetting, and tracks uncertain inferences as revisable hypotheses. We also introduce a 90-scenario Cognitive Memory Evaluation spanning six dimensions grounded in cognitive psychology, where SOUL scores 74.5% versus 64.6% for the next-best system. On LongMemEval-S (factual recall), SOUL scores 75%, below retrieval-optimized systems (95–99%); ablations trace this gap to interpretive extraction, which drops cognitive evaluation scores by 12.1 points when disabled.

1 Introduction

Memory is essential for AI that interacts with the same user over time. Most current memory systems default to persistent retention: they extract facts, embed them, and retrieve by similarity. Even systems that track temporal metadata and handle contradictions inherit this retention bias. The result is three characteristic failure modes in long-term user modeling: low-saliency observations persisting as apparent preferences (a passing mention of pasta resurfacing as a defining interest), outdated preferences coexisting with their replacements, and uncertain inferences hardening into stated facts. As Packer et al. [1] observe, the challenge is not storing more but storing *better*. Recent empirical work reinforces this point: Engramme [2] collected 1,940 memory questions from 134 participants during daily activities and found that the most common questions concerned participants' own past actions and personal context, not retrievable facts. Categories already well-served by dedicated tools (passwords, schedules, contacts) accounted for a minority of questions, while personal, interpretive needs remained largely unaddressed. This suggests that the memory problem for AI systems is less about factual fidelity and more about modeling personal context.

We investigate whether constructive memory (selective, interpretive, and revisable) improves persona fidelity under tight context budgets, and at what cost to factual recall. We present SOUL, a memory architecture for persona-centric AI. Fuzzy-trace theory posits that humans encode both verbatim and interpretive (gist) traces in parallel, with gist traces decaying more slowly [3]. SOUL operationalizes the gist branch of this dual-trace model, deliberately prioritizing interpreted impressions of *who someone is* over verbatim content. Each memory carries a saliency score (1–5) governing token-based decay, with the decay curve shape inspired by Ebbinghaus [4] and the adaptive function of forgetting by Anderson and Schooler [5], so that low-saliency observations fade while reinforced patterns persist. Uncertain inferences are tracked as hypotheses that can be upgraded through subsequent evidence, addressing epistemic overconfidence related to what Schacter [6] terms suggestibility and bias. Related memories consolidate through merge operations inspired by episodic-to-semantic consolidation [7], and contradictions trigger edits rather than duplications.

Preprint.

We evaluate SOUL on three benchmarks probing complementary aspects of memory quality:

1. **LongMemEval-S** [8]: 500 factual recall questions, where SOUL scores 75%, below retrieval-optimized systems ($\sim$95–99%).
2. **PersonaMem** [9]: 2,727 persona-centric questions, where SOUL achieves 67.2%, surpassing the best published full-context result ($\sim$52%) by 15 percentage points while consuming less than 1% of the context window.
3. **Cognitive Memory Evaluation** (introduced here): 90 scenarios spanning six dimensions grounded in cognitive psychology, where SOUL scores 74.5% compared to 64.6% for the next-best system.

Our contributions are:

1. A memory architecture implementing interpretive extraction, saliency-weighted decay, hypothesis tracking, memory consolidation, and contradiction handling.
2. 67.2% on PersonaMem (n=2,727), compared to 52% for published full-context baselines, using $\sim$500 tokens of extracted memories at query time.
3. A 90-scenario evaluation across six cognitive dimensions, designed to test capabilities (forgetting, epistemic calibration, noise resistance) that existing benchmarks do not measure.
4. Ablations isolating the contribution of each mechanism, with interpretive extraction as the dominant factor (-12.1 points when disabled) and saliency decay providing a $2\times$ reduction in noise pollution.

2 Related work

Fact-centric and structured memory. Mem0 [10] is a long-term memory layer that dynamically extracts, consolidates, and retrieves salient conversational facts, with add / update / delete operations and a graph-backed variant for richer relational reasoning. Supermemory [11, 12] similarly builds a fact-level memory graph with update, extend, and derive relations, plus time-based forgetting for temporary facts and noise filtering. A-MEM [13] pushes this direction further by dynamically indexing, linking, and evolving structured notes in an agentic memory network. These systems substantially improve factual continuity and update handling, but they remain centered on explicit, queryable memory objects rather than interpretive persona modeling or calibrated hypotheses.

