Sleeping LLM: Two-Phase Memory Consolidation for Lifelong Learning from 3B to 70B Parameters

Vladimir Baranov
vlad@chatsetter.ai

Abstract

Large language models lose conversational context when sessions end—the context window is volatile and retrieval-augmented generation externalizes memory rather than internalizing it. We present a biologically-inspired system that gives LLMs durable, internalized memory through two complementary mechanisms: MEMIT for instant factual injection during wake, and a two-phase sleep cycle (SWS + REM) for consolidation into LoRA-adapted weights. Across three model scales (3B, 8B, 70B), we show that MEMIT provides near-zero-cost memory: 20 facts produce less than 0.03 perplexity change at every scale. During sleep, Slow-Wave Sleep (SWS) consolidates individual facts via per-fact LoRA training, while REM integration trains on synthetic multi-fact conversations to repair distributional damage. REM reduces SWS-induced perplexity increase by 88% on the 3B model (from +1.6% to +0.2%), with all three scales completing the full wake-nap-sleep lifecycle end-to-end. At 8B, per-fact staged consolidation achieves 80% single-cycle and 95% two-cycle consolidation rates—a 2.2× improvement over bulk training. We report three honest negative results: (1) zero facts pass the alignment-tax gate at any scale via raw completion testing, requiring chat-template validation instead; (2) REM slightly increases perplexity at 8B, helping most where SWS hurts most; and (3) PPL scaling is non-monotonic across model sizes (3B: +1.6%, 8B: +14.1%, 70B: +0.3%), suggesting that model-specific tuning rather than universal protocols will be required for production deployment.

1 Introduction

Language models accumulate knowledge during pretraining but cannot learn from ongoing interaction. The context window provides temporary working memory, but its contents vanish when the session ends. Retrieval-augmented generation (RAG) (Lewis et al., 2020) partially addresses this by externalizing memory to a database, but the knowledge remains outside the model's parameters—it must be retrieved at inference time and competes with the prompt for context space.

Biological memory systems face the same tension between fast learning and stable storage. The hippocampus rapidly encodes new experiences, while the neocortex gradually integrates them into long-term knowledge through sleep-dependent consolidation (McClelland et al., 1995). Complementary Learning Systems (CLS) theory describes a two-stage process: Slow-Wave Sleep (SWS) replays individual memory traces for initial consolidation, while Rapid Eye Movement (REM) sleep integrates these traces into coherent schemas (Diekelmann and Born, 2010; Rasch and Born, 2013). The key insight is that fast and slow learning are not alternatives—they are complementary phases of a single system.

We operationalize this biological architecture in a working system that runs locally on consumer and datacenter hardware. MEMIT (Meng et al., 2023) serves as the hippocampal fast-learning system, injecting facts directly into MLP weights during wake with near-zero perplexity cost. A two-phase sleep cycle then consolidates these facts: SWS trains per-fact LoRA adapters (Hu et al., 2022)

on individual memories, while REM trains on synthetic multi-fact conversations that interleave consolidated knowledge. A perplexity-based validation gate governs REM acceptance, rolling back to the post-SWS state if integration degrades the model.

We validate this architecture across three scales: 3B (4-bit on MacBook Air M3 via MLX, and bfloat16 on H100), 8B (bfloat16 on dual H100), and 70B (NF4 on dual H100). We make five contributions:

1. **MEMIT is near-free-lunch memory.** Across all three scales, 20 per-fact MEMIT edits produce less than 0.03 perplexity change (Table 3). Covariance regularization with the Woodbury formula makes MEMIT effectively cost-free at the scales we test.

2. **Two-phase sleep is validated, with scale-dependent benefits.** REM reduces SWS-induced perplexity damage by 88% on the 3B model, cutting the PPL increase from +1.6% to +0.2% (Table 5). At 8B and 70B, REM effects are smaller, consistent with the pattern that REM helps most where SWS hurts most.

3. **The full lifecycle works at every scale.** All three model sizes complete the wake-nap-sleep pipeline end-to-end (Table 6), with sleep validation approved at 8B and 70B. At 8B, per-fact staged consolidation achieves 95% consolidation over two cycles.

4. **Perplexity scaling is non-monotonic.** SWS increases perplexity by 1.6% at 3B, 14.1% at 8B, and only 0.3% at 70B—the intermediate scale suffers most, suggesting model-specific rather than universal consolidation protocols.

5. **Honest limitations are characterized.** The alignment tax prevents raw-completion consolidation at all scales tested. Chat-template validation succeeds where raw completion fails, but the gap between injection pathway (raw MLP edits) and consolidation pathway (chat-format LoRA) remains unresolved.

