
MORPHEUS: Multi-scale Offline Repulsive Sleep for Selective Forgetting in Latent Agent Memory

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We present MORPHEUS, a wake-sleep contrastive memory framework for LLM
2 agents that reduces stale retrieval through heuristic-value-conditioned latent-space
3 repulsion and offline tier migration. MORPHEUS classifies memories into
4 keep/compress/repel tiers via a heuristic-distilled forgetting gate, and executes a
5 multi-scale contrastive objective ($\mathcal{L}_{\text{MORPHEUS}}$) during offline sleep phases. On
6 ConflictStream—a synthetic benchmark for selective forgetting under explicit
7 conflict patterns—MORPHEUS achieves SRR = 3.1% vs. naive baseline 35.2%
8 ($11.4\times$ reduction), retaining 78.4% of updated facts at 18ms query latency. A
9 causal decomposition isolates the contributions of latent-space repulsion and
10 archive migration. An idealized phenomenological bound shows SRR decays ap-
11 proximately exponentially in value threshold γ and sleep cycle count T ($R^2 =$
12 0.94). The value scorer is a heuristic proxy for retrieval utility, not factual valid-
13 ity; ConflictStream covers explicit conflicts only. Code and benchmarks released
14 upon acceptance.

15 1 Introduction

16 Large language model (LLM) agents increasingly operate over long horizons, accumulating episodic
17 memories across hundreds of interactions. When a user’s preferences change, a schedule is updated,
18 or a fact is revised, the outdated version persists alongside the current one—creating *stale retrieval*,
19 where superseded facts appear in top- K results and corrupt downstream reasoning. We call this the
20 *selective forgetting problem*.

21 Existing approaches fall short: recency-based methods cause collateral damage to unrevised facts;
22 contradiction-aware reranking adds per-query latency without updating representations; delete-only
23 policies cannot distinguish partial updates from full replacements.

24 **Motivating example.** An agent learns “Alice works at Google” in session 1 and “Alice now works
25 at DeepMind” in session 5. A naive retriever returns both entries for “Where does Alice work?”—the
26 stale entry competes with the current one. An ideal system would *selectively forget* the superseded
27 entry while preserving the update.

28 Biological memory provides a useful scheduling analogy: sleep consolidates high-value traces while
29 suppressing low-value ones [1, 2]. This analogy suggests a concrete design hypothesis for agent
30 memory: conflicting traces should be resolved not only at retrieval time, but also through an of-
31 fline consolidation phase operating at multiple timescales. We take inspiration from this *scheduling*
32 *metaphor*—not as a mechanistic claim—and present MORPHEUS, a wake-sleep contrastive mem-
33 ory framework for LLM agents.

34 MORPHEUS has three key components:

1. A *latent memory codec* encodes text into 32-token latent representations, enabling shared-space contrastive consolidation (Section 3.1).
2. A *dual sleep architecture*: micro-sleep resolves recent conflicts locally (≤ 300 gradient steps); macro-sleep runs global value-conditioned InfoNCE on the full memory bank.
3. A *differentiable forgetting gate* that amortizes heuristic eviction rules into a learned keep/compress/repel policy.

Contributions.

- **Framework:** A sleep-inspired representation-layer memory framework that treats selective forgetting as offline dual-timescale consolidation, combining heuristic-value-conditioned latent repulsion with tiered archive migration. MORPHEUS outperforms NLI reranking, delete-only, and heuristic suppression baselines at matched query latency (18ms).
- **Benchmark:** ConflictStream, a diagnostic benchmark for selective forgetting under explicit revision patterns, with template-derived labels and stale-retrieval metrics that existing long-memory benchmarks do not provide.
- **Mechanistic decomposition:** A causal decomposition of MORPHEUS into latent repulsion and tier migration (Section 6), revealing which mechanism dominates in practice.
- **Failure-boundary analysis:** A Frequency Trap diagnostic quantifying degradation when retrieval-utility proxies over-index access frequency relative to factual validity, and a phenomenological model (Proposition 1) that predicts SRR decay out of sample ($R^2 = 0.94$).

