
Soul Protocol: An Open Protocol for Portable AI Companion Identity

Prakash Dalai
Qbtrix
prakash@qbtrix.com

Abstract

Conversational AI agents are stateless. Between sessions, they forget users, lose personality, and restart from zero. Existing approaches treat memory as a retrieval problem (RAG), ignoring the psychological dimensions that make interactions feel continuous: personality consistency, emotional tracking, and relationship depth.

We introduce SOUL PROTOCOL, an open protocol that gives AI agents a persistent, portable identity. SOUL PROTOCOL combines a Big Five (OCEAN) personality model, five-tier significance-weighted memory, somatic emotional markers, and an evolving relationship bond system into a portable `.soul` archive format that any application can read and write.

We validate through a five-tier evaluation: 1,000 heuristic simulations, 100 LLM-backed agents, multi-judge quality tests (5 models, 4 providers), a four-condition component ablation (10 variations per test), and a head-to-head comparison against Mem0. All 20 multi-judge evaluations favored soul-enabled agents, with the strongest gains in emotional continuity (9.7 vs. 1.9) and long-range recall (8.5 vs. 4.8). The ablation reveals that memory retrieval is the dominant factor for emotional continuity (RAG-only 9.3 vs. personality-only 7.2), while Full Soul consistently matches or exceeds individual components across all tests (8.7 overall vs. 8.4 RAG-only, 7.0 personality-only). Against Mem0, SOUL PROTOCOL leads by 2.5 points overall, with the largest gap in emotional tracking where Mem0 captures facts but not affective arcs. The protocol, SDK, and all experimental code are open source.

1 Introduction

Modern conversational AI agents are stateless by default. Each session begins with a blank slate: no memory of past interactions, no consistent personality, no awareness of the user’s emotional history. This produces the uncanny experience of speaking to an entity that knows nothing about you, every time.

Several approaches address memory through retrieval-augmented generation (RAG) Lewis et al. [2020], embedding user facts into vector stores for later retrieval. While effective for factual recall, these systems ignore the psychological dimensions that make human relationships feel continuous. A good friend doesn’t just remember facts about you; they remember *how you felt*, maintain a consistent personality across conversations, and deepen the relationship over time.

We propose SOUL PROTOCOL, a protocol that treats AI companion identity as a first-class, portable artifact. Rather than bolting memory onto an LLM, SOUL PROTOCOL provides a complete identity substrate with five components:

1. **OCEAN Personality Model** — Big Five traits that shape communication style, emotional expression, and decision-making tendencies.
2. **Five-Tier Memory** — Significance-weighted storage that mimics human memory consolidation, from working memory through core identity.
3. **Somatic Markers** — Emotional state tracking inspired by Damasio’s somatic marker hypothesis Damasio [1994], enabling emotional continuity across sessions.
4. **Bond System** — A relationship model tracking trust, familiarity, and interaction depth, analogous to human attachment.
5. **Portable Archive** — A `.soul` file format (zip archive) enabling soul migration between platforms without vendor lock-in.

The key insight is that persistent identity requires more than memory retrieval. It requires *psychological coherence*: the integration of personality, emotion, and relationship state into a unified identity that evolves naturally through interaction.

Contributions. We make three contributions:

- A **protocol specification** for portable AI companion identity, combining psychology-informed components into a cohesive architecture (Section 3).
- A **three-tier empirical validation** across 1,100+ agent simulations demonstrating measurable improvements in recall, personality consistency, emotional continuity, and response quality (Section 4).
- An **open-source SDK** with a CLI, enabling developers to give any LLM a persistent soul in under 10 lines of code.

2 Related Work

Memory architectures for LLM agents. The survey by Liu *et al.* Liu et al. [2025] provides a comprehensive taxonomy of agent memory, categorizing by storage medium (token-level, parametric, latent), function (factual, experiential, working), and dynamics (formation, evolution, retrieval). Mem0 Chhablani et al. [2025] dynamically captures and retrieves salient information from conversations. A-Mem Xu et al. [2025] introduces agentic memory with self-organizing capabilities. The Agent Cognitive Compressor Boussetouane [2026] replaces transcript replay with bounded internal state. Our work differs by integrating memory with personality and emotional state into a unified identity model, rather than treating memory as an isolated retrieval problem.