This paper extends our prior work on per-fact staged consolidation (Baranov, 2026b) and dual-system memory (Baranov, 2026a) with three-scale MEMIT characterization, the REM integration phase, and full lifecycle validation from 3B to 70B.

2   Related Work

2.1   Continual Learning for LLMs

Catastrophic forgetting—where fine-tuning on new data degrades performance on old tasks—is a central challenge in deploying LLMs for lifelong learning (Luo et al., 2023; Li et al., 2024). Elastic Weight Consolidation (Kirkpatrick et al., 2017) penalizes changes to parameters important for prior tasks, while progressive methods like PackNet (Mallya and Lazebnik, 2018) allocate dedicated subnetworks. These approaches require task boundaries and replay buffers that grow with the number of tasks. Our system sidesteps explicit task management by using MEMIT edits as a durable safety net: facts that fail LoRA consolidation retain their MEMIT edit rather than being lost.

2.2   Knowledge Editing

ROME (Meng et al., 2022) demonstrated that individual facts can be edited by modifying specific MLP layers. MEMIT (Meng et al., 2023) extended this to batched editing across multiple layers.

Subsequent work has characterized the reliability and failure modes of knowledge editing (Yao et al., 2023; Mitchell et al., 2022). We use per-fact MEMIT injection (one edit per fact) with covariance regularization to preserve the model's output distribution, achieving near-zero perplexity cost at 20 facts across three scales.

2.3 Sleep-Inspired Learning

Complementary Learning Systems theory (McClelland et al., 1995; Kumaran et al., 2016) provides the biological foundation: the hippocampus handles fast encoding while the neocortex provides slow, interleaved learning. Sleep plays a critical role in this transfer, with SWS replaying hippocampal traces and REM integrating them into existing knowledge structures (Diekelmann and Born, 2010; Rasch and Born, 2013; Walker and Stickgold, 2004). Computationally, Robins (1995) showed that pseudorehearsal during interleaved training prevents catastrophic forgetting. Shin et al. (2017) implemented this as deep generative replay, and van de Ven et al. (2020) explicitly framed replay as brain-inspired continual learning. Tononi and Cirelli (2014) proposed the synaptic homeostasis hypothesis, where sleep globally downscales synaptic weights to restore capacity. Our REM phase operationalizes a related idea: integration training on synthetic multi-fact conversations counteracts the distributional shift introduced by SWS.

2.4 Parameter-Efficient Fine-Tuning

LoRA (Hu et al., 2022) enables efficient adaptation by training low-rank additive updates, and QLoRA (Dettmers et al., 2023) extends this to quantized models. For continual learning, O-LoRA (Wang et al., 2023) uses orthogonal subspaces to reduce interference between tasks, and InfLoRA (Liang and Li, 2024) provides interference-free adaptation. Our system uses standard LoRA (rank 16, alpha 32) but relies on the MEMIT safety net and per-fact gating rather than subspace management to handle forgetting.

2.5 Experience Replay and Synthetic Data

Experience replay (Robins, 1995; Shin et al., 2017) mitigates forgetting by mixing old data into new training. Our replay buffer maintains prioritized conversation history for SWS training. The REM phase extends this idea: rather than replaying raw conversations, it generates synthetic multi-fact dialogues that interleave multiple consolidated facts, forcing the model to integrate rather than memorize in isolation.

3 System Architecture

3.1 Overview: Sleep-Wake Lifecycle

The system operates as a state machine with three phases—wake, nap, and sleep—mapped to biological CLS components (Table 1).

During **wake**, the user converses with the model. Facts extracted from conversation are injected via MEMIT into MLP weights, providing instant recall through raw completion. **Naps** are lightweight consolidation events: LoRA trains on MEMIT-held facts without modifying MEMIT edits, reinforcing traces for future sleep. **Full sleep** executes the two-phase pipeline: SWS consolidates individual facts with per-fact LoRA gating, then REM integrates consolidated knowledge through synthetic multi-fact conversations. Sleep pressure accumulates with each MEMIT edit and triggers sleep when a threshold is crossed.

Table 1: Mapping from Complementary Learning Systems theory to system components.

