MEMORIA: A Self-Evolving Agentic Framework with Transparent, User-Controlled Memory

Qiran Hu [† 1] **Amy Bisalputra** [* 1] **Ke Ding** [* 1] **Min Kim** [* 1] **Kewen Xia** [* 1]

Abstract

We present Memory Enhanced Multi-modal Orchestration Reasoning Intelligence Architecture (MEMORIA), a web application that enables self-evolving AI assistants with transparent, user-controlled memory. Our system addresses fundamental inefficiencies in current AI interactions, where users constantly have to re-establish context across sessions, resulting in significant token waste and degraded performance. MEMORIA learns user work patterns through procedural learning rather than extensive conversation history, achieving personalization comparable to fine-tuning at approximately 1% of the token cost. Users can visualize, rate, and modify the AI's learned memory, creating unprecedented transparency in personalized AI systems.

Team Name: MIRA (Memory Incremental Reasoning Architecture) **Team Number:** 4

1. Introduction

Large language models have demonstrated remarkable capabilities across diverse tasks, yet they face a fundamental architectural limitation: the absence of persistent, adaptive memory that accumulates knowledge from user interactions over time. Each new conversation session requires users to re-explain their preferences, context, and working style, creating substantial friction in human-AI collaboration. Research demonstrates that LLMs suffer a 39% performance drop in multi-turn conversations as they fail to maintain coherent context over extended interactions [1].

2. Motivation

The development of persistent, adaptive memory systems for AI assistants represents a critical frontier in human-AI interaction research, driven by fundamental architectural limitations and growing demand for personalized, efficient AI collaboration.

2.1. Background and Related Work

Recent advances in LLM memory systems have explored diverse architectural approaches. Behrouz et al. [2] introduced ATLAS, a long-term memory module that optimizes memory based on current and past tokens, achieving significant improvements on recall-intensive tasks. Lu and Li [2] proposed dynamic affective memory management employing Bayesian-inspired updates with memory entropy minimization. Chen et al. [3] presented TeleMem, achieving 19% accuracy improvement over Mem0 while reducing token usage by 43% through narrative dynamic extraction. Zhang et al. [4] introduced ACE (Agentic Context Engineering), yielding 10.6% gains on agent benchmarks while reducing adaptation latency by 86.9%.

However, these approaches predominantly employ static mechanisms for memory updates and deletion, making memory management as a deterministic process governed by fixed thresholds or recency heuristics. As conversation histories extend, important contextual information can easily be deleted due to the First-In-First-Out approach.

Inspired by the mechanisms of human long-term memory, MEMORIA introduces a dynamic memory architecture that learns user preferences, workflows, and communication patterns through continuous observation rather than explicit instruction. Unlike prior systems that treat memory as an archive subject to periodic pruning, our approach enables memory nodes to evolve through a confidence-weighted update mechanism that preserves high-utility information while slowly deprecating outdated or contradictory information.

2.2. The Context Inefficiency Crisis

The computational implications are equally significant. Enterprise teams utilizing AI tools now spend an average of $85,521 monthly on AI-native applications [6] with substantial resources wasted re-establishing context every session. Traditional approaches to AI personalization are fine-tuning

†First author and algorithm originator. *These authors contributed equally to application development. [1] University of Illinois Urbana-Champaign, USA. Correspondence to: Qiran Hu <qiranh2@illinois.edu>.

Preprint. February 2, 2026.

and post-training, which require significant investment that often cost $10,000–$30,000 for full model training [7], demand weeks of expertise, and produce opaque systems where users cannot inspect what the model has learned.

2.3. Market Validation

The recent addition of memory features to ChatGPT, Claude, and Gemini in 2024–2025 validates the market need for persistent AI context. Gartner predicts that 40% of enterprise applications will integrate task-specific AI agents by 2026, up from less than 5% in 2025 [8]. However, current implementations focus predominantly on declarative memory on explicit facts, preferences, and conversation history rather than procedural learning that captures how users work.

3. Proposed Features

MEMORIA introduces three key capabilities that distinguish it from existing memory-augmented AI systems.

Transparent Memory Interface. Unlike systems that treat learned information as opaque internal state, MEMORIA provides a dashboard where users can inspect stored memories in natural language, view confidence scores, trace memory origins to specific conversations, and directly edit or remove entries. This transparency addresses the critical limitation of current approaches where users have no visibility into what models learn about them.

Procedural Learning Engine. The core innovation lies in extracting *how* users work rather than merely storing *what* users have said. Through dual-channel analysis, MEMORIA identifies both explicit preferences and implicit behavioral patterns (workflows, temporal regularities, communication styles). This approach, validated in prior educational deployments with 101.5% improvement across metrics [9], enables personalization beyond surface-level preference matching.

Efficient Memory Retrieval. The system loads only contextually relevant memory nodes at inference time, reducing token consumption from approximately 100K tokens (full history) to approximately 5K tokens (relevant subset), which is a 95% reduction that enables cost-efficient operation while maintaining personalization quality.

4. Application Functionality

MEMORIA functions as a web-based AI assistant with integrated memory management. The system performs three primary operations during user interactions.

Memory Extraction. As users engage with the assistant, MEMORIA continuously analyzes conversations to identify preference signals, distinguishing between declarative information (stated preferences) and procedural patterns (workflow sequences, contextual behaviors) to construct an evolving personalized memory graph.

Context Augmentation. At inference time, the system retrieves relevant memory nodes based on semantic similarity, temporal recency, and historical utility scores. Retrieved memories augment the prompt context, enabling responses informed by accumulated user history without requiring full conversation logs.

Feedback Integration. User ratings and corrections propagate through the memory graph via Bayesian updates, adjusting confidence scores and retrieval priorities. This feedback loop enables continuous refinement without model retraining.

Users interact with MEMORIA through two interfaces: a *conversational interface* mirroring standard AI chat applications, where memory operates transparently in the background, and a *memory dashboard* for users to view, rate, manage, and guide how the system remembers and adapts. This dual-interface design balances seamless interaction with fine-grained user control, addressing both deployment efficiency and transparency requirements.

5. References

[1] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024.

[2] A. Behrouz et al. ATLAS: Learning to optimally memorize the context at test time. *arXiv:2505.23735*, 2025.

[3] J. Lu and Y. Li. Dynamic affective memory management for personalized LLM agents. *arXiv:2510.27418*, 2025.

[4] C. Chen et al. TeleMem: Building long-term and multimodal memory for agentic AI. *arXiv:2601.06037*, 2025.

[5] Q. Zhang et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv:2510.04618*, 2025.

[6] CloudZero. The state of AI costs in 2025. Technical Report, 2025.

[7] Educative Team. What is the cost of fine-tuning LLMs? *Dev Learning Daily*, 2025.

[8] Gartner, Inc. Gartner predicts 40% of enterprise apps will feature task-specific AI agents by 2026. Press Release, August 2025.

[9] Q. Hu. LTMBSE-ACE: Long-term memory-based self-evolving adaptive context engine. NOODEIA Project, SALT Lab, 2024.

6. Acknowledgement & Authorship

This project originates from the LTMBSE-ACE algorithm developed in the NOODEIA project at the SALT Lab for Human-AI Interaction [6], which demonstrated fine-tuning-level personalization at 1% token cost. Course concepts extend the project through production-ready web development, scalable backend architecture, memory visualization UI, and secure deployment.

Authorship: Qiran Hu originated the algorithm (first author). Amy Bisalputra, Ke Ding, Min Kim, and Kewen Xia contribute equally to application development.