AI-native Memory

A Pathway from LLMs Towards AGI

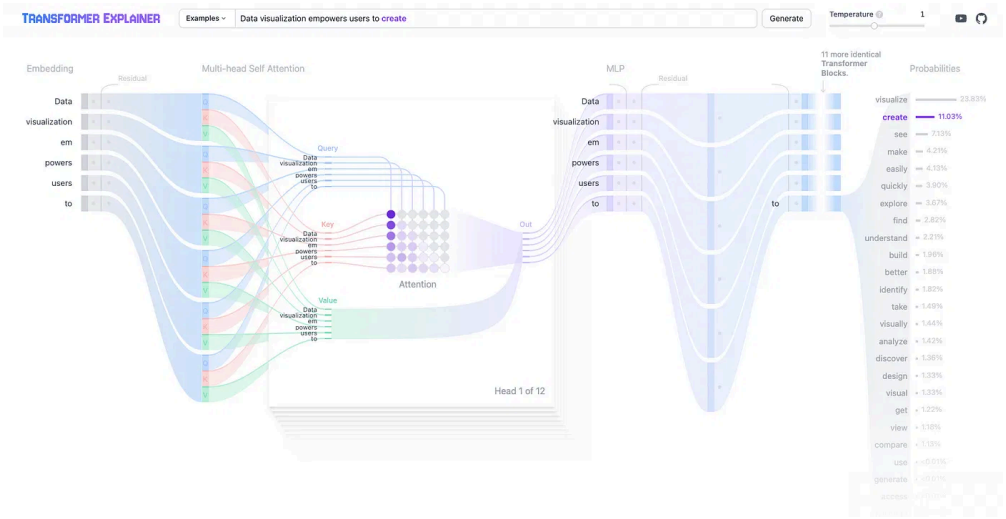
Based on research by Jingbo Shang, Zai Zheng, Jiale Wei, Xiang Ying, Felix Tao, and the Mindverse Team

arXiv:2406.18312 (Aug 28, 2024)

The Current LLM Landscape

- Remarkable progress in **large language models** with capabilities beyond language modeling
- Key models: **GPT series, Gemini, Claude, Llama, Mixtral** demonstrating significant potential in general task solving
- Capabilities include following complex instructions, reasoning, and performing diverse tasks
- Industry consensus: LLMs are becoming **fundamental building blocks** toward artificial general intelligence (AGI)

> Are we on the right path to AGI with current LLM approaches?



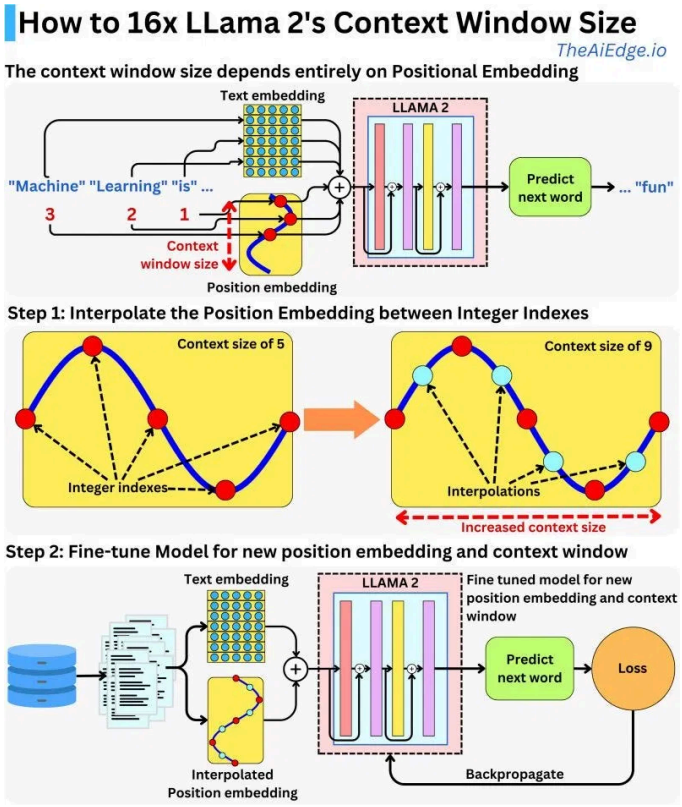
The Long-Context Assumption

- Industry belief: **Unlimited context length = AGI**
- Context evolution: **32K → 128K → 1M+** tokens
- The strategy: Put all raw data into context, let LLM handle everything in one step