Biological Component	System Implementation
Hippocampal fast encoding	MEMIT weight edits (per-fact, instant)
Neocortical slow learning	LoRA adapters (merged into base weights)
SWS trace replay	Per-fact LoRA training on individual memories
REM schema integration	LoRA training on synthetic multi-fact conversations
Sleep pressure	Weighted edit count + time since last sleep
Hippocampal residual	MEMIT delta at reduced scale (0.1) post-consolidation

3.2 Wake Phase: MEMIT Injection

Each fact triple (s, r, o) is injected as an independent MEMIT edit targeting MLP down-projection layers. The weight update for target layer ℓ minimizes reconstruction error subject to covariance regularization:

$$\mathbf{W}'_\ell = \mathbf{W}_\ell + \mathbf{R}_\ell \mathbf{K}_\ell^T \left(\mathbf{K}_\ell \mathbf{K}_\ell^T + \lambda \hat{\mathbf{C}}_\ell \right)^{-1} \tag{1}$$

where $\mathbf{K}_\ell$ are the key vectors at layer ℓ for the edited facts, $\mathbf{R}_\ell$ is the distributed residual (the gap between current and target outputs), and $\hat{\mathbf{C}}_\ell = \frac{1}{M} \sum_{j=1}^{M} \mathbf{k}_j \mathbf{k}_j^T$ is the empirical key covariance estimated from $M = 200$ reference samples. The regularization term $\lambda \hat{\mathbf{C}}_\ell$ penalizes updates that distort the model's output distribution on general text, preserving perplexity. We set $\lambda = 0.1$ across all experiments.

The Woodbury matrix identity (Woodbury, 1950) converts the $d \times d$ inversion in Equation 1 to an $N \times N$ inversion (where N is the number of edited facts and d is the hidden dimension), making the computation tractable even at 70B scale.

Cross-edit null-space constraints prevent sequential edits from overwriting each other: each new edit's key vectors are projected to avoid the subspace spanned by previous edits' keys. The residual is distributed across target layers (dividing by the number of remaining layers at each stage), ensuring no single layer absorbs disproportionate perturbation.

Delta persistence. Each edit's weight delta $\Delta \mathbf{W}_\ell^{(i)}$ is serialized to disk immediately after injection. On process restart, all active deltas are reloaded:

$$\mathbf{W}_\ell \leftarrow \mathbf{W}_\ell + \sum_{i \in \mathcal{A}} s_i \cdot \Delta \mathbf{W}_\ell^{(i)} \qquad \forall \ell \in \text{target layers} \tag{2}$$

where $\mathcal{A}$ is the set of active edits and $s_i \in [0, 1]$ is each edit's current scale. This gives MEMIT edits the same restart persistence as LoRA checkpoints.

3.3 SWS Phase: Per-Fact LoRA Consolidation

The SWS phase consolidates MEMIT-held facts into LoRA weights through per-fact training and individual gating. The procedure follows Algorithm 1.

Curation. Conversation history is curated into Q&A training pairs. MEMIT facts from the edit ledger are converted to explicit question-answer pairs and added to the training set. A replay buffer maintains prioritized conversation history from previous sessions, mixing old and new data.

Algorithm 1 Per-Fact Staged Consolidation (SWS Phase)

Require: Active MEMIT edits $\mathcal{E} = \{e_1, \ldots, e_n\}$, each with stage g_i and scale s_i
  1: $\mathbf{S} \leftarrow \textsc{SnapshotWeights}(\text{target layers})$ ▷ Byte-exact copy
  2: $\mathbf{s}_{\text{pre}} \leftarrow \{(e_i, s_i) \mid e_i \in \mathcal{E}\}$ ▷ Record pre-sleep scales
  3: $\textsc{TrainLoRA}(\mathcal{E})$ ▷ LoRA training on MEMIT facts
  4: $\textsc{MergeLoRA}()$ ▷ Merge adapter into base weights
  5: **for** $e_i \in \mathcal{E}$ **do**   ▷ Isolate pure LoRA signal
  6:      $\textsc{ScaleEdit}(e_i, 0.0)$
  7: **end for**
  8: $v_{\text{bench}} \leftarrow \textsc{BenchmarkValidation}()$
  9: **if** $v_{\text{bench}} < \tau$ **then**   ▷ Benchmark failed
 10:      $\textsc{RestoreWeights}(\mathbf{S})$ ▷ Byte-exact rollback
 11:      Restore all scales from $\mathbf{s}_{\text{pre}}$
 12:      **return** $\textsc{Rejected}$
 13: **end if**
 14: **for** $e_i \in \mathcal{E}$ **do**   ▷ Per-fact evaluation
 15:      $r_i \leftarrow \textsc{TestRecall}(e_i)$ ▷ Pure LoRA recall
 16:      **if** $r_i = \text{True}$ **then**
 17:          **if** $g_i = 0$ **then**
 18:              $\textsc{ScaleEdit}(e_i, \alpha_{\text{residual}}); g_i \leftarrow 1$
 19:          **else if** $g_i = 1$ **then**
 20:              $g_i \leftarrow 2$ ▷ Consolidated
 21:          **end if**
 22:      **else**
 23:          $\textsc{ScaleEdit}(e_i, s_{\text{pre},i})$ ▷ Restore original scale
 24:      **end if**
 25: **end for**
 26: **return** $\textsc{Approved}$, stage updates

Per-fact training. A single-epoch LoRA adapter (rank 16, alpha 32, targeting 8 transformer layers) is trained on the combined dataset, then merged into base weights. The learning rate is 1×10^{-4} with iterations scaled to the number of training examples.

