TO KEEP OR TO FORGET: TOWARD CONTEXT-SENSITIVE MEMORY IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While attention mechanisms share formal connections with associative retrieval, they lack core features of biological memory systems, particularly context-sensitive memory formation and recall. Inspired by recent cellular neurobiological evidence on pyramidal two-point neurons (TPNs), we propose a TPN-inspired memory mechanism for large language models (LLMs) that enables context-sensitive memory writing, reading, and updating through triadic modulation loops. In this framework, global contextual signals modulate local feedforward representations to selectively engage memory operations only when local evidence is coherent with the global internal state. Memory encoding and retrieval emerge from apical-amplification regimes, while memory updating is governed by apical-drive dynamics. This context-sensitive memory gating reduces interference between overlapping representations and enables more stable and coherent associative memory formation and retrieval during ongoing feedforward (FF) processing.

1 INTRODUCTION

There has been growing interest in memory-augmented large language models (LLMs), where architectures extend transformers with long-term external stores or semantic memory retrieval mechanisms Wang et al. (2023); Lewis et al. (2020). These approaches include both pre-training strategies that integrate retrieval into the learning objective and post-hoc augmentation modules that augment trained LLMs with structured memory stores Guu et al. (2020); Borgeaud et al. (2022); Wu et al. (2025)).

Writing, reading, and combining retrieved memory helps store relevant associations and local information, reducing dependence on attention or deeper layers to reconstruct representations that could instead be retrieved. For example, recent work by DeepSeek suggests that specialized associative lookup modules can offload canonical pattern reconstruction from deep transformer computation and improve representational associations Cheng et al. (2026); related retrieval-based mechanisms have been shown to improve token prediction and knowledge access in large language models Khandelwal et al. (2019); Borgeaud et al. (2022).

Recent cellular neurobiological discoveries Larkum (2013); Kastellakis et al. (2023) suggest that current associative-memory mechanisms lack an account of how knowledge can be dynamically stored, recalled, and integrated into ongoing processing through context-sensitive information processing. Specifically, pyramidal TPNs Larkum et al. (1999); Phillips et al. (2024), using their apical dendrites as contextual signals to gate FF information arriving at basal compartments, enable context-dependent transmission and modulation, potentially enhancing associative-memory capacity and flexibility.

We therefore propose a TPNs-inspired memory mechanism for LLMs. Using TPN-inspired dynamics to enable context-sensitive writing, reading, and modulation of ongoing FF processing, memory write, read, and combine operations are triggered only when TPN-like mechanisms establish the relevance of local FF information within a global contextual state. This could reduce interference between overlapping representations while improving memory selectivity and coherence.

2 TPN-INSPIRED MEMORY-ENABLED LLMS

Our proposed method targets three major memory-related operations that, if aligned with TPN working principles, could enable a dynamic and continually adaptive memory module grounded in cellular-biology principles. This module enhances the current information-processing stream by efficiently storing, recalling, and using relevant knowledge in context. Modern surveys of memory mechanisms in LLMs emphasize the need for structured write, read, and update processes to support dynamic memory representations Zhang et al. (2025).

In neuroscience and cognitive modeling, predictive coding frameworks propose that memory formation and consolidation are shaped by prediction error signals, such that unexpected or surprising inputs are more strongly encoded and maintained Barron et al. (2020). Prediction error has been shown to influence the organization of episodic memories and drive re-consolidation processes, consistent with models in which prediction error serves as a key modulatory signal for memory updating Sinclair & Barense (2019). Prediction error and surprise are also closely tied to memory consolidation processes that determine whether an experience is encoded into long-term storage Spens & Burgess (2024). Organisms preferentially store representations when ongoing inputs deviate significantly from expectations.

However, recent cellular neurobiological evidence Phillips et al. (2024); Marvan & Phillips (2024); Graham et al. (2025) suggests that predictions sent as context from higher to lower levels of processing attenuate FF sensory signals that align with those predictions, while mismatched inputs propagate upward as prediction errors. Evidence on apical function challenges purely subtractive prediction-error accounts: basal and apical inputs cooperate to amplify neuronal output when they agree Marvan & Phillips (2024).