Two Critical Assumptions:

- **Needle-in-haystack capability:** LLMs can find necessary information from super long context
- **Long-context reasoning capability:** LLMs can conduct all required inferences in one step

Both must be true simultaneously

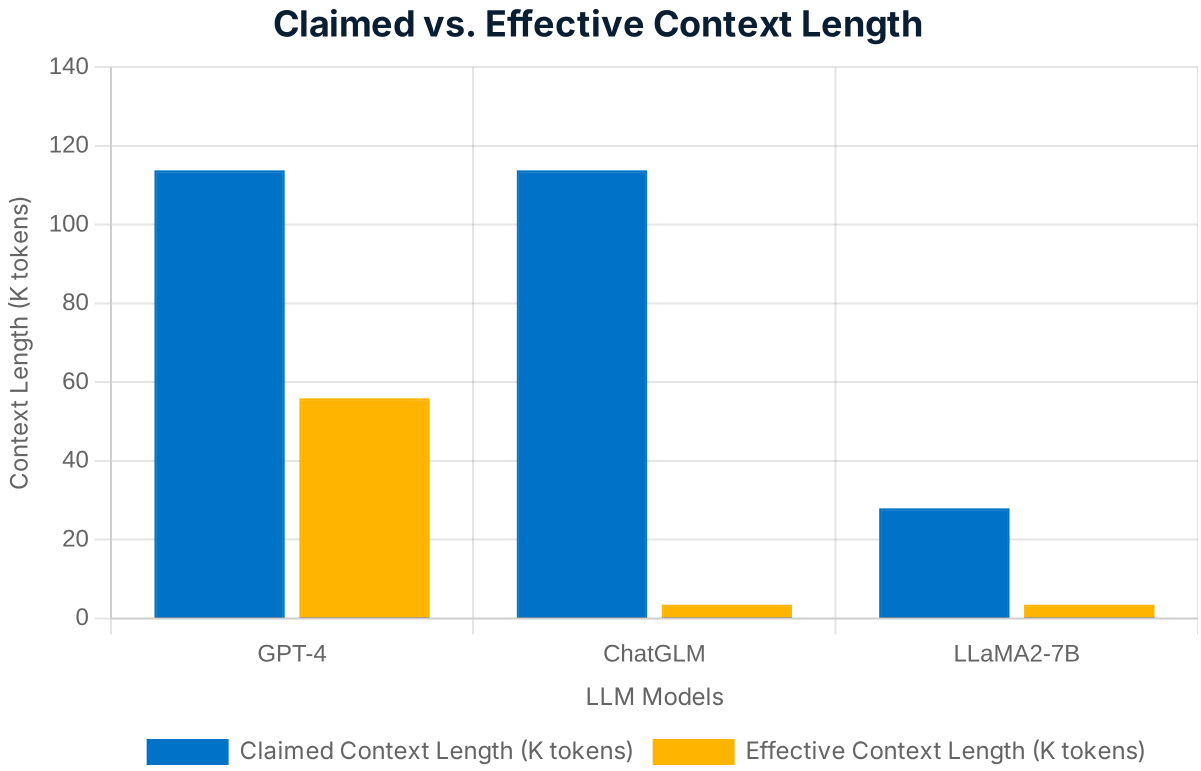


Reality Check: Effective vs. Claimed Context

- Research shows a significant **gap** between effective and claimed context lengths
- Effective context length is defined as the maximum length where LLM outperforms a strong baseline

GPT-4:	Claims **128K**, effective **~64K**
ChatGLM:	Claims **128K**, effective **~4K**
LLaMA2-7B:	Claims **32K**, baseline **4K**