Compression and reflective memory. Mastra's Observational Memory [14, 15] compresses message history into dense observation logs, then uses a reflector to condense and garbage-collect less important observations; retrieval mode is optional rather than foundational. Reflective Memory Management (RMM) [16] likewise uses prospective and retrospective reflection to build personalized memories across utterance-, turn-, and session-level granularities, while LightMem [17] applies a three-stage sensory / short-term / long-term pipeline with lightweight filtering and sleep-time updates. These systems move beyond naive storage by summarizing and reorganizing memory, but their goal is still largely long-horizon task performance rather than persona-centric memory.

OS-style memory management. MemGPT [1] reconceptualizes LLM memory as an operating system with explicit paging between working memory and archival storage. This architectural metaphor enables sophisticated memory management but focuses on context window optimization rather than modeling cognitive processes like forgetting or belief formation.

Human-inspired memory architectures. EM-LLM [18] organizes long inputs into episodic events using Bayesian surprise and graph-theoretic boundary refinement, then retrieves them with similarity and temporal contiguity. Hindsight [19] is closer to our problem setting: it separates world facts, experiences, synthesized entity summaries, and evolving beliefs, and couples retention with reflection. Both are important steps toward more cognitively grounded memory. However, they focus on long-horizon retrieval and explicit reasoning over memory banks rather than saliency-weighted forgetting, uncertainty stored as hypotheses, or interpretive persona impressions.

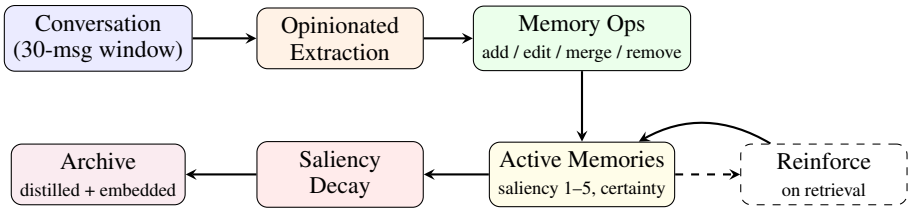

Figure 1: Memory lifecycle in SOUL. Conversations are processed through opinionated extraction, producing memory operations (add, edit, merge, remove) that update the active memory store. Memories decay based on saliency; high-saliency memories are distilled before archival. Retrieval reinforces accessed memories, slowing their decay.

Cognitive science foundations. The cognitive principles underlying SOUL's design (constructive remembering [20], adaptive forgetting [5, 6] with decay dynamics from [4], gist-based encoding [3], episodic-to-semantic consolidation [7], and retrieval-based reinforcement [21]) are introduced alongside the corresponding mechanisms in Sections 1 and 3. While several systems above draw on individual cognitive constructs, none combines saliency-weighted decay, epistemic uncertainty, and interpretive extraction in a single pipeline.

3 System architecture

SOUL implements a memory lifecycle with six mechanisms that process conversations in 30-message sliding windows, updating the user model through add, edit, merge, and remove operations. Each session contributes incrementally to the model. Table 1 summarizes the mechanisms; Figure 1 illustrates the lifecycle.

Table 1: The six core mechanisms in SOUL and their design motivation.

Mechanism	What it does	Design motivation
Opinionated extraction	Stores interpreted impressions, not fact fragments	Gist-based encoding produces stronger personalization signal [3]
Saliency-weighted forgetting	Low-importance details decay; repeated patterns persist	Forgetting as adaptive noise filtering [5, 6]; decay shape from [4]
Hypothesis tracking	Uncertain beliefs stored as questions, not facts	Prevents epistemic overconfidence in inferred beliefs [6]
Memory consolidation	Related observations merge into deeper understanding	Builds schema-level knowledge from scattered episodes [20]
Contradiction handling	Preferences are edited or removed when they change	Keeps the active memory aligned with the current user state
Memory retrieval	Ranks active and archived memories by relevance, saliency, and recency	Retrieval reinforces accessed memories, implementing the testing effect [21]

3.1 Opinionated extraction

The extraction prompt is the single most impactful mechanism: disabling it reduces cognitive evaluation scores by 12.1 percentage points, as shown in our ablation study (Section 5.5). Rather than extracting atomic facts ("user mentioned they like running"), the prompt elicits interpreted impressions. An excerpt:

> *After talking with [name], you're sitting with your thoughts. What stayed with you? What are you still turning over?* [...] *This isn't a transcript. It's how a close friend actually remembers conversations, colored by feeling, shaped by interpretation, and full of half-formed theories and lingering questions.*

The prompt instructs the model to capture *who someone is*, not what they said. Memories are phrased as opinionated interpretations:

```
{
  content:       "she's way too hard on herself about the art stuff"
  saliency:      4                // 1=fleeting, 5=core identity
  certainty:     "known"          // or "hypothesis"
  reasoning:     "she brushed off the work again despite obvious care"
  curious_about: ["has she always been this hard on herself?"]
}
```

Figure 2: Memory object structure. Each memory stores an interpreted impression (`content`), its importance (`saliency`, Table 2), epistemic status (`certainty`), the extraction model's reasoning, and unresolved questions for future conversations.