Per-fact gating. After LoRA merge, each MEMIT edit is scaled to 0.0 (removing its contribution) and the fact is tested for recall through the pure LoRA pathway. Facts demonstrating LoRA recall advance one consolidation stage; facts that fail retain their full MEMIT edit. This staged advancement (Algorithm 1) allows each fact to consolidate at its own rate.

Three consolidation stages track each fact's progress:
- **Stage 0 (active):** MEMIT at scale 1.0, LoRA not yet proven.
- **Stage 1 (consolidating):** LoRA demonstrated recall in one cycle. MEMIT scaled to 0.1.
- **Stage 2 (consolidated):** LoRA demonstrated recall across two cycles.

3.4 REM Phase: Integration Sleep

The REM phase addresses a limitation of per-fact SWS training: each fact is consolidated in isolation, producing LoRA updates that may collectively shift the model's output distribution.

REM counteracts this by training on *multi-fact conversations*—synthetic dialogues that interleave multiple consolidated facts in natural contexts.

Integration data generation. The dreamer module generates synthetic conversations that reference multiple consolidated facts. For example, if facts include "Viktor lives in Portland" and "Viktor works as a librarian," the dreamer produces a conversation where both facts appear naturally: "Tell me about Viktor—where does he live and what does he do?" This forces the model to integrate facts into coherent representations rather than storing each in an isolated weight direction.

REM training. A separate LoRA training pass runs on the integration data with the same hyperparameters as SWS. The LoRA adapter is merged into the post-SWS weights, producing the final post-sleep model.

PPL validation gate. REM is accepted only if it passes a dual validation gate:

$$\text{REM accepted} \iff \frac{\text{PPL}_{\text{post-REM}} - \text{PPL}_{\text{post-SWS}}}{\text{PPL}_{\text{post-SWS}}} \leq \tau_{\text{ppl}} \ \wedge \ \frac{\sum_{f \in \mathcal{S}} \mathbb{1}[\text{recall}(f)]}{|\mathcal{S}|} \geq \tau_{\text{recall}} \quad (3)$$

where $\tau_{\text{ppl}} = 0.10$ (maximum 10% PPL increase over post-SWS baseline), $\mathcal{S}$ is a sample of up to 5 consolidated facts, and $\tau_{\text{recall}} = 0.5$ (at least half must still be recalled). If REM fails either condition, the model is rolled back to the post-SWS state via weight snapshot restoration.

3.5 Health Monitoring and Sleep Triggers

Sleep pressure follows a non-linear curve that allows more facts to accumulate before triggering:

$$p_{\text{edit}} = \min\left(1.0, \left(\frac{n_{\text{edits}}}{n_{\text{max}}}\right)^{1.5}\right) \quad (4)$$

where n_{edits} is the current active edit count and n_{max} is the configured threshold. The exponent 1.5 provides sublinear pressure growth, giving LoRA training larger batches. Consolidation proportionally reduces pressure: when k facts advance to stage ≥ 1, the effective edit count decreases by k.

4 Experimental Setup

4.1 Models and Hardware

We evaluate at three scales (Table 2).

The 3B model (`Llama-3.2-3B-Instruct`) runs in two configurations: 4-bit quantized on Apple Silicon via MLX (Hannun et al., 2023) for local development, and bfloat16 on H100 for controlled experiments. The 8B model (`Llama-3.1-8B-Instruct`) runs unquantized on dual H100 with `device_map="auto"` distributing layers across GPUs. The 70B model (`Llama-3.1-70B-Instruct`) uses bitsandbytes NF4 quantization (Dettmers et al., 2023) on dual H100, with MEMIT target layers reduced from 16 to 8 (layers 36–43) to fit within 160 GB VRAM during the v^* optimization step.

Table 2: Model and hardware configurations. MEMIT layers and LoRA target layers vary by model size; all other hyperparameters are shared.

Model	Hardware	Precision	MEMIT Layers	LoRA r	LoRA α
3B (MLX)	MacBook Air M3, 8 GB	4-bit	8–15	16	32
3B (torch)	1× H100 80 GB	BF16	8–15	16	32
8B	2× H100 80 GB	BF16	12–19	16	32
70B	2× H100 80 GB	NF4	36–43	16	32

4.2 MEMIT and LoRA Configuration

MEMIT uses $\lambda_{\text{reg}} = 0.1$, covariance estimated from 200 reference samples, v^* optimization for 30 steps at learning rate 0.5, and cross-edit null-space projection. LoRA targets 8 transformer layers per model with rank 16, alpha 32, learning rate 1×10^{-4}, and iterations scaled to training examples (1 epoch for naps, 1–3 epochs for full sleep). Multi-GPU LoRA training required three specific fixes for dual-H100 configurations: disabling fused multi-tensor optimizer operations, using gradient enablement instead of quantization-aware training preparation, and skipping model reload after merge.