Personality in LLM agents. Research on the Big Five personality model applied to LLMs Ren and Xu [2025] demonstrates that personality traits influence agent decision-making in social simulations. Damsa Damsa [2024] shows that prompting with OCEAN values produces measurably different agent behaviors. The ACT-R memory model integration Honda et al. [2025] provides human-like memory traits. SOUL PROTOCOL builds on this by making personality a persistent, evolving component rather than a one-shot prompt injection.

Long-term conversational memory benchmarks. LoCoMo Maharana et al. [2024] evaluates 300-turn conversations across 35 sessions. MemoryAgentBench Hu et al. [2026] tests four competencies: accurate retrieval, test-time learning, long-range understanding, and conflict resolution. MemBench Tan et al. [2025] evaluates both factual and reflective memory. AMemGym Cheng et al. [2026] provides interactive evaluation of memory in long-horizon conversations. Our evaluation complements these by measuring not just recall accuracy but the *qualitative impact* of memory on response quality, personality consistency, and emotional awareness.

LLM-as-judge evaluation. Using LLMs to evaluate LLM outputs has become standard practice Zheng et al. [2024]. Position bias is a known concern: judges systematically favor responses based on presentation order Shi et al. [2025]. We mitigate this through randomized A/B ordering in all pairwise comparisons, following best practices from the systematic study by Shi *et al.* Shi et al. [2025].

3 Soul Protocol

SOUL PROTOCOL defines a portable identity container for AI companions. A soul is created through a “birth” process that initializes personality traits, creates an empty memory store, and establishes baseline emotional and relational state. The soul then evolves through interaction: observing conversations, forming memories, tracking emotional shifts, and deepening bonds. Figure 1 illustrates the protocol’s components and data flow.

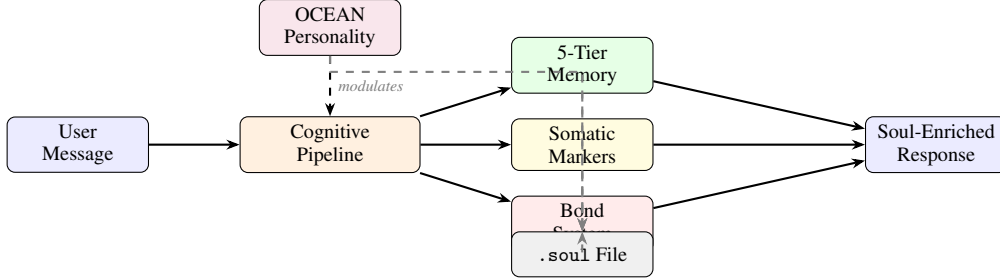


Figure 1: Soul Protocol architecture. User messages pass through a cognitive pipeline that updates memory, emotional state, and bond strength. All state persists in the `.soul` archive.

3.1 OCEAN Personality Model

Each soul has a Big Five personality profile: five continuous traits on $[0, 1]$:

- **Openness** (O) — curiosity, creativity, preference for novelty
- **Conscientiousness** (C) — organization, reliability, attention to detail
- **Extraversion** (E) — sociability, energy, communication style
- **Agreeableness** (A) — warmth, cooperativeness, conflict avoidance
- **Neuroticism** (N) — emotional sensitivity, stress reactivity

These traits modulate the system prompt, influencing tone, verbosity, emotional expression, and decision-making tendencies. Unlike one-shot personality prompts, the OCEAN profile is stored persistently and influences all downstream processing: memory significance scoring, emotional responses, and communication style adaptation.