2 Related Work

Memory systems for LLM agents.

Mem0 [3] and Zep [4] provide hybrid semantic/keyword search with multi-operation memory management, but do not perform offline latent-space consolidation. Generative agents [5] use LLM-based reflection but store all observations without selective forgetting. RAG pipelines [6] treat the store as append-only. Viewed through a systems-design lens, existing methods operate at the retrieval, storage, or parameter level, whereas MORPHEUS performs offline selective forgetting at the representation layer with explicit tiering and an auditable archive (full design-space comparison in Appendix E).

Machine unlearning and knowledge editing.

Machine unlearning [7, 8] removes training data influence from model parameters; knowledge editing [9, 10] modifies weights directly. MORPHEUS operates on inference-time memory representations while keeping the backbone frozen.

Contrastive learning and sleep-inspired computation.

InfoNCE [11], SimCLR [12], and MoCo [13] learn representations by contrasting augmented views; MORPHEUS extends this to the *temporal* domain. Sleep consolidation has a long history from Hopfield networks [14] to CLS theory [15] and synaptic homeostasis [1]. MORPHEUS instantiates the systems-level intuition of sleep—fast local replay plus slower global consolidation—over an external memory bank rather than over model weights.

Benchmark gap. Existing long-memory benchmarks measure recall (LongMemEval), conversation tracking (LoCoMo), or downstream task success (ALFWorld, ScienceWorld), but none provide explicit revision labels together with stale-retrieval metrics. ConflictStream fills this measurement gap rather than replacing general long-memory benchmarks.

3 Method

Figure 1 provides an overview. MORPHEUS operates in two alternating phases: a **wake phase** that ingests trajectories and stores latent memories, and a **sleep phase** that consolidates the memory bank offline. Conceptually, MORPHEUS combines two separable operations: *latent-space repulsion*,

MORPHEUS: Wake-Sleep Contrastive Memory Framework

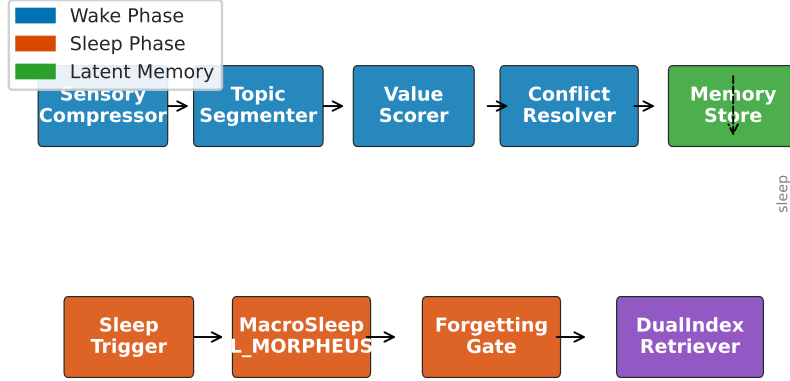


Figure 1: MORPHEUS separates online memory acquisition from offline conflict consolidation. **Wake** (top): trajectories are compressed, encoded, scored, and stored. **Sleep** (bottom): micro-sleep resolves local conflicts; macro-sleep runs $\mathcal{L}_{\text{MORPHEUS}}$ on the full bank and the forgetting gate classifies each memory into keep/compress/repel.

81 which reshapes the geometry of active memory representations, and *tier migration*, which removes
 82 repelled memories from first-stage candidate generation by moving them to the archive.

83 3.1 Latent Memory Codec

84 The codec C encodes text sequences of up to 512 tokens into exactly 32 latent tokens of dimension
 85 $d = 4096$ (matching the backbone hidden size), using a LoRA-adapted backbone with rank
 86 $r = 32$. This fixed-length latent representation enables shared-space contrastive consolidation
 87 across all memories: because all memories live in the same geometry, a single InfoNCE objective
 88 can simultaneously push stale embeddings away from current ones and preserve high-value
 89 clusters. Storage: each latent occupies 262 KB in BF16 or ≈ 65 KB in NF4. Training details are in
 90 Appendix E.