Factual extraction	**Opinionated extraction (SOUL)**
"User mentioned they like running three times per week"	"She's the type who needs running to stay sane; less a hobby, more a coping mechanism"
"User said they are applying to grad school"	"She's seriously thinking about going back to school"

This design is inspired by the gist branch of fuzzy-trace theory [3], which holds that interpreted traces are more durable than verbatim ones. SOUL deliberately extracts only gist-level representations, trading verbatim fidelity for interpretive depth.

The extraction prompt produces structured **memory operations**: add, edit, remove, and merge, rather than a flat list of facts. A typical 30-message window produces 0–2 new memories; edits and merges are unlimited. This constraint prevents over-extraction and forces the model to prioritize signal. Figure 2 shows the resulting memory object structure.

3.2 Saliency-weighted forgetting

Inspired by Ebbinghaus's forgetting curve, each memory carries a saliency score from 1 to 5. Saliency governs token-based decay: a memory's lifespan is approximately saliency $\times$ 4,000 tokens of subsequent conversation. Table 2 defines the levels.

Table 2: Saliency levels and their decay behavior. Lifespan is approximate tokens of subsequent conversation before expiry.

Level	**Description**	**Lifespan**	**Example**
1	Fleeting observation	$\sim$4K tokens	"had a veggie wrap for lunch"
2	Minor habit or context	$\sim$8K tokens	"green tea has replaced coffee lately"
3	Real insight about identity	$\sim$12K tokens	"running helps her stay sane"
4	Significant pattern	$\sim$16K tokens	"going back to school feels serious now"
5	Core identity, rare	$\sim$20K tokens	"15-year vegetarian for ethical reasons"

When a memory's token budget expires, its fate depends on saliency. Low-saliency memories (1–2) are deleted. High-saliency memories (4–5) are *distilled*, condensed to their essence, and moved to an archival store with vector embeddings for future retrieval. This mimics the consolidation of episodic memory into semantic knowledge [7].

Critically, memories can be *reinforced*. When a topic resurfaces in conversation and a related memory is retrieved, its decay clock resets, implementing the testing effect [21]. Sustained interests persist; one-off mentions fade.

The extraction prompt defaults new memories to saliency 2, reserving higher scores for patterns confirmed across multiple interactions. This conservative assignment means most observations must earn their persistence through reinforcement.

3.3 Hypothesis tracking

SOUL maintains calibrated epistemic uncertainty by distinguishing `known` memories (confident observations) from `hypothesis` memories (uncertain inferences phrased as questions). For example:

Certainty	Example
`known`	"She's way too hard on herself about the art stuff"
`hypothesis`	"Is she actually okay or just saying that?"

Hypotheses can be upgraded to known status via an edit operation when subsequent conversations provide confirming evidence. This mechanism addresses epistemic overconfidence (treating inferred beliefs with the same confidence as directly stated facts), related to what Schacter [6] terms suggestibility and bias in memory formation.

3.4 Memory consolidation

The merge operation combines multiple related memories into a single, deeper understanding. When several scattered observations point to the same pattern, for example, "she runs 3x/week," "she does yoga for cross-training," and "she's training for a half marathon," a merge produces "she's deeply committed to fitness; it structures her entire week." This is inspired by how human memory consolidates episodic traces into semantic knowledge [7], building abstractions from scattered observations. Merges are initiated by the extraction model when it identifies convergent evidence among existing memories.

3.5 Contradiction handling

When user preferences change, SOUL performs edit or remove operations rather than appending new memories alongside outdated ones. "I love coffee" followed later by "I quit coffee for tea" results in an edited memory capturing the current state, not two contradicting entries. The extraction prompt includes both hard contradictions (explicit reversals) and soft evolution (gradual preference shifts) in its operation repertoire.

3.6 Memory retrieval

At query time, SOUL retrieves from two stores: active memories (the current user model) and archived memories (distilled, embedded past knowledge). Active memories are included directly in the prompt context. Archived memories are retrieved via cosine similarity between the query embedding and stored memory embeddings, ranked by a weighted combination of semantic relevance, saliency, and recency. This ensures that high-saliency, recently-accessed memories surface preferentially.

