A COGNITIVE LENS ON SELECTIVE MEMORY IN NEURAL SEQUENCE MODELS: SURPRISE, REPLAY, AND CONSOLIDATION

Sohyung Kim[1,*], **Jea Kwon**[2]
[1]Independent Researcher, [2]Max Planck Institute for Security and Privacy (MPI-SP)
ks000225@gmail.com, onlytojay@gmail.com

ABSTRACT

Recent sub-quadratic sequence models—including Mamba, DeltaNet, and Titans—have independently converged on selective memory mechanisms allocating write intensity proportional to input surprise, mirroring surprise-gated encoding in human episodic memory. We argue this convergence reflects shared computational pressures studied under complementary learning systems (CLS) theory. However, current models capture only encoding; the critical post-encoding processes of consolidation, replay, and reconsolidation remain absent. Drawing triangular connections among (i) prioritized experience replay in reinforcement learning, (ii) surprise-gated writing in sequence models, and (iii) sleep-dependent consolidation in cognition, we identify three architectural gaps: lack of offline replay across memory systems, absence of importance-sampling correction for selective writes, and missing reconsolidation pathways. We formalize each gap, propose mechanisms, and derive testable predictions at both machine learning and cognitive modeling levels, charting a design space for cognitively grounded sequence architectures.

1 INTRODUCTION

A striking pattern has emerged: architectures achieving best performance in the sub-quadratic regime consistently employ selective memory control. Mamba's data-dependent gating (Gu & Dao, 2024), DeltaNet's error-correcting delta rule (Yang et al., 2024; 2025), and Titans' prediction-error-based surprise (Behrouz et al., 2025) all instantiate the same principle—not all inputs deserve equal treatment in memory.

This principle has deep roots in cognitive science. Encoding strength is modulated by novelty, prediction error, and arousal (Tulving, 1972; McGaugh, 2004). The complementary learning systems (CLS) framework (McClelland et al., 1995; Kumaran et al., 2016) provides theoretical grounding: a dual-system architecture with fast episodic store (hippocampus) and slow semantic store (neocortex) optimally resolves stability–plasticity tradeoffs.

Recent work has begun mapping CLS onto hybrid models. Dong et al. (2025) interpret attention as "snapshot memory" and SSM as "fading memory." Irie et al. (2025) validate this through CLS at scale. Behrouz et al. (2025) posit short-term, long-term, and persistent memory modules.

The gap. These mappings focus on *encoding*—what to write and with what intensity. In human memory, encoding is only the beginning. Post-encoding processes—consolidation during sleep (Stickgold, 2005), reconsolidation upon retrieval (Nader et al., 2000), and prioritized replay (Schaul et al., 2016)—are equally critical. These have computational counterparts in reinforcement learning's prioritized experience replay (PER), yet remain absent from sequence models.

[*]Corresponding author. This independent research was conducted prior to the author joining the KAIST Graduate School of AI.

Table 1: Encoding vs. post-encoding: what's present and what's missing in current models.

Human Memory	Sequence Model	Credit	✓/×
Episodic/working	Attention (KV cache)	D, I	✓
Semantic/procedural	SSM/linear recur.	D, I	✓
Surprise-gated write	Selective gates	B, Y	✓
Sleep consolidation	—	—	×
IS correction	—	—	×
Reconsolidation	—	—	×

D=Dong et al. (2025); I=Irie et al. (2025); B=Behrouz et al. (2025); Y=Yang et al. (2025).

This paper. We identify three post-encoding architectural gaps (§3), formalize mechanisms (§4), and derive testable predictions (§5). Our contribution is a design space grounded in the triangular correspondence among human memory, RL replay, and sequence models.

2 Background: Encoding-Level Convergence

Modern sub-quadratic models maintain fixed-size recurrent state $\mathbf{S}_t \in \mathbb{R}^{d \times d}$ updated input-dependently.

Gated SSMs. Mamba (Gu & Dao, 2024) and Mamba-2 (Dao & Gu, 2024) make discretization parameters Δ_t, $\mathbf{B}_t$, $\mathbf{C}_t$ functions of $\mathbf{x}_t$, modulating write strength.

Error-correcting rules. DeltaNet (Yang et al., 2024) and Gated DeltaNet (Yang et al., 2025) compute prediction errors and write corrections—online gradient descent with surprise as prediction error magnitude.

Explicit surprise. Titans (Behrouz et al., 2025) gate write intensity by gradient of local prediction loss with momentum.

Table 1 summarizes this mapping. Bottom rows: post-encoding mechanisms essential to human memory and RL have no counterparts in current models.