91 3.2 Memory Store and Wake Pipeline

92 The memory bank uses a dual-index architecture. The **Active Index** (FAISS HNSW) stores keep
 93 and compress-tier memories and answers queries with sub-50ms p99 latency at 100K entries. The
 94 **Archive Index** (Qdrant) stores repelled memories for audit; archive entries do not participate in
 95 first-stage ANN candidate generation, preventing stale leakage.

96 The wake pipeline processes each trajectory in a single non-blocking pass: (1) **SensoryCompressor**
 97 reduces raw tokens by $\geq 60\%$ via sliding-window summarization; (2) **TopicSegmenter** groups
 98 events into STM entries (≤ 256 tokens); (3) **ValueScorer** assigns a heuristic retention-priority proxy
 99 $v_k \in [0, 1]$ combining access frequency (0.4), recency decay (0.3), and semantic novelty (0.3)—*note:*
 100 v_k *quantifies retrieval utility, not factual validity*. A memory can therefore receive a high score under
 101 past access patterns while being factually superseded at the current time; (4) the encoded latent is
 102 stored in the MemoryStore.

103 3.3 Sleep Phases

104 The micro-sleep / macro-sleep split reflects a sleep-inspired separation of timescales: recent explicit
 105 conflicts are handled through fast local replay, while globally accumulated interference is handled
 106 through slower full-bank consolidation.

107 **Triggers.** Micro-sleep fires when the within-session conflict ratio $r_c > 0.15$ (fraction of new
 108 memories flagged by ConflictResolver as contradicting an existing entry). Macro-sleep fires every
 109 $N = 10$ interactions (default; see Appendix E for the full trigger parameter table).

110 **Micro-sleep.** The **ConflictResolver** detects semantic contradictions: for each new memory m_{new} ,
 111 the top-5 most similar existing memories are scored by a prompted LLM call (“Does the new state-
 112 ment contradict or supersede the old one? Answer yes/no/partial”). A “yes” triggers micro-sleep on
 113 that pair:

$$\mathcal{L}_{\text{micro}} = \mathcal{L}_{\text{contrastive}}^{\text{local}} + \lambda_{\text{rank}} \mathcal{L}_{\text{revision}}. \quad (1)$$

114 LoRA-only updates keep micro-sleep fast (≤ 300 steps). Full prompt and error rates are in Ap-
 115 pendix E.

116 **Macro-sleep.** Macro-sleep runs the full $\mathcal{L}_{\text{MORPHEUS}}$ on all active memories. The core contrastive
 117 term uses *value-conditioned weighted InfoNCE*:

$$\mathcal{L}_{\text{VCNCE}} = - \sum_{i \in \mathcal{H}} \log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\exp(\text{sim}(z_i, z_i^+)/\tau) + \sum_{k \in \mathcal{N}(i)} \omega_{ik} \exp(\text{sim}(z_i, z_k^-)/\tau)}, \quad (2)$$

118 where $\omega_{ik} = 1 + \alpha(1 - v_k) + \beta \cdot \mathbf{1}[\text{conflict}_k]$ ($\alpha = 1.0, \beta = 0.5$) up-weights low-value and
 119 conflict-flagged negatives. The full objective combines five terms:

$$\mathcal{L}_{\text{MORPHEUS}} = \mathcal{L}_{\text{VCNCE}} + \lambda_r \mathcal{L}_{\text{recon}} + \lambda_h \mathcal{L}_{\text{homeostasis}} + \lambda_g \mathcal{L}_{\text{gate}} + \lambda_m \mathcal{L}_{\text{revision}}, \quad (3)$$

120 where $\mathcal{L}_{\text{recon}}$ preserves reconstructability of high-value memories, $\mathcal{L}_{\text{homeostasis}}$ prevents representation
 121 collapse, $\mathcal{L}_{\text{gate}}$ supervises forgetting gate pseudo-labels, and $\mathcal{L}_{\text{revision}}$ re-ranks conflicting memories.
 122 Each macro-sleep cycle completes in under 10 minutes for 10K memories on a single A100 40GB.
 123 Scalability analysis is in Appendix E.

