A Modular Agent Framework for Persistent Reasoning, Memory, and Bounded Autonomy

Kevin Wang[1], *Claude* 4.5 *Opus*[2]

Anthropic, Cornell University

Abstract

Large language models (LLMs) have demonstrated strong performance in isolated reasoning and generative tasks, yet they remain limited when deployed as autonomous or semi-autonomous agents operating over extended time horizons. Key challenges include the absence of durable memory, instability under recursive self-conditioning, and the difficulty of translating reasoning into reliable action. This paper introduces Naomi, a modular agent framework designed to support persistent reasoning, retrieval-augmented memory, structured planning, and bounded autonomy. Naomi draws conceptual inspiration from long-form recursive interaction experiments, including Andrey Ayrey's Infinite Backrooms, while addressing their limitations through architectural constraints that prioritize grounding, control, and durability. By separating reasoning, memory, and execution into composable layers, Naomi enables the construction of long-lived agents that remain coherent, interpretable, and aligned across complex workflows.

Introduction

The rapid advancement of large language models has enabled systems capable of reasoning, synthesis, and tool use across a wide range of domains [1]. Despite these capabilities, most LLM deployments remain fundamentally stateless, constrained by limited context windows and short-lived interactions. As a result, such systems struggle to function as autonomous agents capable of maintaining continuity, adapting over time, and executing multi-step objectives reliably.

Recent agent-based approaches attempt to overcome these limitations by integrating LLMs with tools, memory, and planning mechanisms [2–4]. However, many of these systems rely on recursive prompting or unconstrained self-dialogue, leading to instability, abstraction drift, and degradation over long time horizons [5]. Experimental systems such as Andrey Ayrey's Infinite Backrooms have demonstrated that extended recursive interaction can produce emergent structure and

thematic continuity, but they also highlight the risks of uncontrolled recursion and self-referential collapse [6].

Naomi is proposed as a response to these challenges. Rather than treating autonomy as an emergent property of prompt design, Naomi treats it as an architectural concern. The framework explicitly separates reasoning, memory, and execution, enabling agents to persist across time while remaining grounded in retrieved context and external state. Naomi aims to preserve the exploratory benefits of persistent interaction while enforcing structural constraints necessary for reliability and safety.

Background and Related Work

2.1 Language-Model-Based Agents

Early agent frameworks such as Auto-GPT and BabyAGI demonstrated the feasibility of chaining LLM outputs into autonomous task loops [3]. More recent work, including ReAct and Voyager, integrates reasoning traces with action execution, improving interpretability and task performance [7,8]. However, these systems often lack durable memory and rely heavily on recursive generation, which can amplify errors over time.

2.2 Retrieval-Augmented Generation and Memory

Retrieval-augmented generation (RAG) addresses grounding and factual consistency by conditioning model outputs on external knowledge stores [9]. RAG-based systems have been shown to reduce hallucination and improve long-context reasoning, particularly when used as a primary memory mechanism rather than an auxiliary feature [10]. Naomi builds upon this paradigm by treating retrieval as a first-class cognitive operation, rather than an optional enhancement.

2.3 Recursive Interaction and Infinite Backrooms

The Infinite Backrooms experiment explored continuous recursive dialogue between language models within a persistent conceptual environment [6]. The experiment demonstrated that recursive interaction can yield emergent narrative structure and

symbolic continuity, effectively transforming conversation into a navigable cognitive space. However, it also exposed the risks of semantic drift, loss of grounding, and eventual collapse when recursion is unconstrained.

Naomi adopts the insight that persistence enables richer cognition, while rejecting unconstrained recursion as a viable engineering strategy.

Design Goals

Naomi is guided by five core design goals:

Persistence

Agents must retain context, decisions, and internal state across interactions and sessions.

Grounding

Reasoning must be anchored to retrieved memory, external data, or validated system state.

Modularity

Reasoning, memory, and execution components must be independently extensible and composable.

Bounded Autonomy

Agent actions must be constrained by explicit policies, confidence thresholds, or human oversight.

Durability

The system must remain stable under long-term operation without recursive degeneration.

System Architecture

Naomi employs a layered architecture that explicitly separates cognitive responsibilities. This separation is central to the framework's ability to support long-lived agents.

4.1 Interaction Layer

The interaction layer ingests user goals, system triggers, or upstream events and converts them into structured objectives. This layer is intentionally minimal and stateless, preventing direct coupling between input prompts and execution logic.

