13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Discrete and Continuous Memory Architectures to Prevent Digital Dementia in LLM Agents

Tibor Tajti

Eszterházy Károly Catholic University, and University of Debrecen, Faculty of Informatics
tajti.tibor@uni-eszterhazy.hu,tajti.tibor@inf.unideb.hu

Abstract

LLM agents suffer from progressive memory failures—context loss, goal drift, confabulation, and strategy amnesia—mirroring cognitive decline in dementia. We call this *digital dementia* and measure it with a per-step symptom rate. To counter it, we compare two memory architectures: a *discrete* design where the agent explicitly picks which memory layer to query, and a *continuous* design where a single vector-similarity search retrieves from all memory stores at once, weighted by relevance. Both designs account for the practical constraints of context-window limits and the token cost of long prompts. An ablation on coding and planning tasks of 10–100 steps shows that continuous attention lowers symptom frequency on longer tasks, while discrete routing remains more interpretable.

Keywords: large language models, autonomous agents, memory architecture, context window management, vector similarity search, digital dementia

1. Introduction

LLM-based agents perform well on short tasks [7], but on longer ones their accuracy drops as the context window fills up [3]—mirroring human cognitive decline. The brain counters this through specialized subsystems: working, episodic, semantic, and procedural memory [1, 2, 6]. Systems like MemGPT [4] and Generative Agents [5] add multi-level memory to agents, but they all rely on *discrete routing*:

the agent or a controller explicitly picks which layer to use. Moreover, none of them treat context-window limits and the associated token cost as a first-class design constraint. To our knowledge, no previous work has framed these failures as a form of measurable cognitive decline, nor compared discrete routing with *continuous, similarity-weighted memory attention*.

2. Digital Dementia: Risk Factors and Metric

We identify five risk factors (Table 1) and define the *Digital Dementia Index* for a task of T agent steps as $\mathrm{DDI}(T) = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} f_i(T)/T$, where $\mathcal{S} = \{s_1, \ldots, s_5\}$ are the risk factors and $f_i(T)$ counts occurrences of symptom s_i. Lower DDI means more effective prevention.

Table 1. Digital dementia risk mapping.

Clinical symptom	**Agent failure**	**Diagnostic indicator**
Short-term memory loss	Context overflow	Task-critical info absent from context
Executive dysfunction	Goal drift	Subtask diverges from objective
Confabulation	Action hallucination	References to non-performed actions
Temporal disorientation	Step repetition	Re-execution of completed steps
Procedural memory loss	Strategy amnesia	Failure to reuse successful patterns

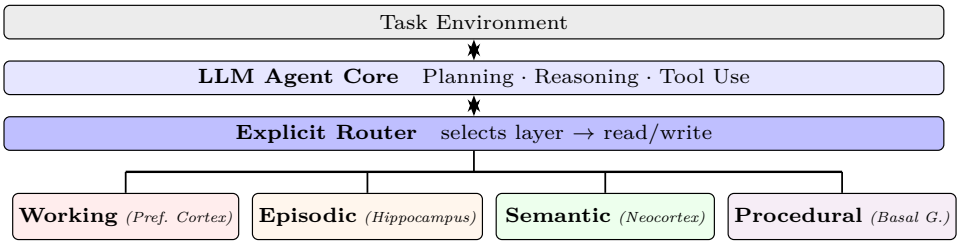
3. Memory Architectures Against Digital Dementia

Both architectures use four biologically inspired stores: **Working** (context manager / prefrontal cortex) prevents *context overflow*; **Episodic** (action log / hippocampus) prevents *goal drift* and *step repetition*; **Semantic** (vector knowledge base / neocortex) prevents *confabulation*; **Procedural** (strategy templates / basal ganglia) prevents *strategy amnesia*. They differ in how the agent *accesses* these stores (Figure 1).

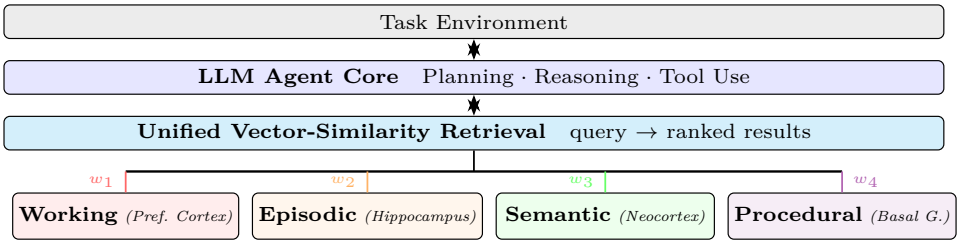
In the **discrete** design, a router picks one layer at a time—easy to interpret and ablate. In the **continuous** design, the agent's context is matched against all stores at once; similarity scores act as natural weights, so the LLM sees a ranked list rather than choosing a layer. This mirrors how human recall works: several memory systems fire together, each contributing in proportion to relevance [2].

4. Conclusions and Future Work

We framed progressive memory failures in LLM agents as *digital dementia* and compared two ways to prevent it: discrete layer routing and continuous similarity-weighted attention. An ablation over tasks of 10–100 steps, measuring DDI, completion rate, and per-symptom frequency, confirms that both architectures reduce cognitive decline, with continuous attention especially effective on longer tasks. Future work will explore adaptive weight tuning and memory transfer across tasks.



(a) Discrete: explicit router selects which layer(s) to query.



(b) Continuous: all stores queried simultaneously; similarity scores yield weights w_i.

Figure 1. Two architectures sharing four memory stores. (a) Discrete routing selects layers explicitly. (b) Continuous attention retrieves from all layers weighted by relevance.

References

[1] R. C. Atkinson, R. M. Shiffrin: *Human Memory: A Proposed System and Its Control Processes*, in: Psychology of Learning and Motivation, vol. 2, Academic Press, 1968, pp. 89–195, doi: 10.1016/S0079-7421(08)60422-3.

[2] A. Baddeley: *The Episodic Buffer: A New Component of Working Memory?*, Trends in Cognitive Sciences 4.11 (2000), pp. 417–423, doi: 10.1016/S1364-6613(00)01538-2.

[3] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang: *Lost in the Middle: How Language Models Use Long Contexts*, Transactions of the Association for Computational Linguistics 12 (2024), pp. 157–173, doi: 10.1162/tacl_a_00638.

[4] C. Packer, V. Fang, S. G. Patil, K. Lin, S. Wooders, J. E. Gonzalez: *MemGPT: Towards LLMs as Operating Systems*, arXiv preprint arXiv:2310.08560 (2023), doi: 10.48550/arXiv.2310.08560.

[5] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein: *Generative Agents: Interactive Simulacra of Human Behavior*, in: Proc. UIST, 2023, pp. 1–22, doi: 10.1145/3586183.3606763.

[6] E. Tulving: *Episodic and Semantic Memory*, in: Organization of Memory, Academic Press, 1972, pp. 381–403.

[7] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, et al.: *A Survey on Large Language Model based Autonomous Agents*, Frontiers of Computer Science 18.6 (2024), p. 186345, doi: 10.1007/s11704-024-40231-1.