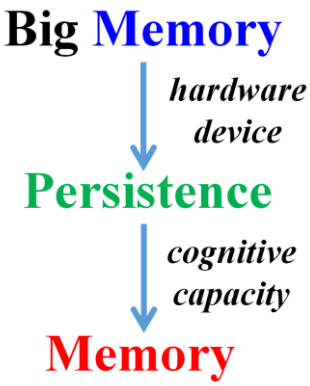
The Persistence of Big Memory is Memory

Yu Hua
Huazhong University of Science and Technology
csyhua@hust.edu.cn

Traditional computer architecture rigidly distinguishes volatile memory for runtime computation and non-volatile storage for persistent data archiving [1] [2]. This long-standing dichotomy has constrained the evolutionary boundary of artificial intelligence (AI), confining large language models (LLMs) to transient, stateless, and short-window intelligence. This paper proposes an academic assertion: ***the persistence of big memory is memory***. Here, the first "memory" refers to computer memory as a hardware device, and the second denotes human memory as the cognitive capacity to retain information. This academic assertion is a fundamental paradigm shift for sustainable artificial intelligence.



Different from conventional DRAM-based memory with volatility, big memory is defined as a heterogeneous, multi-level, cross-node, and capacity-scalable unified memory system that integrates diverse storage-memory devices and provides a consistent abstract view for upper-layer applications [3]. Its essential technological breakthrough lies in persistent memory enhancement, which essentially transforms data storage into cognitive memory in machines. This study further verifies a linear-like correlation between memory persistence scale and AI intelligence level, demonstrating that big memory persistence is the fundamental driver for AI to evolve from transient reactive intelligence to sustainable cognitive intelligence.

For decades, the computer industry has formed a fixed cognitive framework. Specifically, memory refers to volatile runtime caching with high speed and limited capacity, and persistence belongs to disk-based storage systems responsible for long-term data retention [4]. In this traditional division of labor, memory is merely a temporary data carrier for program execution, and persistence is an auxiliary functional attribute of storage devices. Such a rigid partition adapts to the operation

logic of traditional deterministic computing tasks but completely fails to match the cognitive evolution law of artificial intelligence. This architectural design leads to the inherent limitations of early large language models, which are essentially transient intelligent agents lacking effective memory persistence. Early LLMs operate within fixed context windows, with interactive records, user preferences, task experiences, and cognitive traces cleared after session termination. Without persistent memory inheritance, each inference of the model is a zero-start independent calculation, resulting in fragmented interaction logic, poor user adaptation, limited personalized cognition, and constrained performance.

The emergence of big memory changes the traditional definition of computer memory and reconstructs the underlying infrastructure of AI cognition. Different from single-structured traditional memory, big memory is a comprehensive heterogeneous memory system with core characteristics of large-capacity scaling, multi-device heterogeneity, cross-node clustering, and hierarchical storage differentiation [5][6]. It integrates volatile DRAM, non-volatile memory chips, solid-state storage, and cross-server distributed memory resources, shields the hardware differences of underlying heterogeneous devices through unified scheduling and abstract views, and provides upper-layer AI programs and user scenarios with consistent, transparent, and scalable memory services [7][8]. More importantly, the fundamental innovation of big memory is not limited to capacity expansion and hardware integration, but the endogenous integration of persistent capabilities into the memory architecture, thereby implementing the organic unity of runtime calculation and long-term data retention in the memory layer.

The essential logic of intelligence evolution originates from memory accumulation, which is a universal law of cognitive evolution applicable to both biological intelligence and machine intelligence. Biological intelligence presents an obvious positive correlation between memory persistence and intelligent level. A useful illustration is the relationship between memory persistence and intelligence level. For example, goldfish have a brief working-memory window, and their cognition is largely confined to transient reactive behaviors. However, humans accumulate memories over decades. This extended temporal horizon enables logical reasoning, experiential abstraction, preference formation, and iterative learning, collectively supporting advanced cumulative intelligence. The duration of memory persistence serves as a fundamental determinant of cognitive capacity. The intelligent difference between models is essentially the difference in memory scale, memory persistence ability, and memory inheritance efficiency.