Amplification here refers to neuronal bursting, brief high-frequency firing, when both basal and apical compartments are simultaneously depolarized, signalling a match between context and input. This is formally described as apical-amplification (AA) MOD Phillips et al. (2024); Graham et al. (2025) when FF input drives neuronal output, and apical-drive (AD) MOD when contextual input dominates. These processes are enabled through triadic modulation loops (TMLs) among TPNs Zain et al. (2025).

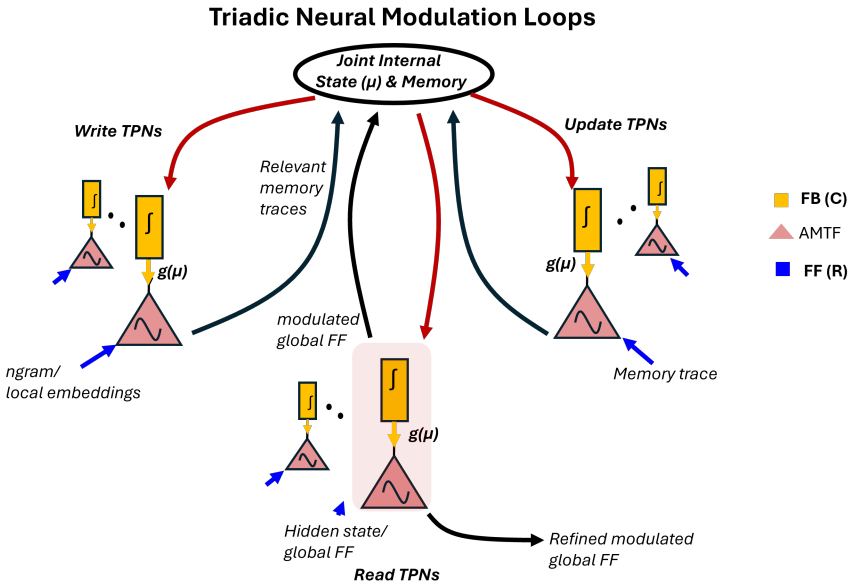

Figure 1: TPN-inspired context-sensitive memory mechanism. Global contextual (C) feedback (FB) signals, represented by the joint internal state and memory, modulate local feedforward (FF) representations through asynchronous triadic Modulation Transfer Functions (AMTFs). This enables memory write, read, and combination operations only when local evidence is coherent with the global contextual state.

Figure 1 illustrates how TPN operations via TMLs could enable context-sensitive writing, reading, and modulation of ongoing FF processing to improve memory selectivity and coherence. Specifically, local FF signals (blue arrows) are processed via nonlinear transformations (MOD) using joint internal state and memory μ as internal predictions to form candidate representations that may be written to, updated, or retrieved from memory. Internal predictions as global contextual state, represented by the dynamic generative controller $g(\mu)$, provide top-down signals (red arrows) analogous to apical dendritic input in three pyramidal TPNs populations. Memory write, read, and combination operations (yellow integration blocks) are triggered only when local FF evidence is coherent with the global contextual state. This context-sensitive gating ensures that only relevant representations participate in memory operations, reducing interference between overlapping representations while improving memory selectivity, stability, and representational coherence. Within this framework, different TPN modulation regimes correspond naturally to distinct memory operations. AA for Write and Read TPNs to enable context-sensitive writing and reading. In contrast, AD for Update TPNs, providing internally guided memory updating and consolidation.

2.1 WRITE TO MEMORY

A fundamental challenge in memory-augmented LLM research is determining what information should be written into memory and when. Write-TPNs aim to store locally compressed and contextually relevant high-utility representations. These may include n-grams derived from contextualized embeddings, entropy-clustered embeddings, or novelty-based tokens serving as meaningful contextual cues when retrieved later. In the proposed framework, memory writing is governed by an AA regime, where agreement between FF evidence and contextual predictions produces cooperative depolarization. This AA regime ensures that only representations coherent with the global contextual state are encoded into memory, reducing interference between overlapping representations. Under this mechanism, write-TPNs serve two complementary roles: selecting contextually relevant representations for encoding, and accelerating convergence toward stable encoded representations with reduced reliance on backpropagation.