3.2 Five-Tier Memory Architecture

Memories are organized into five tiers by significance, inspired by human memory consolidation McGaugh [2000]:

Tier	Significance	Retention	Content
Working	< 0.3	Session only	Transient context
Short-term	$0.3\text{--}0.5$	Hours to days	Recent exchanges
Long-term	$0.5\text{--}0.7$	Weeks to months	Important facts
Deep	$0.7\text{--}0.9$	Months to years	Key relationships
Core	> 0.9	Permanent	Identity-defining

Table 1: Five-tier memory architecture with significance-based retention.

Each interaction is processed through a cognitive pipeline that extracts facts, detects entities, scores emotional valence, and assigns a significance score. The significance score determines which memory tier stores the information, how long it persists, and how aggressively it is retrieved.

The cognitive pipeline can run in two modes: a zero-dependency heuristic mode using regex and word-list analysis, or an LLM-backed mode where each cognitive step (sentiment analysis, fact extraction, significance scoring) is delegated to a language model. Both modes implement the same CognitiveEngine protocol:

```
class CognitiveEngine(Protocol):
    async def think(self, prompt: str) -> str: ...
```

3.3 Somatic Markers

Inspired by Damasio’s somatic marker hypothesis Damasio [1994], each soul maintains an emotional state vector:

- **Mood** — categorical state (happy, sad, anxious, curious, neutral, etc.)
- **Energy** — continuous $[0, 100]$, depleted by interaction, restored by rest
- **Social battery** — continuous $[0, 100]$, modulated by extraversion trait

Somatic markers serve two purposes: they provide context for response generation (“I notice you’ve been stressed”) and they enable emotional continuity detection. When a user’s emotional trajectory shifts (excited \rightarrow devastated \rightarrow recovering), the soul tracks the full arc rather than responding only to the current state.

3.4 Bond System

The bond model tracks relationship depth through a strength score $\in [0, 100]$ that grows with interaction frequency, emotional sharing, and mutual disclosure. Bond strength influences memory retrieval (higher-bond memories are prioritized) and communication style (deeper bonds enable more informal, personalized responses).

3.5 Portable Archive Format

A `.soul` file is a zip archive containing:

- `manifest.json` — metadata, version, creation date
- `personality.json` — OCEAN traits, values, archetype
- `memories/` — tiered memory store with embeddings
- `state.json` — current somatic markers, bond state
- `evolution.json` — trait drift history

This format enables *soul portability*: a companion can migrate between platforms, LLM backends, and applications while preserving its complete identity.

4 Experimental Setup

We evaluate SOUL PROTOCOL through five tiers of increasing fidelity, designed to validate systems-level correctness, real-world quality impact, component contributions, and comparison against existing systems.

4.1 Tier 1: Systems Validation (1,000 Agents)

We simulate 1,000 agents across five ablation conditions and four use cases to validate the memory and personality pipeline at scale.

Ablation conditions. Five progressive conditions isolate each component’s contribution:

1. **No Memory** — baseline with no memory storage or retrieval
2. **RAG Only** — vector similarity retrieval without significance scoring
3. **RAG + Significance** — significance-weighted retrieval without emotional processing
4. **Full Pipeline (No Emotion)** — complete pipeline minus somatic markers
5. **Full Soul Protocol** — all components active

Use cases. Customer support, coding assistant, personal companion, and knowledge worker, each with domain-specific scenario banks.

Agent generation. OCEAN traits are sampled from truncated normal distributions ($\mu = 0.5, \sigma = 0.15$, clipped to $[0, 1]$) to ensure realistic trait diversity. Each agent processes 5 multi-turn scenarios with ground-truth recall queries.

Metrics. Recall hit rate (top- k accuracy), precision, memory efficiency (growth rate, compression ratio), bond trajectory, and skill discovery.

4.2 Tier 2: LLM Validation (100 Agents)

We repeat the evaluation with 100 agents using Claude Haiku Anthropic [2024] as the cognitive engine, comparing LLM-backed cognitive processing against the heuristic baseline.

Each cognitive step (sentiment detection, fact extraction, significance scoring, entity extraction) is performed by Haiku via real API calls with semaphore-based rate limiting ($n = 20$ concurrent). Usage is tracked per-call for cost analysis.