3 Three Post-Encoding Gaps

3.1 Gap 1: Replay Across Memory Systems

During slow-wave sleep, the hippocampus replays episodes, driving gradual neocortical transfer—systems consolidation (McClelland et al., 1995; Rasch et al., 2007). Replay prioritizes high-reward-prediction-error experiences (Diekelmann & Born, 2010).

In RL, experience replay (Lin, 1992) and PER (Schaul et al., 2016) store transitions and re-sample them, with PER sampling proportional to TD error. This acts as functional hippocampus, stabilizing learning.

Current hybrid models process tokens in single passes. Attention (episodic) and recurrence (semantic) write independently with no subsequent transfer. No sleep-like phase re-processes and distills attention cache into recurrent state, contrasting with CLS and PER where fast-slow interplay drives generalization.

3.2 Gap 2: Importance-Sampling Correction

Human memory exhibits bias: emotionally charged events are disproportionately represented (McGaugh, 2004). Reconsolidation partially corrects this, recalibrating memory strength (Nader et al., 2000; Schiller et al., 2010).

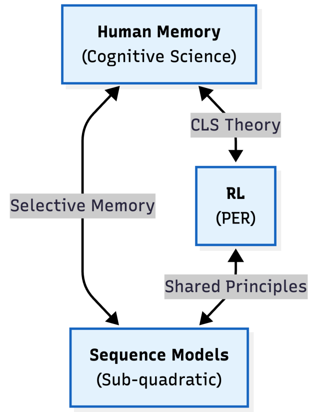

Figure 1: Triangular correspondence among human memory consolidation, prioritized experience replay (PER) in reinforcement learning, and modern sub-quadratic sequence models. All three domains converge on selective memory principles, yet sequence models lack the post-encoding processes—consolidation, importance-sampling correction, and reconsolidation—that are essential in both cognitive science and RL.

In PER, prioritized sampling changes the data distribution, introducing bias. The solution: importance-sampling weights $w_i = (N \cdot P(i))^{-\beta}$. Without correction, agents overfit to high-error transitions (Schaul et al., 2016).

When Titans or DeltaNet modulate write by surprise, high-surprise tokens receive disproportionate representation without compensating normalization—PER without IS correction. We hypothesize this contributes to weaknesses on tail-distribution phenomena like rare token prediction.

3.3 Gap 3: Reconsolidation Upon Retrieval

Nader et al. (2000) showed consolidated memories, when reactivated, return to labile states requiring restabilization. This allows updating given new retrieval-time information.

In PER, the "stale priority" problem arises: after parameter updates, stored TD errors become outdated. If not refreshed, replay distribution misaligns with the current model (Schaul et al., 2016).

In sequence models, once written to $\mathbf{S}_t$, information is never re-evaluated. Reading ($\mathbf{y}_t = \mathbf{S}_t\mathbf{q}_t$) doesn't modify $\mathbf{S}_t$. Information deemed "surprising" at encoding retains influence indefinitely, even if subsequent context renders it irrelevant.

4 Proposed Mechanisms

4.1 Cross-System Consolidation Phase

We propose periodic *consolidation phases* transferring information between systems. For hybrid model with attention $\mathcal{A}$ (episodic) and recurrent $\mathcal{R}$ (semantic), after processing segments, consolidation selects cache entries by importance and re-presents them:

$$\mathbf{S}^{\text{cons}} = \mathbf{S}_T + \eta_c \sum_{i \in \mathcal{C}} \alpha_i \, \Delta\mathbf{S}(\mathbf{k}_i, \mathbf{v}_i), \tag{1}$$

where $\mathcal{C}$ is selected entries, α_i importance weights, $\Delta\mathbf{S}$ state update, η_c consolidation rate. This mirrors hippocampal replay and PER but operates on internal states.

4.2 Importance-Sampling Normalization

To correct distributional bias, we propose normalization during reads. Let s_t denote surprise-based write intensity:

$$\tilde{\mathbf{y}}_t = \frac{\mathbf{S}_t \mathbf{q}_t}{\sum_{\tau=1}^{t} \gamma^{t-\tau} s_\tau + \epsilon}, \tag{2}$$

where $\gamma \in (0, 1)$ is decay, $\epsilon > 0$ prevents division by zero. The denominator estimates cumulative write intensity, analogous to PER IS weights.