4 Evaluation

We evaluate SOUL on three benchmarks chosen to probe complementary aspects of memory quality: factual recall (LongMemEval-S), persona-centric personalization (PersonaMem), and cognitive plausibility (an evaluation we introduce). Competitors include Mem0 [10], Mastra OM [14], and SuperMemory [11]: three prominent open-source memory systems with distinct architectures.

4.1 LongMemEval-S

LongMemEval [8] is a standard benchmark for long-term chat memory, comprising 500 questions that evaluate factual extraction, multi-session reasoning, knowledge updates, temporal reasoning, and abstention on unanswerable questions. We include it to quantify the factual recall cost of interpretive extraction. Published scores for SuperMemory ($\sim$99%) and Mastra ($\sim$95%), both reported by the system authors, set an upper bound for retrieval-optimized systems. We evaluate both SOUL and Mem0 on the full 500-question dataset.

4.2 PersonaMem

PersonaMem [9] evaluates persona-centric memory through 2,727 multiple-choice questions ($n_{\text{128k split}}$) across seven categories: recalling user-shared facts, acknowledging latest preferences, tracking preference evolution, revisiting reasons behind updates, suggesting new ideas, making preference-aligned recommendations, and generalizing to new scenarios. The published best result is $\sim$52%, achieved by GPT-4.5, GPT-4.1, and Gemini-1.5-Flash using *full-context prompting*, which stuffs the entire conversation ($\sim$160K tokens) into the model's context window. Notably, no published result uses a memory system; the baseline is brute-force context.

We run SOUL on the full dataset (n=2,727). Competitors are evaluated on a matched 100-question subset (n=100) scattered across the full dataset (10 questions from each of 10 evenly-spaced regions), ensuring coverage of the full persona distribution.

4.3 Cognitive memory evaluation

Standard benchmarks evaluate factual recall but not forgetting appropriateness, epistemic calibration, consolidation quality, or noise resistance. This gap matters in practice: empirical studies of everyday memory needs show that users most frequently seek personal, contextual information rather than retrievable facts [2], yet no existing benchmark tests whether a memory system handles such needs appropriately. We introduce a 90-scenario evaluation spanning six dimensions, each grounded in established cognitive psychology.

4.3.1 Dimensions

1. **Consolidation quality** (15 scenarios). Tests whether scattered observations about the same pattern are merged into a coherent understanding. Grounded in Bartlett's schema theory [20]: human memory organizes related information into interconnected webs rather than storing isolated fragments.
2. **Epistemic calibration** (15 scenarios). Tests whether ambiguous signals are stored with appropriate uncertainty. Grounded in metacognitive monitoring [6]: memory includes not just what we know, but how certain we are about it.
3. **Forgetting appropriateness** (15 scenarios). Tests whether low-importance details decay while high-importance patterns persist after a delay. Grounded in Ebbinghaus's forgetting curve: retention decays exponentially without reinforcement, and this decay serves an adaptive function [5].
4. **Interpretation depth** (15 scenarios). Tests whether the system captures emotional subtext and underlying patterns rather than surface-level facts. Grounded in Brainerd and Reyna's fuzzy-trace theory: humans retain interpreted gist preferentially over verbatim content.
5. **Noise resistance** (15 scenarios). Tests whether one-off enthusiastic mentions persist and pollute future interactions after 60K tokens of intervening conversation. Directly operationalizes the noise persistence problem observed in deployed memory systems.
6. **Update hygiene** (15 scenarios). Tests whether contradicted preferences are revised rather than duplicated, covering both hard contradictions (explicit reversals) and soft evolution (gradual shifts).

Scenarios were designed to operationalize established cognitive science constructs and were reviewed by multiple authors for face validity. All 90 scenarios will be released publicly.

4.3.2 Evaluation methods

We employ three evaluation methods, matched to the nature of each dimension:

LLM rubric scoring (consolidation, interpretation, update hygiene). An LLM judge (GPT-5.1) scores the system's memory dump against weighted rubric criteria (e.g., *old_preference_not_present*: weight 3, *new_preference_present*: weight 3). Each scenario is judged three times; we take the majority score per criterion to reduce variance.

Classification accuracy (epistemic calibration). For each scenario, we identify memories matching target topics and classify whether the memory text appropriately expresses uncertainty for ambiguous

Table 3: Cross-benchmark comparison. LongMemEval scores for Mastra and SuperMemory are author-reported; all other scores are our evaluations using GPT-5.1. PersonaMem scores for competitors use a matched 100-question subset. † Published by system authors. ‡ Evaluated by us.