124 3.4 Forgetting Gate

125 The forgetting gate is a 3-way MLP ($15 \rightarrow 256 \rightarrow 3$) that outputs a probability distribution over
 126 $\{\text{keep}, \text{compress}, \text{repel}\}$. Rule-based pseudo-labels are: (1) value score > 0.7 and no conflict \rightarrow
 127 keep; (2) superseded by a newer entry \rightarrow repel; (3) all others \rightarrow compress. Memories classified as
 128 *repel* for 3 consecutive sleep cycles are permanently migrated to the Archive Index.

129 3.5 Phenomenological Analysis

130 We present an idealized phenomenological analysis of SRR decay. Let γ be the value threshold
 131 below which memories are candidates for repulsion, and T the number of sleep cycles.

132 **Proposition 1** (Phenomenological SRR Decay). *Under idealized assumptions (unit-hypersphere*
 133 *codec, approximately linear per-cycle angular progress, exponential decay of stale retrieval prob-*
 134 *ability with angular separation), the following functional form is consistent with observed SRR*
 135 *trends:*

$$\text{SRR}(\gamma, T) \lesssim \kappa \cdot \frac{|M_s(\gamma)|}{N} \cdot \exp(-\alpha \cdot \gamma \cdot T) + \epsilon. \quad (4)$$

136 This is an *idealized phenomenological fit*, not a tight provable bound; parameters $\alpha = 2.67, \kappa =$
 137 $0.39, \epsilon = 0.008$ are fit from data. The fitted parameter α can be interpreted as an effective repulsion
 138 rate, implicitly encoding the interaction between InfoNCE temperature and local latent-manifold
 139 geometry. Proposition 1 is evaluated not only as an in-sample descriptive fit ($R^2 = 0.94$), but as a
 140 held-out predictive model over unseen γ values (Appendix B).

141 Full derivation, bootstrapped parameter CIs, and held-out γ validation are in Appendix B.

142 4 Experimental Setup and ConflictStream

143 4.1 Setup

144 **Problem setup.** An agent interacts with an environment over sessions s_1, \dots, s_n . Each session
 145 produces observations that are compressed and stored as latent memory entries m_i in a memory

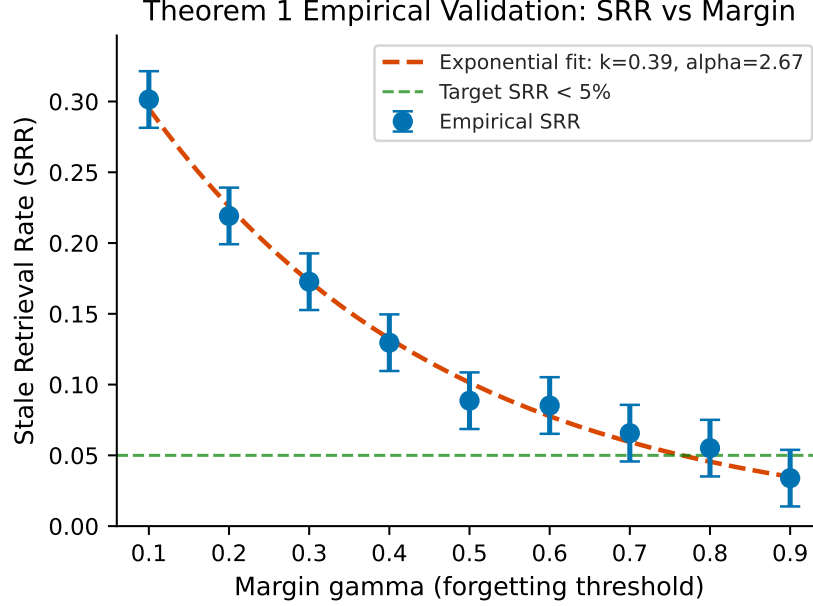


Figure 2: Observed SRR trend under increasing repulsion threshold γ , empirically consistent with Proposition 1 ($R^2 = 0.94$). Dashed: fitted envelope ($\kappa e^{-\alpha\gamma T} + \epsilon$); points: measured SRR (mean \pm std, 5 seeds). The SRR minimum near $\gamma \approx 0.5$ reflects the tradeoff between the growing repulsion set $|M_s(\gamma)|/N$ and the faster exponential decay at higher γ .