> Super long effective context is very difficult to achieve, and the effective context size in existing solutions has not fundamentally improved

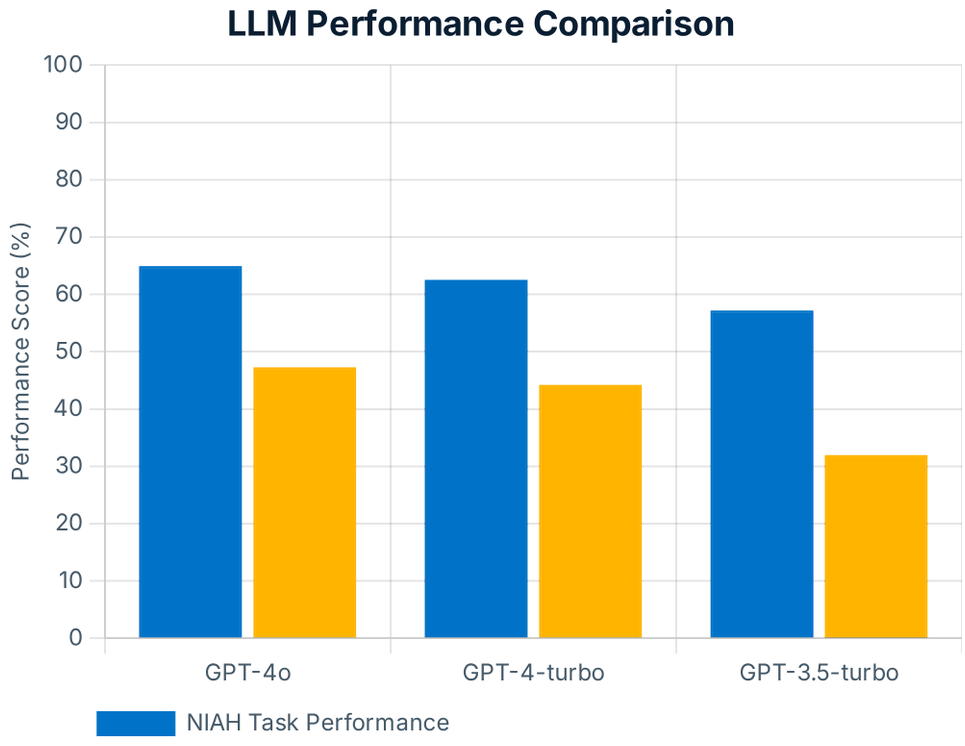


Beyond Needle-in-Haystack

- **Traditional NIAH:** Simple retrieval tasks that only test finding information
- **Real-World Challenge:** Reasoning-in-a-haystack requires both finding information AND conducting complex reasoning
- **Experimental Evidence:** Current LLMs struggle significantly with combined tasks
- **Implication:** Long-context approach has fundamental limitations for AGI

NIAH vs. Reasoning-in-a-Haystack

Traditional NIAH	Reasoning-in-a-Haystack
• Find specific information • Single-step retrieval • No complex reasoning	• Find multiple pieces of information • Multi-hop reasoning • Complex inference required