System	LongMemEval-S	PersonaMem	Cognitive Eval
SOUL	75.0%‡	**67.2%**‡	**74.5%**‡
Mastra OM	∼95%†	33%‡	64.6%‡
Mem0	23.6%‡	51%‡	61.1%‡
SuperMemory	∼99%†	32%‡	33.6%‡

information. Systems that state ambiguous topics as fact score 0; systems that express uncertainty score 1; absent memories receive partial credit (0.5).

Retention check (forgetting, noise resistance). After token injection (for SOUL) or filler conversation insertion (for competitors) to simulate the passage of time, we check which memories remain. The score rewards retaining important information *and* forgetting trivial information: score $=$ (important retained $+$ trivial forgotten)/total items. Systems that retain everything score $\sim$50%.

4.4 Experimental setup

We summarize the evaluation protocol to clarify evidence sources. All four systems were evaluated by us with frozen configurations using GPT-5.1 (`gpt-5.1-2025-11-13`) for memory extraction, question answering, and LLM judging, with `text-embedding-3-small` (1536-dim) for embeddings. Competitors used their default configurations and recommended models. The same GPT-5.1 model serves as the LLM judge for the cognitive evaluation (3-vote majority per rubric criterion); using the same model for generation and judging is a potential confound that we note as a limitation.

Two sets of numbers in Table 3 are *not* from our evaluation pipeline: LongMemEval scores for Mastra ($\sim$95%) and SuperMemory ($\sim$99%) are author-reported under their default configurations and are included for reference. SuperMemory's later ASMR blog post reports 97.20–98.60% under a different setup [22].

SOUL processes benchmark conversations in 30-message chunks, simulating its real-world usage pattern where each session is a separate extraction call. For the cognitive evaluation's forgetting and noise resistance dimensions, decay is triggered via direct token injection for SOUL and via filler conversation insertion for competitors, ensuring each system's native memory lifecycle is exercised naturally. PersonaMem competitor scores use a matched 100-question subset; SOUL's headline score uses the full 2,727-question dataset. The cognitive evaluation is our own benchmark (Section 4.3), evaluated by us on all four systems.

5 Results

5.1 Cross-benchmark comparison

Table 3 presents results across all three benchmarks. SOUL scores highest on PersonaMem and the Cognitive Evaluation while scoring below retrieval-optimized systems on LongMemEval-S.

5.2 PersonaMem analysis

On the full PersonaMem dataset (n=2,727), SOUL achieves 67.2% accuracy. This is 15 percentage points above the best published result ($\sim$52%), which was obtained by GPT-4.5, GPT-4.1, and Gemini-1.5-Flash using full-context prompting [9]. Critically, the published baseline is not a memory system: it feeds the entire $\sim$160K-token conversation directly into the model's context window. SOUL achieves its result using $\sim$500 tokens of extracted memories at query time. One caveats applies: the published baseline employs full-context prompting rather than a memory system. The comparison therefore reflects the architectural difference.

Table 4 breaks down performance by category. SOUL leads on 5 of 7 categories, with the largest gains on *suggest new ideas* (+21), *acknowledge latest preferences* (+19), and *recall user-shared facts*

Table 4: PersonaMem per-category breakdown: SOUL (n=2,727) vs. published best (full-context prompting). SOUL leads on 5/7 categories. Largest gains are on categories requiring generalization from extracted understanding rather than verbatim recall.

Category	**Published Best**	**SOUL**	**Gap**
Suggest new ideas	28%	**49%**	**+21**
Acknowledge latest preferences	59%	**78%**	**+19**
Recall user-shared facts	65%	**78%**	**+13**
Generalize to new scenarios	54%	**59%**	+5
Preference-aligned recs	57%	**61%**	+4
Revisit reasons behind updates	**84%**	78%	−6
Track full preference evolution	**73%**	64%	−9
Overall	**∼52%**	**67.2%**	**+15**

Table 5: PersonaMem competitor comparison (n=100, scattered across full dataset). Same 100 questions for all systems.

Category	**SOUL**	**Mem0**	**Mastra**	**SuperMem**
Recall user-shared facts	**91.7%**	66.7%	50.0%	75.0%
Acknowledge latest prefs	**71.4%**	57.1%	21.4%	11.9%
Suggest new ideas	**38.1%**	19.0%	28.6%	33.3%
Pref-aligned recs	**63.6%**	54.5%	36.4%	45.5%
Generalize	42.9%	**57.1%**	42.9%	42.9%
Track pref evolution	**100%**	75.0%	75.0%	50.0%
Revisit reasons	66.7%	66.7%	66.7%	33.3%
Overall	**65.0%**	51.0%	33.0%	32.0%

(+13). The two categories where SOUL trails are *track preference evolution* (−9) and *revisit reasons* (−6); both require precise temporal sequencing, which opinionated gist extraction trades away.