4.3 Reconsolidation Gate

We propose *retrieval-triggered updates*. Let $\mathbf{q}_t$ be query, $\hat{\mathbf{v}}_t = \mathbf{S}_t \mathbf{q}_t$ retrieved value. Reconsolidation gate $g_t^{\text{recon}} \in [0, 1]$ determines update extent:

$$\mathbf{S}_t^{\text{recon}} = \mathbf{S}_t + g_t^{\text{recon}} \cdot \beta_r \left(\mathbf{v}_t^{\text{target}} - \hat{\mathbf{v}}_t \right) \mathbf{q}_t^\top, \tag{3}$$

where $\mathbf{v}_t^{\text{target}}$ is context-suggested value, β_r reconsolidation rate, and the gate activates when retrieved information is inconsistent with expectations (Sevenster et al., 2013).

5 Testable Predictions

5.1 Machine Learning Predictions

1. **Replay improves rare-event recall.** Consolidation phases should show disproportionate gains on tasks requiring recall of rarely-appearing information, such as entity-based QA from long documents.
2. **IS correction reduces tail perplexity.** Importance-sampling normalization should reduce perplexity on low-frequency tokens without harming high-frequency performance.
3. **Reconsolidation enables graceful adaptation.** Reconsolidation gates should enable graceful degradation under mid-sequence distributional shifts as the state self-corrects when retrieving outdated associations.

5.2 Cognitive Modeling Predictions

1. **Serial position effects.** Consolidation replay should produce stronger primacy effects as early items receive more replay passes (Gais et al., 2006).
2. **Testing effect.** Reconsolidation gates should reproduce the testing effect (Roediger & Karpicke, 2006) as retrieval triggers state updates reinforcing accurate associations.
3. **Spacing effect.** Consolidation should yield better retention with spaced presentations as consolidation integrates each presentation before the next arrives.

These predictions are falsifiable. If a mechanism improves engineering metrics but fails to reproduce the corresponding cognitive phenomenon, it suggests a useful heuristic rather than a genuine principle.

6 Discussion

Roadmap for empirical validation. While presenting a design space without experiments, we propose a concrete roadmap. **Phase 1:** Implement consolidation on small hybrid model ($\sim$350M parameters), test on MQAR benchmark measuring (i) early vs. late position accuracy, (ii) rare entity recall, (iii) training stability. **Phase 2:** Run on cognitive psychology tasks (serial position, testing effect, spacing effect), compare to human data. **Phase 3:** If successful, scale to 1B+ parameters, evaluate on MMLU, BBH, long-context tasks. This ensures both engineering validation and cognitive grounding.

Limitations. Consolidation increases computational cost. The human-model analogy is imperfect: biological processes operate over different timescales, involve neurochemical mechanisms without computational analogs, serve survival not next-token prediction. As a blue-sky paper without experiments, our contribution rests on design space novelty and specificity. Some mechanisms may already be implicit—deep Transformer layers may perform inter-system transfer via residuals.

Conclusion. Sequence model convergence on surprise-gated memory signals importance from a cognitive perspective. The next step: look beyond encoding toward post-encoding processes making human memory robust. By connecting human memory consolidation, prioritized experience replay, and sequence design, we identify a concrete design space with falsifiable predictions. We hope this encourages cross-pollination among communities studying memory in brains, RL agents, and neural architectures.

ACKNOWLEDGMENTS

SoHyung Kim would like to express profound gratitude to Jea Kwon for his invaluable mentorship and steadfast support throughout this independent research. Additionally, an AI assistant was utilized solely for English language polishing and LaTeX formatting, strictly in accordance with the ICLR 2026 LLM policy.

REFERENCES

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024.

Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. In *International Conference on Learning Representations*, 2025.

Steffen Gais, Brandon Lucas, and Jan Born. Sleep after learning aids memory recall. *Learning & Memory*, 13(3):259–262, 2006.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2024.

Kazuki Irie, Morris Yau, and Samuel J. Gershman. Blending complementary memory systems in hybrid quadratic-linear transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3–4):293–321, 1992.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

James L. McGaugh. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27:1–28, 2004.

Karim Nader, Glenn E. Schafe, and Joseph E. Le Doux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726, 2000.

Björn Rasch, Jan Börn, and Steffen Gais. Maintaining memories by reactivation. *Current Opinion in Neurobiology*, 17(6):698–703, 2007.

Henry L. Roediger and Jeffrey D. Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3):181–210, 2006.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2016.

Daniela Schiller, Marie-H. Monfils, Candace M. Raio, David C. Johnson, Joseph E. LeDoux, and Elizabeth A. Phelps. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, 463(7277):49–53, 2010.

Dieuwke Sevenster, Tom Beckers, and Merel Kindt. Prediction error governs pharmacologically induced amnesia for learned fear. *Science*, 339(6121):830–833, 2013.

Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005.

Endel Tulving. *Episodic and semantic memory*. Academic Press, New York, 1972.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.