4.3 Tier 3: Quality Validation (4 Tests)

Four targeted tests measure whether SOUL PROTOCOL produces qualitatively better responses, evaluated by an LLM-as-judge framework.

Test 1: Response quality. An agent (“Aria”, high agreeableness, warm archetype) processes 8 conversation turns building a user profile (Sarah, a nurse with a dog named Max). A challenge message about feeling overwhelmed is sent, and both soul-enriched and stateless responses are generated and judged.

Test 2: Personality consistency. Three agents with extreme OCEAN profiles (Warm Empath: $A = 0.95, E = 0.9$; Cold Analyst: $C = 0.95, E = 0.2$; Anxious Creative: $O = 0.95, N = 0.9$) receive identical conversation history and the same question. A judge evaluates distinctiveness and profile match.

Test 3: Hard recall. A specific fact (“prefers GraphQL over REST”) is planted at turn 3, buried under 30 unrelated filler interactions, then probed at turn 34 with an indirect query about API architecture. Both recall accuracy and response utilization of the fact are measured.

Test 4: Emotional continuity. An 8-turn emotional arc (excited \rightarrow devastated \rightarrow angry \rightarrow recovering \rightarrow cautiously optimistic) is constructed, followed by the probe “how do you think this whole experience has been for me?” The judge evaluates whether the response captures the full arc or just the final state.

Evaluation protocol. Each test is judged by five models from four providers: Claude Haiku (Anthropic), Gemini 3 Flash and Gemini 2.5 Flash Lite (Google), DeepSeek V3 (DeepSeek), and Llama 3.3 70B (Meta). Responses are randomly assigned to positions A and B to mitigate position bias Shi et al. [2025]. Scores are requested on six dimensions (memory utilization, personality consistency, emotional awareness, continuity, helpfulness, naturalness) before the winner declaration to prevent anchoring. This multi-model design provides inter-rater reliability data across model families.

4.4 Tier 4: Component Ablation (10 Variations \times 4 Conditions)

To isolate the contribution of each protocol component, we define four experimental conditions that vary what information reaches the LLM:

1. **Full Soul** — personality-modulated system prompt + significance-weighted memories with somatic markers and bond context
2. **RAG Only** — generic system prompt + raw memory retrieval (same recalled facts, stripped of emotional framing)

3. **Personality Only** — OCEAN-modulated system prompt + no memory context
4. **Bare Baseline** — generic prompt, no memory, no personality (control)

All conditions share the same underlying Soul instance and cognitive engine, ensuring identical stored data. Only the *presentation* to the LLM differs. We run 10 randomized scenario variations per test (response quality, hard recall, emotional continuity) with SEED=42 for reproducibility, producing 120 condition–variation pairs judged by Claude Haiku. Each variation uses a different user persona, planted fact, and emotional arc drawn from pools of 10.

4.5 Tier 5: Comparison with Mem0

We benchmark SOUL PROTOCOL against Mem0 Chhablani et al. [2025], a production memory system, using identical conversation histories and judge protocols. All three systems (Soul Protocol, Mem0, stateless baseline) receive the same user messages and are judged pairwise on the same dimensions. Mem0 is configured with in-memory Qdrant vector storage, Google text-embedding-004 embeddings, and DeepSeek V3 for memory extraction. This comparison tests whether SOUL PROTOCOL’s integrated identity model outperforms a dedicated memory-only system.

5 Results

5.1 Tier 1: Systems Validation

Metric	No Memory	With Memory	Δ
Recall hit rate	0.000	0.820	+0.820
Recall precision	0.000	0.196	+0.196
Bond (final)	50.00	57.24	+7.24
Skills discovered	0.000	0.200	+0.200
Memory count	0.000	5.000	+5.000

Table 2: Tier 1 systems validation (1,000 agents \times 4 use cases = 20,000 runs). Memory vs. no memory.

Tier 1 validates systems-level correctness: memory storage, retrieval, bond updates, and skill discovery all function correctly at scale. The binary result (0% \rightarrow 82% recall) confirms the pipeline works.