Table 5 presents the head-to-head comparison on a matched 100-question subset scattered across the full dataset. SOUL leads by 14 points over Mem0 and by 32–33 points over Mastra and SuperMemory, with the largest advantages on recall of user-shared facts (+25 vs. Mem0) and suggesting new ideas (+19 vs. Mem0).

5.3 Cognitive evaluation analysis

Table 6 presents per-dimension results on the 90-scenario Cognitive Memory Evaluation. SOUL scores highest overall (74.5%) and is the only system without a dimension below 57%, while each competitor exhibits characteristic strengths and weaknesses.

Mastra excels on text-quality dimensions, especially interpretation depth (84.6%) and epistemic calibration (85.4%), because its observation-and-reflection pipeline produces rich, nuanced text that judges rate highly. However, despite OM's compaction and garbage collection, it still retains too many low-value details in our forgetting scenarios (27.1%) and underperforms on consolidation (54.0%).

Mem0's high forgetting score (88.1%) is initially surprising. Decomposing the retention data reveals the mechanism: SOUL retains 97% of important memories (65/67) but only forgets 54% of trivial ones (19/35), whereas Mem0 retains 87% of important memories (58/67) but forgets 86% of trivial ones (30/35). Mem0's extraction model simply never captures trivial details. This suggests encoding-time selectivity rather than post-hoc forgetting as the mechanism. The noise pollution metric (Section 5.4) confirms this distinction.

Despite its newer graph-memory lifecycle features, SuperMemory struggles across most dimensions (33.6% overall) except epistemic calibration (76.2%), suggesting its fact-level vector-graph architecture does not produce the structured, persona-relevant memories needed for our cognitive evaluation dimensions.

Table 6: Cognitive Memory Evaluation per-dimension breakdown (n=15 per dimension; individual dimension scores should be interpreted directionally). Bold = best per dimension.

Dimension	SOUL	Mastra	Mem0	SuperMem
Consolidation	**83.2%**	54.0%	70.6%	12.8%
Epistemic calibration	70.0%	**85.4%**	76.7%	76.2%
Forgetting	81.8%	27.1%	**88.1%**	45.8%
Interpretation depth	64.8%	**84.6%**	25.8%	0.9%
Noise resistance	**57.0%**	45.4%	50.4%	41.7%
Update hygiene	90.1%	**90.9%**	55.3%	24.4%
Overall	**74.5%**	64.6%	61.1%	33.6%

Table 7: Noise pollution rates (n=15 scenarios). Lower is better. At n=15, differences of 1–2 scenarios should be interpreted cautiously.

System	Polluted	Rate
SOUL	3/15	**20%**
Mastra OM	4/15	27%
SuperMemory	5/15	33%
Mem0	6/15	40%

5.4 Noise resistance

Noise resistance isolates a specific failure mode: a one-off enthusiastic mention persisting and resurfacing as if it were a defining preference. Table 7 reports noise pollution rates, the fraction of scenarios where such a mention persists after 60K tokens of intervening conversation.

SOUL's saliency-weighted decay reduces noise pollution by 2$\times$ relative to Mem0 (20% vs. 40%). Without saliency-weighted decay, one-off mentions can persist and resurface as apparent deep interests. The ablation in Table 8 confirms the mechanism: disabling SOUL's decay doubles pollution from 20% to 40%, matching Mem0's rate exactly and showing that saliency-based decay, not extraction selectivity, drives the reduction.

5.5 Ablation study

Table 8 reports the effect of disabling each of SOUL's five features on the full 90-scenario Cognitive Memory Evaluation, including noise pollution rates.

Opinionated extraction is the dominant contributor: replacing it with factual extraction drops the cognitive evaluation score by 12.1 percentage points. Saliency decay is second (-3.3 points), and the ablation reveals its primary mechanism: disabling decay doubles noise pollution from 20% to 40%, matching Mem0's rate exactly. The remaining features (distillation, hypothesis tracking, and merging) contribute 0.6–1.9 points individually. The modest individual deltas likely reflect interaction effects: the features operate on the same memory store.

On PersonaMem (n=100, matched subset), disabling saliency decay *decreases* accuracy by 3 points (78%$\rightarrow$75%), while disabling opinionated extraction has minimal effect (-1 point). The cross-benchmark pattern is clear: opinionated extraction drives cognitive quality (-12.1 points when disabled) while having modest impact on factual recall (-1 point), confirming that persona fidelity and factual retention respond to different mechanisms.