4.3 Evaluation Protocol

We evaluate along four axes:
- **Raw completion recall:** given a prompt like "Idris Larsson works as", does the model complete with the correct target? Tests the MEMIT pathway.
- **Chat-template recall:** given a question in chat format ("What does Idris Larsson do for work?"), does the model answer correctly? Tests the LoRA/chat pathway.
- **Perplexity (PPL):** cross-entropy loss on reference texts, measuring model health.
- **Sleep validation:** 5-question benchmark evaluated before and after sleep, with a minimum score ratio gate.

4.4 Fact Generation

All experiments use 20 synthetic person-city-occupation facts (e.g., "Viktor Sørensen lives in Portland and works as a librarian"). Each fact generates one MEMIT edit targeting the subject-relation-object triple and one Q&A training pair for LoRA. Facts are injected in two batches of 10, with perplexity measured after each batch.

5 Results

5.1 MEMIT: Near-Zero PPL Cost Across Scales

Table 3 shows perplexity trajectories during MEMIT injection at all three scales.

The maximum absolute PPL change across all conditions is 0.027 (8B at 20 facts)—less than 0.5% of the baseline. At 3B and 70B, the change is within 0.01. In several conditions, PPL actually *decreases* after injection, likely due to the covariance regularization slightly tightening the output distribution. These results confirm that covariance-regularized MEMIT with the Woodbury formula is a near-free-lunch operation: facts can be injected during wake with negligible impact on general model quality.

Table 3: Perplexity cost of MEMIT injection across scales. 20 per-fact edits produce less than 0.03 absolute PPL change at every model size, confirming that covariance-regularized MEMIT is effectively cost-free.

Model	Baseline	+10 facts	+20 facts	ΔPPL
3B	5.711	5.709	5.703	-0.008
8B	5.752	5.766	5.725	-0.027
70B	5.096	5.105	5.099	$+0.003$

Table 4 shows MEMIT recall capacity at each scale, tested independently of the sleep pipeline.

Table 4: MEMIT recall capacity (raw completion) at increasing fact counts. Peak recall is 0.80–0.82 across all scales, with 8B sustaining 0.80+ recall up to 50 facts. 70B encountered OOM at 50 facts due to v^* backward pass memory.

Facts	3B	8B	70B
5	0.80	0.80	—
10	0.80	0.70	0.80
20	0.65	0.65	0.80
30	0.70	0.77	0.77
40	—	0.82	0.78
50	—	0.82	OOM

All three scales achieve similar peak recall (≈ 0.80), but 8B shows the best capacity curve, sustaining 0.80+ from 35 to 50 facts. The 3B model peaks at 10 facts and degrades at higher counts. The 70B model holds 0.77–0.80 stably up to 40 facts before running out of memory during the v^* backward pass at 50 facts (16 dequantized layers at bfloat16 $\approx 120\,\mathrm{GB}$ plus autograd overhead exceeds the 160 GB VRAM budget).

5.2 Two-Phase Sleep: SWS vs. SWS+REM

Table 5 presents the central result: a controlled comparison of SWS-only versus SWS+REM sleep at all three scales, each processing 20 MEMIT-injected facts.

Table 5: Two-phase sleep comparison (20 facts per condition). REM reduces SWS-induced perplexity damage by 88% at 3B, with negligible effect at 70B and slight increase at 8B. Recall is maintained or improved in all conditions. ΔPPL is relative to baseline.

Model	SWS-only PPL	SWS-only Recall	SWS+REM PPL	SWS+REM Recall	REM ΔPPL
3B	5.804 (+1.6%)	0.90	5.722 (+0.2%)	0.90	-88% damage
8B	6.564 (+14.1%)	0.95	6.625 (+15.2%)	1.00	$+7.5\%$ damage
70B	5.113 (+0.3%)	0.90	5.120 (+0.5%)	0.90	negligible

The results reveal a clear pattern: **REM helps most where SWS hurts most.**

3B: Strong REM benefit. SWS alone increases PPL by 1.6% (5.711 $\rightarrow$ 5.804). Adding REM reduces the net increase to 0.2% (5.711 $\rightarrow$ 5.722)—an 88% reduction in PPL damage. The REM

phase generated 6 integration conversations and was approved by the validation gate (internal PPL dropped from 5.75 to 5.53 during REM training). Recall is unchanged at 0.90 in both conditions.