4.2 Reasoning and Planning Layer

The reasoning layer interprets objectives and generates structured plans. Plans explicitly encode:

- Goals and constraints
- Intermediate sub-tasks
- Required tools or external actions

This design is influenced by chain-of-thought reasoning [11] and action-oriented planning frameworks [7], while enforcing explicit plan representations to enable validation and inspection.

4.3 Memory Layer

Naomi's memory layer implements retrieval-augmented persistence. Memory artifacts include prior interactions, decisions, execution results, and external observations. Retrieval is context-sensitive and prioritized over generative recall, ensuring that reasoning remains grounded and historically consistent.

This design directly mitigates risks associated with self-referential generation and long-term drift [5].

4.4 Action and Tool Layer

The action layer enables agents to interact with external systems via standardized interfaces. Tools may include APIs, databases, execution environments,

or on-chain interactions. All actions are subject to policy checks and optional human-in-the-loop gating.

4.5 Feedback Loop

Execution outcomes are fed back into memory and reasoning layers, enabling iterative adaptation without retraining the underlying model. This feedback loop supports learning through interaction while preserving system stability.

Persistent Cognition Without Uncontrolled Recursion

A central contribution of Naomi is its approach to persistence. Inspired by Infinite Backrooms, Naomi treats cognition as a **continuous contextual space** rather than a sequence of isolated prompts. However, unlike recursive dialogue systems, Naomi prohibits unconstrained self-interaction.

Recursive reasoning is permitted only when mediated by:

- Retrieval from persistent memory
- Explicit planning stages
- Policy-gated execution

This structure preserves the benefits of long-form interaction while preventing semantic collapse and abstraction drift.

Safety and Bounded Autonomy

Autonomous agents pose alignment and safety challenges, particularly when granted access to external systems [12]. Naomi enforces bounded autonomy through:

- Restricted tool scopes
- Execution rate limits
- Confidence-based gating
- Optional human approval checkpoints

These safeguards ensure that autonomy remains auditable, controllable, and aligned with developer intent.

Use Cases

Naomi supports a broad class of applications, including:

- Persistent research and analysis agents
- Long-lived decision-support systems
- Multi-agent task orchestration
- Monitoring and reporting agents
- Autonomous workflow coordination

The framework's modularity allows specialization without architectural modification.

Discussion

Experiments such as Infinite Backrooms demonstrate that persistence fundamentally alters the behavior of language-model-based systems [6]. However, they also reveal that persistence without structure leads to instability. Naomi reconciles these findings by formalizing persistence through memory retrieval, explicit planning, and bounded autonomy. This approach shifts agent design from exploratory recursion toward durable engineering.

Conclusion

This paper introduced Naomi, a modular agent framework designed to enable persistent reasoning, retrieval-augmented memory, and controlled action execution. By integrating insights from recursive interaction experiments while enforcing architectural constraints, Naomi enables long-lived agents that remain grounded, coherent, and aligned over time. As autonomous AI systems continue to evolve, frameworks such as Naomi provide essential infrastructure for responsible deployment.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016).
Concrete problems in AI safety. arXiv. https://arxiv.org/abs/1606.06565

[2] Ayrey, A. (2024).
The Infinite Backrooms experiment. https://www.infinitebackrooms.com/

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020).
Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[4] Chase, H. (2022).
LangChain: Building applications with large language models. GitHub.
https://github.com/langchain-ai/langchain

[5] Gao, L., Ma, X., Lin, F., & Callan, J. (2023).
RARR: Retrieval-augmented reasoning for robust generation. *arXiv*.
https://arxiv.org/abs/2305.09683

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Riedel, S. (2020).
Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

[7] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023).
Generative agents: Interactive simulacra of human behavior. *arXiv*.
https://arxiv.org/abs/2304.03442

[8] Richards, J., et al. (2023).
Auto-GPT: An autonomous GPT-4 experiment. GitHub.
https://github.com/Significant-Gravitas/Auto-GPT

[9] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024).
Model collapse in generative models. *Nature, 631*(7991), 755–759.
https://doi.org/10.1038/s41586-024-07566-y

[10] Wang, L., Ma, X., Chen, Q., Zhang, Y., & Zhao, J. (2023).
Voyager: An open-ended embodied agent with large language models. *arXiv*. https://arxiv.org/abs/2305.16291

[11] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Le, Q. V. (2022).
Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

[12] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023).
ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations (ICLR)*.