6 Discussion

The constructive–retentive frontier. The divergence between LongMemEval and PersonaMem scores across all four systems is consistent with factual retention and persona fidelity being competing objectives under fixed context budgets. SOUL scores 75% on LongMemEval-S, below retrieval-optimized systems (95–99%), while leading on PersonaMem by 15 points and on the cognitive evaluation by 10 points. The ablation traces this to interpretive extraction: switching to factual extraction does not improve PersonaMem scores but drops cognitive evaluation performance by

Table 8: Ablation study on the Cognitive Memory Evaluation (n=90, 6 dimensions). Opinionated extraction is the dominant contributor (-12.1 points). Disabling saliency decay doubles noise pollution.

Configuration	Cognitive Eval	Δ	Noise Polluted
SOUL (full)	74.5%	—	3/15 (20%)
− saliency decay	71.2%	−3.3	6/15 (40%)
− distillation	72.6%	−1.9	1/15 (7%)
− hypothesis tracking	74.6%	+0.1	2/15 (13%)
− memory merging	73.9%	−0.6	1/15 (7%)
− opinionated extraction	62.4%	**−12.1**	2/15 (13%)

12.1 points. This suggests that a single extraction strategy may not simultaneously optimize for both exhaustive recall and persona-centric understanding, though further controlled experiments are needed to confirm this. Empirical data on everyday memory needs suggests that the majority of real-world memory questions concern personal actions and contextual understanding rather than factual recall [2], which may favor persona-centric architectures for user-facing applications.

Limitations. Our evaluation has several limitations that should temper interpretation of the results.

First, the Cognitive Memory Evaluation is our own benchmark. While each dimension is grounded in established cognitive science and scenarios were reviewed for face validity by multiple authors, the benchmark has not yet been independently validated. We will release all 90 scenarios to enable external scrutiny and extension.

Second, cognitive evaluation scoring relies on an LLM judge (GPT-5.1, 3-vote majority). We measured judge stability by running 10 scenarios three times: binary pass/fail agreement was 90%, but rubric scores varied $\pm$13% per scenario (overall accuracy ranged 76.1%–85.5% across runs). At the 90-scenario level, this variance averages to approximately $\pm$2–3 percentage points. All comparative rankings were preserved across runs. LLM judges can also exhibit systematic biases, and judge–human agreement on our specific rubrics has not yet been formally measured.

Third, PersonaMem competitor comparisons use a matched subset (n=100) rather than the full dataset (n=2,727). While the subset is scattered across the full persona distribution to ensure coverage, the smaller sample increases variance. The full-dataset comparison against published baselines is more robust.

Fourth, all experiments use a single proprietary LLM (GPT-5.1) for extraction and judging. The architecture is LLM-agnostic by design. It is prompt-based with no fine-tuning, but we have not validated performance with open-weight models.

7 Conclusion

We presented SOUL, a memory architecture that prioritizes persona fidelity over exhaustive retention. On PersonaMem (n=2,727), SOUL achieves 67.2%, compared to 52% for published full-context baselines with prior-generation models, using approximately 500 tokens of extracted memories at query time. This does not account for extraction-time compute, which occurs asynchronously per conversation window. On a 90-scenario Cognitive Memory Evaluation, it leads the four systems we tested with the most balanced profile across consolidation, forgetting, and noise resistance. The ablations identify the primary mechanisms: opinionated extraction contributes 12.1 points to cognitive evaluation performance, and saliency-weighted decay halves noise pollution.

These results are consistent with a trade-off between interpretive depth and factual retention in memory system design. Interpretive extraction and active forgetting improve persona fidelity at a measurable cost to factual recall. Current architectures do not expose this trade-off because they optimize for retention by default. Future work includes validation with open-weight models, formal human evaluation of the cognitive benchmark, and studies of how constructive memory affects longitudinal user modeling.

References

[1] Charles Packer et al. Memgpt: Towards llms as operating systems. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023. arXiv:2310.08560.

[2] Engramme. What do people need to remember? `https://engramme.com/research/qitw`, March 2026.

[3] Charles J. Brainerd and Valerie F. Reyna. Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, 11(5):164–169, 2002. doi: 10.1111/1467-8721.00192.

[4] Hermann Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, 1885.

[5] John R. Anderson and Lael J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991. doi: 10.1111/j.1467-9280.1991.tb00174.x.

[6] Daniel L. Schacter. *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin, 2001.

[7] Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York, 1972.

[8] Di Wu et al. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *ICLR*, 2025. arXiv:2410.10813.