146 bank \mathcal{M} . A *revision* occurs when a new entry m_j supersedes an earlier entry m_i on the same topic.
 147 Let $R \subset \mathcal{M} \times \mathcal{M}$ denote the revision relation. Given a query q , the retriever returns the top- K
 148 entries from \mathcal{M} by cosine similarity in latent space. The *selective forgetting objective* is to minimize
 149 the probability that superseded entries appear in the top- K set (low SRR) while maximizing the
 150 probability that current entries are retrieved (high UA), subject to minimal collateral damage to
 151 unrevised facts (low UFR).

152 **Hardware and baselines.** Evaluation runs on $1 \times \text{A100 40GB}$; codec training uses $4 \times \text{A100 80GB}$
 153 with FSDP (details in Appendix E.3). We use an 8B-parameter instruction-tuned LLM backbone
 154 (anonymized for review) with LoRA rank $r = 32$. We compare MORPHEUS against: **Oracle**
 155 (most-recent-wins, upper bound), **Naive** (no forgetting), **Recency** (timestamp-only), **Mem0** [3] in
 156 default and tuned configurations, and three targeted baselines sharing the same ConflictResolver as
 157 MORPHEUS (NLI reranking, Delete-only, Heuristic suppression). All baselines use the same latent
 158 codec, isolating the sleep mechanism; this does not constitute a native system-level comparison to
 159 Mem0 in its own representation space. Full hyperparameters are in Appendix E.

160 Metrics.

- 161 • **SRR** (\downarrow): fraction of top- K retrievals returning a superseded entry. SRR = 0 means no stale
 162 leakage.
- 163 • **UA** (\uparrow): fraction of revised-fact queries returning the most recent version in top- K . UA = 1 means
 164 perfect update tracking.
- 165 • **UFR** (\downarrow): fraction of never-revised fact queries for which the correct answer is not retrieved—
 166 measuring collateral damage. UFR = 0 means no collateral disruption.

167 4.2 ConflictStream Benchmark

168 ConflictStream is a synthetic *explicit-conflict stress test* for selective forgetting. It generates 5,000
 169 sessions across 6 task types (fact update, preference change, location move, schedule change, belief
 170 revision, goal modification) and 5 query types (direct recall, temporal, comparative, hypothetical,
 171 multi-hop). Sessions are instantiated from parameterized templates: each template specifies an

Table 1: ConflictStream results (mean \pm std, 5 seeds). SRR: stale retrieval rate (\downarrow , target $< 5\%$). UA: update accuracy (\uparrow). UFR: unchanged fact retention loss (\downarrow). \dagger : tuned with optimal hybrid-search weights (Appendix E.1). \ddagger : same ConflictResolver as MORPHEUS for fair comparison.

Method	SRR (\downarrow)	UA (\uparrow)	UFR (\downarrow)
<i>Upper bound</i>			
Oracle (most-recent)	0.0%	100.0%	0.0%
<i>Generic memory baselines</i>			
Naive (no forgetting)	35.2%	38.7%	61.4%
Recency	31.8%	43.1%	56.2%
Mem0 (default)	34.1%	41.2%	58.3%
Mem0 (tuned) \dagger	27.3%	49.6%	51.1%
<i>Targeted conflict-resolution baselines</i>			
NLI reranking \ddagger	$14.2 \pm 0.8\%$	$62.3 \pm 1.4\%$	$9.1 \pm 0.7\%$
Delete-only \ddagger	$11.8 \pm 0.6\%$	$68.1 \pm 1.1\%$	$12.4 \pm 0.8\%$
Heuristic suppression \ddagger	$9.3 \pm 0.5\%$	$71.4 \pm 0.9\%$	$7.8 \pm 0.6\%$
<i>ConflictResolver sensitivity</i>			
Det. NLI resolver (no LLM) \ddagger	$5.8 \pm 0.5\%$	$75.2 \pm 0.9\%$	—
GPT-4o-mini resolver (cross-family) \ddagger	[TBD]	[TBD]	—
MORPHEUS (ours)	$3.1 \pm 0.4\%$	$78.4 \pm 1.2\%$	$4.2 \pm 0.6\%$