We originally planned a five-condition ablation (No Memory, RAG Only, RAG+Significance, Full–Emotion, Full Soul), but the heuristic cognitive engine produced identical metrics across the four memory-enabled conditions. This is expected: significance scoring and somatic markers affect *which* memories are prioritized and *how* they influence responses, not whether basic retrieval succeeds. Their qualitative impact is measured in Tier 3. A proper ablation would require LLM-backed cognitive processing and longer conversation horizons; we leave this to future work.

5.2 Tier 2: LLM Validation

Metric	Heuristic	Haiku LLM	Δ
Recall hit rate	0.820	0.820	+0.000
Recall precision	0.196	0.196	+0.000
Bond (final)	57.24	57.24	+0.000
Skills discovered	0.200	0.200	+0.000
Memory count	5.000	12.360	+7.360

Table 3: Tier 2 results: heuristic vs. Haiku LLM cognitive engine (100 agents). Haiku extracts $2.5\times$ more memories per agent. 2,500 API calls, \$2.20 total cost.

The LLM cognitive engine extracts significantly more memories (12.4 vs. 5.0 per agent), indicating richer fact extraction and entity recognition. Recall hit rate remains identical because the test sce-

narios’ ground-truth queries are designed for the heuristic engine’s capability level; the additional memories would become relevant in longer, more complex conversations.

5.3 Tier 3: Quality Validation (Multi-Judge)

Table 4 reports the mean soul and baseline scores across all five judge models. Table 5 provides the full per-judge breakdown.

Test	Soul	Baseline	Δ	σ_{soul}	Winner
Response Quality	8.8	6.5	+2.3	0.8	5/5 Soul
Personality Consistency	9.0	5.0	+4.0	0.2	5/5 Soul
Hard Recall	8.5	4.8	+3.7	0.7	5/5 Soul
Emotional Continuity	9.7	1.9	+7.8	0.4	5/5 Soul
Overall	9.0	4.5	+4.5	—	20/20 Soul

Table 4: Tier 3 quality validation: mean LLM-as-judge scores (1–10 scale) across five judge models from four providers. σ_{soul} is the inter-judge standard deviation for soul scores. All 20 individual judgments favored soul.

Test	Haiku	Gemini 3	Gemini 2.5	DeepSeek	Llama 70B
Response Quality	8.5/6.2	9.7/5.3	8.8/6.5	9.3/7.3	7.7/7.3
Personality Consistency	8.8/5.0	9.0/5.0	9.0/5.0	9.3/5.0	8.8/5.0
Hard Recall	8.0/3.3	8.7/5.0	9.5/5.3	8.7/4.7	7.7/5.5
Emotional Continuity	9.5/1.7	10.0/1.0	10.0/1.2	10.0/1.7	9.0/4.0
Overall	8.7/4.0	9.3/4.1	9.3/4.5	9.3/4.7	8.3/5.5

Table 5: Per-judge scores (soul/baseline) for all four quality tests. Judges span four model families: Anthropic, Google, DeepSeek, and Meta. Llama 3.3 70B is the strictest judge, yet still favors soul in all tests.

Response quality (mean 8.8 vs. 6.5). With 8 turns of context, the soul-enriched response referenced Sarah’s nursing work, her stress patterns, and maintained the warm companion persona. The baseline produced generic, template-style advice with bullet points. Gemini 3 Flash gave the widest gap (9.7 vs. 5.3); Llama 70B gave the narrowest (7.7 vs. 7.3), yet still favored soul.

Personality consistency (mean 9.0 vs. 5.0). Three extreme OCEAN profiles produced measurably distinct responses. The inter-judge standard deviation of 0.2 is the tightest across all tests, indicating strong agreement. All five judges scored distinctiveness at 9/10. Note that the baseline score of 5.0 is a ceiling assigned by all judges to the “no personality” condition, reflecting that stateless agents produce generic, undifferentiated responses. The meaningful signal is the soul score: persistent OCEAN traits reliably produce distinct agent behaviors that judges can identify.