The core dimensionality-reduction innovation of big memory persistence is to break the context window limitation of traditional LLMs and implement the transformation from transient memory to long-term sustainable memory. Traditional LLMs only have working memory limited by token windows, belonging to typical short-term transient memory, which cannot retain historical interaction experience and user feature information. In contrast, the multi-level persistent memory system built by big

memory enables AI agents to permanently retain user preferences, historical task records, interactive semantic features, and domain learning experiences. This fundamental change constructs a linear-like correlation between memory window length and AI intelligent level. The richer the persistent memory reserves, the more sufficient the cognitive basis of the model, the stronger the reasoning iteration and adaptive learning ability, and the higher the upper limit of artificial intelligence.

Big memory persistence is not an auxiliary storage optimization, but the essential implementation form of machine memory. The traditional view regards persistence as an additional technical feature of memory devices. This study redefines its essence: the persistence of big memory is memory. The persistent capability of big memory improves the simple data storage attribute of traditional storage hardware and endows machines with real cognitive memory attributes. Data retained by big memory persistence is no longer cold archived data, but active cognitive resources that can be called, learned, and iterated in real time by AI models. It is precisely this persistent memory capability that enables AI to avoid the dilemma of one-time transient inference, achieve continuous accumulation of experiences, iterative optimization of cognition, and sustainable evolution of capabilities, and complete the qualitative leap from passive reactive intelligence to active cognitive intelligence.

Big memory builds the underlying cognitive foundation for sustainable artificial intelligence through heterogeneous integration and persistent memory reconstruction. Persistence is not a peripheral function of big memory, but its core essence and intrinsic value. The assertion that the persistence of big memory is memory reveals the underlying physical and logical law of AI intelligence evolution, provides a brand-new theoretical paradigm for the design of next-generation AI infrastructure, and points out the core technical direction for breaking through the bottleneck of current artificial intelligence and implementing advanced sustainable cognitive intelligence.

Reference

[1]. Shihang Li, Matthew Giordano, Tushar Garg, Rohan Kadekodi, Daniel S. Berger, Baris Kasikci, Thomas Anderson, Simon Peter, "Finding NEMO: Nimble and Expressive Memory Observability", Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2026.

[2]. Yueyang Pan, Yash Lala, Musa Unal, Yujie Ren, Seung-seob Lee, Abhishek Bhattacharjee, Anurag Khandelwal, Sanidhya Kashyap, "Scalable Far Memory: Balancing Faults and Evictions", Proceedings of ACM Symposium on Operating Systems Principles (SOSP), 2025.

[3]. Yu Hua, Xue Liu, Ion Stoica, "The Golden Rule of Big Memory: Persistence is not Harmful", Communications of the ACM (CACM), Vo.69, No.5, 2026, Pages: 51-55.

[4]. Vinay Banakar, Suli Yang, Kan Wu, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Kimberly Keeton, "OBASE: Object-Based Address-Space Engineering to Improve Memory Tiering", Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2026.

[5]. João Barreto, Daniel Castro, Paolo Romano, Alexandro Baldassin, "FUR: Fast and Unlimited Reads on Persistent Memory Transactions", Proceedings of European Conference on Computer Systems (EuroSys), 2026.

[6]. Jiyu Hu, Seokjoo Cho, Landon Johnson, Kiran Hombal, Shreesha Gopalakrishna Bhat, Marcos K. Aguilera, Ramnatthan Alagappan, Aishwarya Ganesan, "MEGALON: Efficient Data Sharing for Partly Coherent CXL Memory", Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2026.

[7]. Yu Hua, "Big Memory Systems", Springer Nature, January 2026.

[8]. Nanqinqin Li, Yuhong Zhong, Asaf Cidon, Michael J. Freedman, "Harvesting Sub-Microsecond CXL Memory Stalls with LiteSwitch", Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2026.