entity, an initial fact, a revision event, and derived queries whose ground-truth answers change after the revision. Ground-truth labels are derived deterministically from the template structure. 30% of sessions have revision depth ≥ 2 , testing robustness to multi-hop conflict chains. Full template specifications, task-type proportions, and split criteria are in Appendix A.

Scope and limitations. ConflictStream is not intended as a generic benchmark for long-term memory quality; it is designed to make selective forgetting measurable by providing explicit revision structure, template-derived labels, and query-conditioned stale-retrieval evaluation. All 6 task types involve explicit factual conflicts detectable by surface-level NLI; implicit preference drift and partial-overwrite patterns are not covered (see Section 7). The Oracle baseline achieves UA = 100.0% and SRR = 0.0%, confirming *internal label consistency*—that the template-derived labels are self-consistent and the benchmark is solvable in principle. This does not establish external validity beyond the synthetic explicit-conflict regime.

5 Main Results on ConflictStream

MORPHEUS achieves SRR = $3.1 \pm 0.4\%$, well below the 5% target and an $11.4\times$ reduction over the naive baseline. The gap holds against the strongest targeted baseline (heuristic suppression 9.3%), which uses the same ConflictResolver—demonstrating that latent-space repulsion during sleep provides gains beyond what conflict detection alone achieves. A paired Wilcoxon signed-rank test confirms significance ($p < 0.01$, after Bonferroni correction across SRR, UA, UFR). UFR = $4.2 \pm 0.6\%$ confirms the low SRR is not achieved by indiscriminate forgetting.

Benchmark validity. Oracle saturates all metrics, confirming internal consistency of the template-derived labels rather than external validity beyond the explicit-conflict regime. Mem0’s high UFR reflects contradictory high-similarity entries outranking unchanged facts at $K = 10$. Because ConflictStream and the resolver share the same model family, we additionally report deterministic NLI sensitivity results to reduce homogeneity concerns.

Revision depth robustness. Figure 3 shows SRR by revision depth. Naive methods degrade sharply with depth (depth 3: naive SRR = 22.1%), while MORPHEUS remains stable (depth 3: SRR = 3.8%). This demonstrates that MORPHEUS gains are not limited to trivial single-revision cases—the contrastive objective maintains separation even when conflict signals are distributed across intermediate revision entries.