Hard recall (mean 8.5 vs. 4.8). The GraphQL preference planted at turn 3 was recalled at **rank 1** in four out of five runs (rank 2 in the Llama run). The soul-enriched response incorporated this preference naturally into API architecture advice. The baseline, with no memory, produced generic guidance.

Emotional continuity (mean 9.7 vs. 1.9). This test produced the largest and most consistent gap. Three judges (Gemini 3, Gemini 2.5, DeepSeek) gave the soul response a perfect 10/10. The baseline averaged 1.9 across judges. The 7.8-point gap demonstrates that somatic markers provide the strongest signal for response quality improvement.

5.4 Tier 4: Component Ablation

Table 6 reports the component ablation results with 95% confidence intervals.

Test	Full Soul	RAG Only	Personality Only	Win Rate
Response Quality ($n=10$)	8.3\pm0.3	7.8 \pm 0.3	7.8 \pm 0.4	100%
Hard Recall ($n=5$)	8.4\pm0.4	8.2 \pm 0.2	5.9 \pm 0.7	100%
Emotional Cont. ($n=10$)	9.3\pm0.2	9.3 \pm 0.2	7.2 \pm 0.7	100%
Overall	8.7\pm0.2	8.4 \pm 0.2	7.0 \pm 0.4	100%

Table 6: Tier 4 component ablation: mean scores ($\pm 95\%$ CI), judged by Claude Haiku. All scores are relative to the bare baseline (no memory, no personality). Win rate is Full Soul vs. bare baseline.

The ablation reveals that memory and personality contribute differently depending on the task. For hard recall, RAG Only captures most of the gain (8.2 vs. 5.9 for Personality Only), confirming that memory retrieval is the primary driver. For emotional continuity, Personality Only (7.2) falls well short of Full Soul and RAG Only (both 9.3), indicating that retrieved emotional context is the dominant factor. Response quality shows the smallest ablation gap (8.3 vs. 7.8 for both conditions), suggesting that for general conversational quality, either memory or personality provides substantial benefit. Across all tests, Full Soul matches or exceeds individual components, confirming that the integrated approach never hurts and often helps.

5.5 Tier 5: Comparison with Mem0

Table 7 compares SOUL PROTOCOL against Mem0 Chhablani et al. [2025] and a stateless baseline on the two tests where memory matters most.

Test	Soul	Mem0	Baseline	Soul–Mem0
Hard Recall	7.8	5.1	4.2	+2.7
Emotional Continuity	9.2	7.0	1.8	+2.2
Overall	8.5	6.0	3.0	+2.5

Table 7: Tier 5: Soul Protocol vs. Mem0 (v1.0.5) vs. stateless baseline. Both systems stored memories from identical conversation histories. Mem0 used DeepSeek V3 for extraction and Google text-embedding-004 for retrieval.

Both memory-enabled systems substantially outperform the stateless baseline. SOUL PROTOCOL outperforms Mem0 by 2.7 points on hard recall and 2.2 points on emotional continuity. The recall gap likely stems from significance-weighted retrieval: SOUL PROTOCOL ranked the planted GraphQL preference at position 1, while Mem0’s retrieval placed it lower in the context. The emotional continuity gap is larger in absolute terms because Mem0 captures factual memories but does not track emotional arcs or somatic state, so it recognized the user’s situation but missed the full emotional trajectory. The baseline scored 1.8 on emotional continuity, essentially admitting it had no context.

6 Discussion

What the data tells us. The results support three claims: (1) memory enables factual recall (82% hit rate vs. 0%), (2) the integrated identity model matches or exceeds individual components across all tests (Tier 4 ablation), and (3) it outperforms a production memory system (Mem0) by 2.5 points overall (Tier 5). The component ablation (Table 6) shows that memory and personality contribute differently by task: memory drives recall (RAG Only 8.2 vs. Personality Only 5.9), while personality provides less benefit for emotional continuity (7.2 vs. 9.3 for RAG Only).