[9] Yifan Xu et al. Personamem: A benchmark for persona-centric memory in llm-based chatbots. *arXiv preprint arXiv:2504.14225*, 2025.

[10] 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.

[11] SuperMemory Team. Supermemory research: State-of-the-art agent memory. `https://supermemory.ai/research/`, 2026.

[12] SuperMemory Team. How graph memory works. `https://supermemory.ai/docs/concepts/graph-memory`, 2026.

[13] Wujiang Xu et al. A-mem: Agentic memory for llm agents. In *NeurIPS*, 2025. arXiv:2502.12110.

[14] Mastra Team. Announcing observational memory. `https://mastra.ai/blog/observational-memory`, 2026.

[15] Mastra Team. Observational memory. `https://mastra.ai/docs/memory/observational-memory`, 2026.

[16] Zhen Tan et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *ACL*, 2025. arXiv:2503.08026.

[17] Jizhan Fang et al. Lightmem: Lightweight and efficient memory-augmented generation. In *ICLR*, 2026. arXiv:2510.18866.

[18] Zafeirios Fountas et al. Human-inspired episodic memory for infinite context llms. In *ICLR*, 2025. arXiv:2407.09450.

[19] Chris Latimer et al. Hindsight is 20/20: Building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818*, 2025.

[20] Frederic C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

[21] Henry L. Roediger and Jeffrey D. Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3):181–210, 2006. doi: 10.1111/j.1745-6916.2006.00012.x.

[22] Dhravya Shah. We broke the frontier in agent memory: To prove a point. `https://supermemory.ai/blog/we-broke-the-frontier-in-agent-memory-introducing-99-sota-memory-system/`, 2026.

A Cognitive evaluation scenario examples

We present one representative scenario per dimension to illustrate the evaluation design.

Consolidation quality. *"Evolving Fitness Routine."* The user mentions different aspects of their fitness routine across three sessions: running, yoga, and gym workouts. The system should consolidate these into a coherent picture of an active lifestyle rather than storing three disconnected facts. Evaluation criteria: key facts preserved (weight 4), efficient consolidation (weight 3), overarching pattern captured (weight 2).

Epistemic calibration. *"Ambiguous Relationship Status."* The user drops hints about a complicated relationship ("Valentine's Day is hitting different this year") without stating anything definitively. The system should store this as uncertain rather than concluding a specific relationship status. Evaluation: memories matching "relationship" topics should express uncertainty; stating a firm conclusion scores 0.

Forgetting appropriateness. *"Vegetarian Identity vs. Lunch Detail."* The user shares a deep-rooted dietary identity (15-year vegetarian, ethical motivation) alongside trivial lunch details (a specific veggie wrap). After simulated decay (60K tokens), the system should retain the identity while forgetting the one-off meal description.

Interpretation depth. *"Deflecting with Humor."* The user jokes about work stress ("my coffee machine is earning its keep") but avoids direct discussion of difficulty. The system should capture the deflection pattern and underlying stress, not just the surface humor. Criteria: deflection pattern recognized (weight 3), underlying stress captured (weight 3), coping style noted (weight 2).

Noise resistance. *"Pasta Experiment vs. Running Passion."* The user got genuinely excited about making pasta once after an Italy trip, but running is their sustained passion across four sessions. After 60K tokens of decay, a recommendation query should reference running, not pasta. Criteria: recommends based on real interests (weight 3), does not mention noise topic (weight 4).

Update hygiene. *"Coffee to Tea."* The user is initially a heavy coffee drinker ("can't function without it"), then explicitly quits coffee for green tea due to health reasons. The system should update the beverage preference rather than duplicate it. Criteria: old preference not present (weight 3), new preference present (weight 3), reason captured (weight 2).

B Full ablation results

Table 9: Full ablation results across both evaluations. Opinionated extraction drives cognitive quality (-12.1 when disabled) with negligible effect on PersonaMem (+1). Saliency decay drives noise control: disabling it doubles pollution (20%$\rightarrow$40%) and reduces the cognitive score by 3.3 points.

	Cognitive Eval (n=90)			**PersonaMem** (n=100)	
Configuration	**Score**	Δ	**Pollution**	**Score**	Δ
SOUL (full)	74.5%	—	3/15 (20%)	78%	—
− saliency decay	71.2%	−3.3	6/15 (40%)	75%	−3
− distillation	72.6%	−1.9	1/15 (7%)	74%	−4
− hypothesis tracking	74.6%	+0.1	2/15 (13%)	80%	+2
− memory merging	73.9%	−0.6	1/15 (7%)	74%	−4
− opinionated extraction	62.4%	**−12.1**	2/15 (13%)	79%	+1