The Memory-Centric Vision

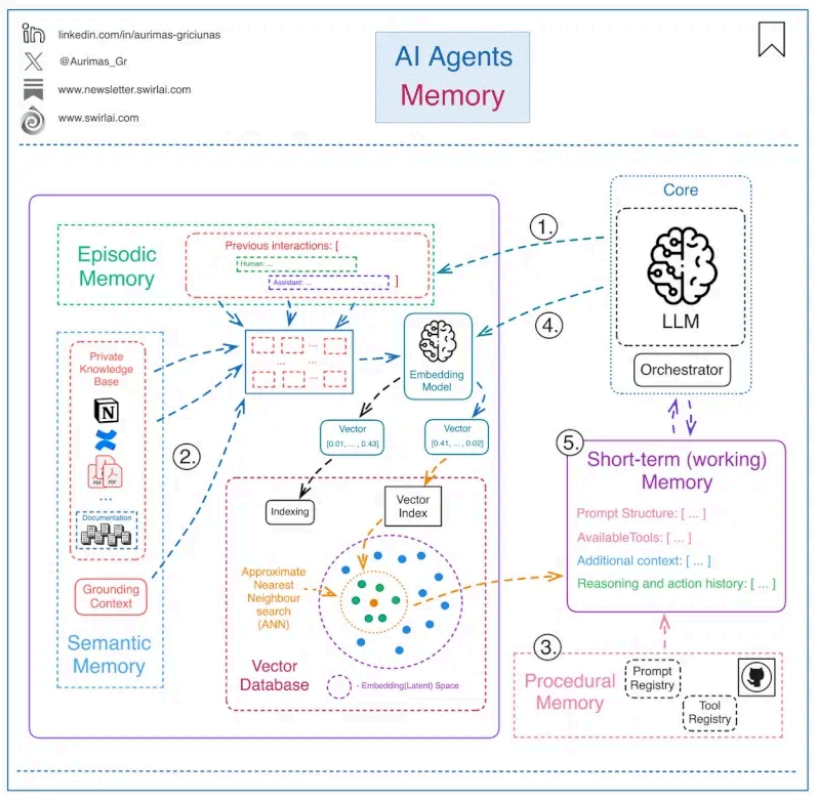
- **New Paradigm:** Memory as the key to AGI, not just unlimited context
- **Beyond RAG:** Not just raw data retrieval, but organized, processed information
- **User Accessibility:** Memory should be directly consumable by users

Architecture Analogy:

LLMs = Processors
Core reasoning engines

Context = RAM
Short-term working memory

Memory = Disk Storage
Long-term organized knowledge



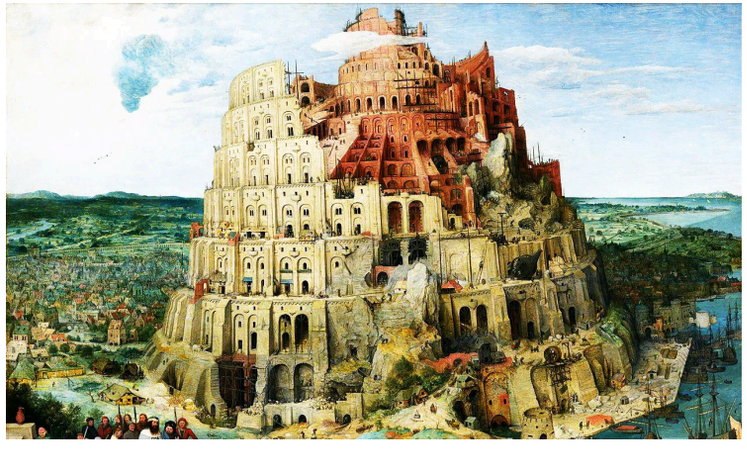
AI-native Memory Characteristics

Two Approaches to Memory Implementation

Approach 1: Information Extraction/Generation

The "Memory Palace" Approach

- Constructing organized storage of processed information
- Structured, searchable information repository
- Enhanced retrieval-augmented generation
- Natural language interface for direct user consumption

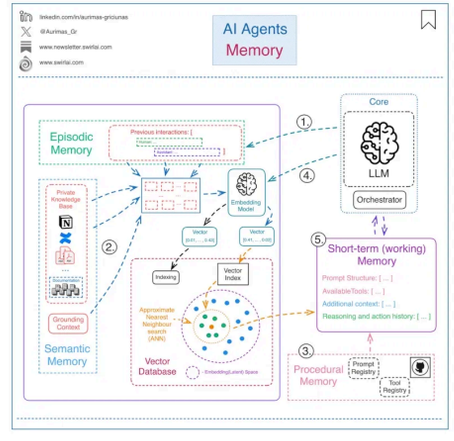


VS

Approach 2: Neural Network Compression

The "Neural Memory" Approach

- Compressing memory as neural networks
- Potentially using LLMs themselves as memory
- Deep integration with reasoning capabilities
- Efficient parameter storage of processed information