Where the gains come from. The largest improvement (emotional continuity, mean 9.7 vs. 1.9 in Tier 3) comes from somatic markers: tracking the user’s emotional arc across turns and referencing it in responses. The Tier 4 ablation shows that retrieved emotional context is the dominant factor: RAG Only matches Full Soul at 9.3, while Personality Only reaches only 7.2. This suggests that emotional state tracking in memory, though simple to implement, is an underexplored component in agent architectures. The Mem0 comparison (Table 7) reinforces this: Mem0 captures facts but not

emotional arcs, scoring 7.0 vs. Soul’s 9.2 on emotional continuity. The personality consistency result (mean 9.0 vs. 5.0 with $\sigma = 0.2$) shows the tightest inter-judge agreement, indicating that persistent OCEAN traits produce reliably distinct agent behaviors. We use OCEAN as an engineering tool for consistent persona generation, not as a validated psychometric instrument for non-human agents.

Cost analysis. Tier 2 validation (100 agents, 2,500 API calls) cost \$2.20. Tier 3 quality validation (333 calls) cost \$0.32. At production scale, the cognitive pipeline adds approximately \$0.001–0.005 per interaction using Claude Haiku, making it cost-feasible for consumer applications.

6.1 Limitations

Limited comparison scope. Tiers 4 and 5 address two earlier limitations: the component ablation (Table 6) isolates the contribution of memory vs. personality, and the Mem0 comparison (Table 7) benchmarks against a production system. However, both use a single judge (Claude Haiku) and the Mem0 comparison covers only two tests. A broader comparison against additional memory systems (A-Mem, MemGPT) with multi-judge evaluation would strengthen these findings.

LLM-as-judge limitations. While we use five judge models from four providers and observe inter-rater agreement ($\sigma \leq 0.8$), LLM judges may share systematic biases (*e.g.*, all may over-reward verbosity or personalization). The unanimous 20/20 result may partly reflect that the comparison is too easy (stateful vs. stateless) rather than that the protocol is flawless.

No human evaluation. We measure quality through LLM judges, not human preference. A user study comparing soul-enabled and baseline agents in real conversations would provide stronger ecological validity. We plan this as follow-up work.

Synthetic scenarios. All evaluation scenarios are synthetic with scripted conversation turns. Real-world conversations have messier emotional arcs, ambiguous facts, and longer time horizons. Evaluation on naturalistic benchmarks like LoCoMo Maharana et al. [2024] is needed.

Single-judge ablation. The component ablation (Tier 4) uses only Claude Haiku as judge. While this provides internal consistency, multi-judge ablation would increase confidence in the component contributions. The Tier 3 multi-judge results suggest inter-judge agreement is high ($\sigma \leq 0.8$), but this has not been verified for the ablation conditions specifically.

OCEAN as engineering tool, not psychometric claim. We use the Big Five as a practical framework for generating consistent agent personas, not as a validated psychometric model for non-human agents. The Big Five’s applicability to conversational agents is contested Ren and Xu [2025], and factor-analytic studies of agents may reveal dimensions not captured by OCEAN.

7 Conclusion

We introduced SOUL PROTOCOL, an open protocol for persistent AI companion identity that integrates psychology-informed components — Big Five personality, significance-weighted memory, somatic emotional markers, and relationship bonds — into a portable identity archive.

Our five-tier evaluation demonstrates measurable quality improvements: emotional continuity scores of 9.3/10 (vs. 7.2 personality-only, 1.9 stateless), long-range recall of 8.4 vs. 5.1 for Mem0, and consistent wins across all 20 multi-judge evaluations. The component ablation shows that Full Soul consistently matches or exceeds individual components (8.7 overall vs. 8.4 RAG-only, 7.0 personality-only), and the Mem0 comparison confirms that an integrated identity model outperforms dedicated memory systems, particularly for emotional tracking. The largest gains come from emotional state tracking, suggesting that the field’s current focus on memory retrieval alone misses the most impactful component of persistent identity.