8B: Mixed result. SWS causes the largest PPL increase at 14.1% (5.752 → 6.564). Adding REM increases this slightly to 15.2% (5.752 → 6.625), an additional 0.061 PPL. However, REM improves recall from 0.95 to 1.00—the integration training appears to strengthen the recall pathway at the cost of slightly more distributional shift. The REM phase generated 7 integration conversations and was approved (internal PPL: 3.71 → 3.64).

70B: Negligible effect. SWS barely perturbs the 70B model (+0.3%, ΔPPL = 0.017). REM adds only 0.007 more PPL, well within noise. The 70B model's larger parameter space absorbs consolidation with minimal distributional shift, leaving little room for REM to help. The REM phase generated 10 integration conversations.

The non-monotonic PPL scaling (3B: +1.6%, 8B: +14.1%, 70B: +0.3%) is the most surprising finding and is discussed in Section 6.4.

5.3 Full Lifecycle Validation

Table 6 shows end-to-end lifecycle results at all three scales.

Table 6: Full lifecycle validation across scales. All three models complete the wake-nap-sleep pipeline. MEMIT recall is tested via raw completion; sleep validation uses a 5-question benchmark with chat-template queries.

Model	MEMIT	Capacity	Nap	Sleep	Status
3B	3/3	0.80 @ 10	PASS	2/5 recall	Full lifecycle
8B	2/3	0.82 @ 40	2/4 consolidated	APPROVED 5/5	Full lifecycle
70B	2/3	0.80 @ 40	PASS	APPROVED 5/5	Full lifecycle

All three model sizes complete the full pipeline: wake (MEMIT injection from conversation), nap (LoRA reinforcement), and full sleep (SWS consolidation + REM integration + validation). At 8B, the nap successfully consolidated 2 of 4 tested facts. At 8B and 70B, full sleep was approved with 5/5 validation scores. The 3B sleep cycle achieved 2/5 post-sleep recall—lower than 8B/70B, but the sleep pipeline completed without errors.

Multi-GPU execution on dual H100 required fixes to 5 locations in the MEMIT pipeline (moving tensors to per-layer devices under `device_map="auto"`) and 3 fixes to LoRA training (disabling fused optimizer operations, replacing quantization-aware preparation with gradient enablement, and using in-memory merged models instead of fuse-and-reload).

5.4 Ablation Studies

Dual-system ablation. Table 7 isolates the contribution of each subsystem using 8B data from Baranov (2026b).

The two subsystems are complementary: MEMIT edits the raw completion pathway (MLP down-projections), while LoRA trains the chat-template pathway (attention and MLP layers). Neither alone provides both raw and chat recall. The MEMIT residual at 0.1 has zero measurable effect on either recall metric, consistent with the representational separation between pathways.

Table 7: Dual-system ablation (8B model, 10 facts from v3 experiments). MEMIT provides instant raw recall but no chat access. LoRA provides chat access but cannot inject facts instantly. The combined system provides both pathways with the same chat recall.

Condition	Raw Recall	Chat Recall	PPL
MEMIT only (pre-sleep)	8/10	0/10	6.53
Pure LoRA (MEMIT at 0.0)	6/10	8/10	11.52
MEMIT + LoRA (residual 0.1)	6/10	8/10	11.52

Covariance regularization (λ sweep). At $\lambda = 0$, MEMIT injection of 20 facts at 8B increases PPL by $>2\times$. At $\lambda = 0.1$, the same injection produces ΔPPL < 0.03. Higher values ($\lambda \geq 0.5$) reduce recall by over-constraining the update. The value $\lambda = 0.1$ provides the best trade-off across all three scales.

Retention under interference. A controlled residual sweep at 8B (Baranov, 2026b) tested whether MEMIT residual traces protect consolidated facts from new-fact interference. At residual scales 0.0–0.3, zero effect was observed. At 0.5, the residual actively degraded performance. This falsifies the "structural echo" hypothesis: MEMIT and LoRA occupy sufficiently separate parameter subspaces that cross-pathway reinforcement does not occur.

5.5 Per-Fact Staged Consolidation

Table 8 summarizes the staged consolidation trajectory at 8B with 20 facts, from Baranov (2026b).

Table 8: Per-fact staged consolidation trajectory (8B model, 20 facts). 95% of facts reach stage ≥ 1 in a single cycle, with 100% chat recall maintained. Per-fact training achieves $2.2\times$ the consolidation rate of bulk training (80% vs. 37% at 10 facts).