Experimental Methodology

Reasoning-in-a-Haystack Evaluation: Testing both retrieval and reasoning capabilities simultaneously

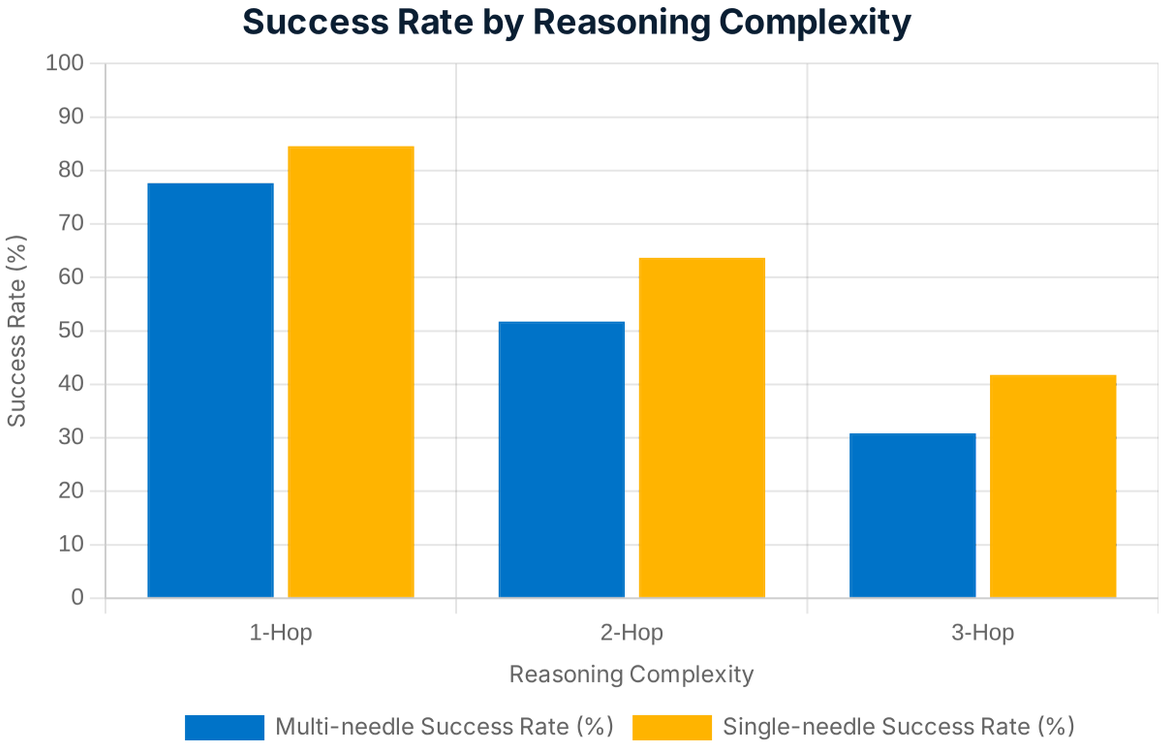
Real-World Data: Based on Mebot user interactions with proper consent and privacy protection

Complexity Levels: 1, 2, and 3 hops of reasoning required to test different depths

Provider LLMs: GPT-4o, GPT-4-turbo, GPT-3.5-turbo as evaluation targets

Experimental Design:

- **Multi-needle:** Information distributed at depths of 20%, 40%, 60%, 80%
- **Single-needle:** All information combined at 40% or 60% depth
- **Evaluation Criteria:** True answers manually refined for accuracy



The Path Forward

> **Memory, not unlimited context, is the pathway to AGI**

Paradigm Shift

From context length to memory architecture as the foundation for AGI development

Research Directions

Memory organization, compression techniques, and efficient retrieval mechanisms

Industry Impact

Rethinking AGI development strategies with memory-centric approaches

Challenges

Privacy, security, scalability, and ethical considerations in memory systems