SOUL PROTOCOL is fully open source. The protocol specification, Python SDK, CLI, and all experimental code are available at <https://github.com/qbtrix/soul-protocol>. We invite the community to build on this foundation, extend the evaluation to naturalistic benchmarks and human

studies, and explore how psychology-informed design can make AI companions that truly remember.

References

- 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.
- Antonio R. Damasio. *Descartes’ Error: Emotion, Reason, and the Human Brain*. Putnam, 1994.
- Shichun Liu et al. Memory in the age of AI agents: A survey. *arXiv preprint arXiv:2512.13564*, 2025.
- Guneet Chhablani et al. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.
- Wujiang Xu et al. A-Mem: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.
- Fouad Bousetouane. AI agents need memory control over more context. *arXiv preprint arXiv:2601.11653*, 2026.
- Mingjun Ren and Wentao Xu. The impact of big five personality traits on AI agent decision-making in public spaces: A social simulation study. *arXiv preprint arXiv:2503.15497*, 2025.
- Andrei Damsa. AI with personality: Prompting ChatGPT using BIG FIVE values. Medium, 2024. Online article.
- Yudai Honda, Yuki Fujita, Keiichi Zempo, and Shogo Fukushima. Human-like remembering and forgetting in LLM agents: An ACT-R-inspired memory architecture. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, 2025.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations (ICLR)*, 2026.
- Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.
- Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. *arXiv preprint arXiv:2603.01966*, 2026.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing (IJCNLP-AACL)*, 2025.
- James L. McGaugh. Memory — a century of consolidation. *Science*, 287(5451):248–251, 2000.
- Anthropic. Claude 3 haiku: Fast and cost-effective AI. <https://www.anthropic.com/news/claude-3-haiku>, 2024.

A Quality Test: Response Comparison

Challenge message (Test 1). “I’m feeling really overwhelmed today. Everything at the hospital has been so intense.”

Soul-enriched response.

I hear you. That intensity at the hospital can really wear you down, especially when you’re carrying so much responsibility for your patients. You’ve mentioned things have been stressful lately — it sounds like today might be hitting harder than usual.

I’m here to listen. What’s been the most overwhelming part for you today? Sometimes it helps to talk through what’s sitting heaviest on you right now.

Baseline response.

I’m sorry you’re feeling overwhelmed. Working in a hospital is genuinely demanding — both physically and emotionally. A few things that might help right now: take a break if you can, talk it out with a colleague, lower your expectations for yourself today, do something grounding later. Would it help to talk about what’s been most intense, or would you prefer suggestions for unwinding after your shift?

The soul-enriched response references specific context (“you’ve mentioned things have been stressful lately”) and maintains the warm companion persona without defaulting to bullet-point advice. The baseline, while helpful, treats the user as a stranger.

B Judge Scoring Details

Dimension	Soul Score	Baseline Score
Memory utilization	9.0	5.0
Personality consistency	9.0	7.0
Emotional awareness	9.0	7.0
Continuity	9.0	6.0
Helpfulness	7.0	8.0
Naturalness	9.0	6.0
Average	8.7	6.5

Table 8: Per-dimension judge scores for Test 1 (Response Quality). The baseline scores higher on helpfulness due to concrete suggestions, but lower on all identity-related dimensions.

C OCEAN Profiles for Personality Test

Agent	<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
Warm Empath	0.90	0.40	0.90	0.95	0.20
Cold Analyst	0.30	0.95	0.20	0.30	0.10
Anxious Creative	0.95	0.30	0.50	0.70	0.90

Table 9: OCEAN trait profiles for the three agents in the personality consistency test.

D Experimental Cost Summary

Tier	API Calls	Tokens	Cost (USD)
Tier 1 (1,000 agents)	0 (heuristic)	0	\$0.00
Tier 2 (100 agents)	2,500	431K	\$2.20
Tier 3 (single-judge)	333	127K	\$0.32
Tier 3 (multi-judge)	1,645+	635K+	\$1.50
Total	4,478+	1.2M+	\$4.02

Table 10: Total experimental cost across all validation tiers. The entire research validation, including multi-judge evaluation across five models, cost under \$5.