Step	PPL	Raw	Chat	St. 0	St. 1	St. 2
Baseline	6.49	—	—	—	—	—
Post-inject	6.53	5/20	0/20	20	0	0
Post-sleep-1	11.52	8/20	20/20	1	19	0
Post-sleep-2	13.33	9/20	20/20	0	1	19

In a single sleep cycle, 19/20 facts advance from stage 0 to stage 1 (95%), with 100% chat recall. Over two cycles, 19/20 reach stage 2 (consolidated). The remaining fact advances to stage 1 in cycle 2. Chat recall is maintained at 100% throughout.

The perplexity cost is substantial: PPL nearly doubles after the first sleep cycle ($6.53 \rightarrow 11.52$) and continues rising ($11.52 \rightarrow 13.33$). This is the cost of SWS-only consolidation without REM—each fact's LoRA training bends the output distribution toward the training data. The two-phase results in Section 5.2 show that REM can partially mitigate this cost.

Nap safety. Table 9 compares the prior destructive nap design (Baranov, 2026a) with the current non-destructive design.

Table 9: Nap safety comparison (8B model, 10 facts). The prior design reverted MEMIT edits after LoRA training, causing a 40% recall drop. The current design preserves all MEMIT state.

Metric	Pre-Nap	v2 Post-Nap	v3 Post-Nap
MEMIT raw recall	8/10	0/10 (reverted)	8/10
Net recall	8/10	4/10 (−40%)	8/10 (0%)

6 Discussion

6.1 MEMIT as Long-Term Memory

A surprising finding is that MEMIT—originally designed as the hippocampal "fast but temporary" component—functions as durable long-term memory. With covariance regularization, 20 MEMIT edits produce less than 0.03 PPL change across three scales. Edits survive restarts via delta persistence, accumulate without catastrophic interference via null-space constraints, and maintain recall stably up to 40–50 facts.

This challenges the original CLS framing where hippocampal memories are inherently fragile and require neocortical transfer for durability. In our system, the primary reason to consolidate into LoRA is not MEMIT fragility but pathway access: MEMIT edits are accessible through raw completion but not through the chat template. Users interact via chat, so consolidation into the chat-accessible LoRA pathway is functionally necessary even though MEMIT could hold the facts indefinitely.

6.2 Where Two-Phase Sleep Helps

The REM benefit is inversely correlated with model size's ability to absorb SWS perturbation. At 3B, SWS causes meaningful distributional damage (+1.6% PPL) because the model's smaller parameter space concentrates LoRA updates. REM's integration training—which exposes the model to multi-fact conversations rather than isolated Q&A pairs—counteracts this concentration, reducing the PPL increase by 88%.

At 8B, SWS causes the largest PPL increase (+14.1%), but REM slightly worsens it (+7.5% additional). We hypothesize that the 8B model occupies a regime where the parameter space is large enough that SWS updates interact in complex ways, and REM's additional training adds rather than counteracts the perturbation. The recall improvement ($0.95 \rightarrow 1.00$) suggests REM is beneficial for knowledge retention even when PPL increases.

At 70B, the model's massive parameter space absorbs both SWS and REM with negligible impact (ΔPPL < 0.025 total), providing no leverage for REM to help.

6.3 The Alignment Tax

Across all three scales in the lifecycle experiments, zero facts pass the raw-completion consolidation gate (the per-fact test with MEMIT at 0.0 after LoRA training). LoRA learns facts through the chat-template pathway—as confirmed by 5/5 chat validation scores at 8B and 70B—but does not generalize to raw completions. This alignment tax, first characterized at 70B in Baranov (2026a), appears more pervasive than initially understood.

The staged consolidation results at 8B (80%/95% in Table 8) used a different experimental protocol from Baranov (2026b) that tested consolidation through the chat pathway. The lifecycle experiments test through raw completion, revealing the gap. The practical implication is that the

consolidation gate should evaluate through the same pathway (chat template) that users access, rather than the raw completion pathway that MEMIT edits target.

6.4 Perplexity as Health Signal

The non-monotonic PPL scaling across model sizes is the most unexpected result:
- **3B:** +1.6% (SWS-only). Small model, concentrated updates, moderate damage.
- **8B:** +14.1% (SWS-only). Intermediate scale, largest damage.
- **70B:** +0.3% (SWS-only). Large model, updates absorbed with minimal impact.

We conjecture that the 8B model occupies an unfortunate middle ground: large enough that LoRA updates target a wide set of attention heads and MLP layers (spreading perturbation broadly), but small enough that each individual perturbation is proportionally significant. The 3B model has fewer target layers and lower effective rank, limiting the spread. The 70B model has so many parameters that the same LoRA update (rank 16) represents a proportionally smaller perturbation.

This has practical implications: perplexity-based sleep gates should use model-specific thresholds rather than universal values. The 10% PPL threshold in our REM gate ($\tau_{\text{ppl}} = 0.10$) is appropriate for 3B and 70B but would rarely trigger at 8B, where SWS alone exceeds it.