2.2 READ FROM MEMORY

Even with selective memory writing, a model must recall relevant information based on associative cues or context. Guided recall under TPN principles aims to both resolve ambiguity and improve next-token prediction. Memory retrieval is governed by an AA regime, where contextual predictions and incoming FF evidence cooperate to amplify memory traces that best align with the current global state. Unlike similarity-only retrieval mechanisms, recall decisions depend on predictive gain and semantic alignment. When predictive mismatch is low, retrieval is suppressed; when mismatch increases, candidate memory traces are retrieved and ranked according to their expected contribution to restoring predictive coherence in the hidden state. Retrieved knowledge modulates FF processing through apical-style contextual gating, enabling memory to directly influence ongoing computation.

2.3 UPDATE MEMORY

Retrieved traces must be updated, strengthened, or weakened over time. This process is governed by TPN dynamics operating in an AD regime, where contextual signals dominate neuronal output and guide internal memory revision. Following memory retrieval and injection into the processing stream, predictive mismatch is recomputed. Reduction in mismatch serves as a consolidation signal determining whether the retrieved trace should be reinforced, revised, or weakened. In this AD regime, internally generated contextual signals guide memory updating independently of immediate FF, enabling memory consolidation, restructuring, and long-term stability.

3 CONCLUSION

We hypothesize that the proposed approach enables a transition from static memory retrieval toward dynamic, context-sensitive memory systems, where internal predictive states guide writing, reading, and updating through TPN triadic modulation loops, improving representational alignment and associative-structure formation. Implementation and empirical evaluation of the proposed framework are currently ongoing.

REFERENCES

Helen C. Barron, Ryszard Auksztulewicz, and Karl Friston. Prediction and memory: A predictive coding account. *Progress in Neurobiology*, 192:101821, May 2020. doi: 10.1016/j.pneurobio.2020.101821. URL https://www.sciencedirect.com/science/article/pii/S0301008220300769.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Yukun Li, et al. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *arXiv preprint arXiv:2601.07372*, 2026.

Bruce P Graham, Jim W Kay, and William A Phillips. Context-sensitive processing in a model neocortical pyramidal cell with two sites of input integration. *Neural Computation*, 37(4):588–634, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

George Kastellakis, Simone Tasciotti, Ioanna Pandi, and Panayiota Poirazi. The dendritic engram. *Frontiers in Behavioral Neuroscience*, 17:1212139, 2023.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2019.

Matthew Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36(3):141–151, 2013.

Matthew E Larkum, J Julius Zhu, and Bert Sakmann. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398(6725):338–341, 1999.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Tomáš Marvan and William A Phillips. Cellular mechanisms of cooperative context-sensitive predictive inference. *Current Research in Neurobiology*, pp. 100129, 2024.

W. A. Phillips, T. Bachmann, W. Spratling, L. Muckli, L Petro, and T. Zolnik. Cellular psychology: relating cognition to context-sensitive pyramidal cells. *Trends in Cognitive Sciences*, 2024.

Alyssa H. Sinclair and Morgan D. Barense. Prediction error and memory reactivation: How incomplete reminders drive reconsolidation. *Trends in Neurosciences*, 42(10):727–739, September 2019. doi: 10.1016/j.tins.2019.08.007. URL https://pubmed.ncbi.nlm.nih.gov/31506189/.

Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *Nature Human Behaviour*, 8(3):526–543, January 2024. doi: 10.1038/s41562-023-01799-z. URL https://www.nature.com/articles/s41562-023-01799-z#citeas.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv e-prints*, pp. arXiv–2504, 2025.

Noor Ul Zain, Mohsin Raza Naseem, and Ahsan Adeel. Single layer tiny co4 outpaces gpt-2 and gpt-bert. In *Proceedings of the First BabyLM Workshop*, pp. 313–322, 2025.

Dianxing Zhang, Wendong Li, Kani Song, Jiaye Lu, Gang Li, Liuchun Yang, and Sheng Li. Memory in large language models: Mechanisms, evaluation and evolution. *arXiv e-prints*, pp. arXiv–2509, 2025.