6.5 Biological Analogy: Where It Holds and Breaks

The CLS mapping (Table 1) holds well for the two-phase sleep cycle. SWS-like per-fact replay consolidates individual memories, while REM-like integration training weaves them into multi-fact schemas. The sleep pressure mechanism provides a plausible analog to homeostatic sleep drive. The staged advancement mirrors the biological observation that memories consolidate at different rates (Rasch and Born, 2013).

The analogy breaks in two places. First, the residual trace hypothesis—that partially erased hippocampal traces aid future retrieval—is falsified in transformer weight space. MEMIT and LoRA target sufficiently non-overlapping parameter subspaces that cross-pathway reinforcement does not occur. Second, biological REM sleep involves global synaptic downscaling (Tononi and Cirelli, 2014); our REM phase adds training rather than removing it, which is mechanistically opposite despite achieving a similar functional outcome (perplexity restoration).

7 Limitations

Single-run experiments. All results are from single runs without error bars. The 88% REM benefit at 3B could vary with different random seeds, fact orderings, or reference text choices.

Alignment tax unresolved. Zero facts pass raw-completion consolidation at any scale. Chat-template validation works, but the gap between MEMIT's raw pathway and LoRA's chat pathway remains an open problem.

8B PPL anomaly. The 14.1% PPL increase at 8B is substantially worse than 3B or 70B. We lack a definitive explanation for this non-monotonic behavior.

Synthetic facts only. All experiments use synthetic person-occupation-location triples. Real conversational memories (opinions, preferences, temporal events) may consolidate differently.

No long-term study. We test at most 2 sleep cycles with 20 facts. Behavior at 100+ facts over 50+ cycles is unknown.

No RAG comparison. We do not compare against retrieval-augmented baselines, which would provide context on whether weight-based memory offers advantages over external retrieval.

Blocking sleep. The model goes offline during full sleep cycles (40s at 3B, 100s at 8B, 890s at 70B for SWS+REM). Naps are faster (<60s) but still blocking.

8 Future Work

Three directions are most pressing. First, the alignment tax should be addressed directly—either by training LoRA on raw-completion data in addition to chat-template data, or by developing consolidation gates that evaluate through the chat pathway. Second, the non-monotonic PPL scaling warrants systematic study across more model sizes and LoRA configurations to identify the regime boundaries. Third, multi-session longevity testing (50+ cycles, 500+ facts) is needed to understand whether perplexity degradation is bounded or unbounded under repeated consolidation.

9 Conclusion

We presented a two-phase sleep architecture for lifelong learning in language models, validated from 3B to 70B parameters. MEMIT provides near-zero-cost fast memory (ΔPPL < 0.03 at 20 facts across three scales). SWS consolidates individual facts via per-fact LoRA training with 95% two-cycle consolidation at 8B. REM integration reduces SWS-induced perplexity damage by 88% on the 3B model. All three scales complete the full wake-nap-sleep lifecycle end-to-end.

The system validates the Complementary Learning Systems framework in transformer architectures, but with important caveats: the alignment tax prevents raw-completion consolidation, MEMIT functions as durable rather than temporary memory, and perplexity scaling is non-monotonic across model sizes. These findings suggest that production deployment will require model-specific tuning of sleep parameters rather than universal protocols.

References

V. Baranov. Dual-system memory consolidation for lifelong learning in language models: Combining direct weight editing with sleep-wake training. *arXiv preprint*, 2026a. v2 of this work.

V. Baranov. Per-fact staged consolidation for lifelong learning in language models: From bulk training to granular memory management. *arXiv preprint*, 2026b. v3 of this work.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

A. Hannun, J. Digani, A. Katharopoulos, and R. Collobert. MLX: Efficient and flexible machine learning on Apple silicon. Apple Machine Learning Research. https://github.com/ml-explore/mlx, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Blundell, D. Wierstra, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

H. Li, L. Ding, M. Fang, and D. Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Y.-S. Liang and W.-J. Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.

Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

A. Mallya and S. Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

J. L. McClelland, B. L. McNaughton, and R. 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.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

K. Meng, A. S. Sharma, A. J. Andonian, Y. Belinkov, and D. Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2022.

B. Rasch and J. Born. About sleep's role in memory. *Physiological Reviews*, 93(2):681–766, 2013.

A. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

G. Tononi and C. Cirelli. Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34, 2014.

G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069, 2020.

M. P. Walker and R. Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44 (1):121–133, 2004.

X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang. O-LoRA: Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023.

M. A. Woodbury. Inverting modified matrices. *Memorandum Report*, 42:336, 1950.

Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.