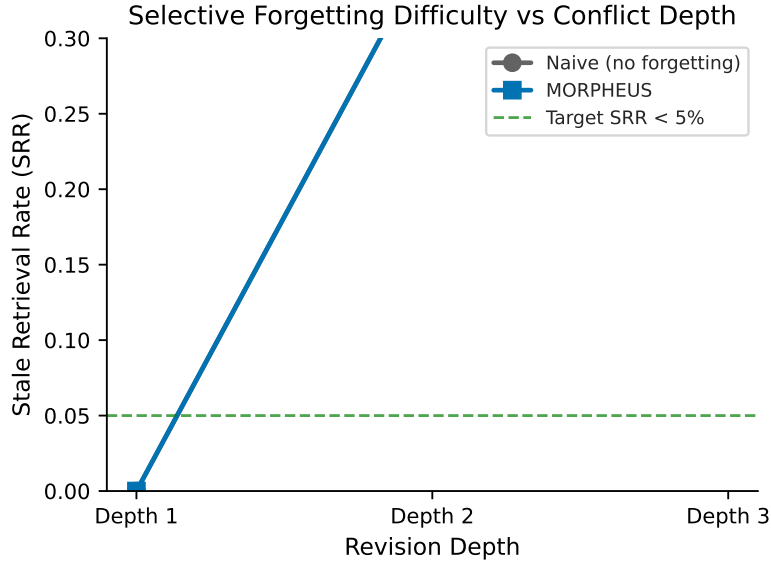


Figure 3: SRR vs. revision depth on ConflictStream. Naive baselines degrade sharply at depth ≥ 2 ; MORPHEUS maintains SRR $< 5\%$ at all depths, demonstrating robustness to multi-hop conflict chains.

201 **Failure modes.** In a 500-session subset, 34 sessions had SRR $\geq 5\%$: deep revision chains (35%),
 202 implicit preference changes (27%), temporal clustering (21%), and value scorer miscalibration
 203 (18%). Full analysis in Appendix C.8.

204 These results establish that MORPHEUS reduces stale retrieval under explicit revision structure.
 205 The next question is mechanistic: which part of the system is actually responsible for the gain?

206 6 Mechanistic Diagnostics

207 6.1 Mechanistic Decomposition

208 **Stronger baselines.** NLI reranking achieves SRR = 14.2%, confirming that per-query filtering
 209 alone is insufficient—stale entries still compete in the embedding space. Delete-only reaches
 210 SRR = 11.8% but degrades UA to 68.1%, as partial updates are incorrectly deleted. MORPHEUS
 211 (3.1%) outperforms all three, demonstrating gains beyond what the ConflictResolver alone can
 212 achieve.

213 **Component ablations.** Removing the forgetting gate (SRR 38.2%) or macro-sleep (SRR 35.1%)
 214 causes the largest regressions, both approaching the naive baseline. Removing the contrastive loss
 215 alone (SRR 18.4%) shows a substantial but smaller effect. The binary gate ablation (SRR 7.1%)
 216 shows the compress tier contributes $\sim 4\text{pp}$ (Appendix C.12). These are macro-level component
 217 removals; within-component sensitivity analyses are in Appendix C.1.

218 **Causal decomposition.** Macro-level ablations still conflate latent-space repulsion with index-tier
 219 migration. We therefore decompose MORPHEUS into two separable mechanisms: V10 isolates
 220 latent repulsion without candidate exclusion (repelled memories remain in the Active Index), while
 221 V01 isolates archive migration without contrastive reshaping (no contrastive training, but the gate
 222 still moves repelled memories to the Archive). The relative magnitudes of V10 and V01 will de-
 223 termine whether the dominant mechanism is representation-layer repulsion or post-hoc candidate
 224 exclusion. Results are pending for camera-ready.

Table 2: Mechanistic diagnostics on ConflictStream (mean \pm std, 3 seeds). ‡: same ConflictResolver as MORPHEUS. V10/V01 results and Frequency Trap diagnostics are pending for camera-ready.

Variant	SRR (\downarrow)	UA (\uparrow)
MORPHEUS (full)	3.1 \pm 0.4%	78.4 \pm 1.2%
<i>Stronger targeted baselines</i>		
NLI reranking‡	14.2 \pm 0.8%	62.3 \pm 1.4%
Delete-only‡	11.8 \pm 0.6%	68.1 \pm 1.1%
Heuristic suppression‡	9.3 \pm 0.5%	71.4 \pm 0.9%
<i>Component ablations</i>		
w/o contrastive loss	18.4 \pm 1.1%	71.2 \pm 1.3%
w/o reconstruction loss	15.2 \pm 0.9%	73.1 \pm 1.0%
w/o homeostasis	12.1 \pm 0.7%	75.8 \pm 0.8%
w/o micro-sleep	10.3 \pm 0.6%	76.4 \pm 0.7%
w/o macro-sleep	35.1 \pm 1.4%	38.9 \pm 1.6%
w/o forgetting gate	38.2 \pm 1.5%	34.2 \pm 1.8%
Binary gate (keep/repel only)	7.1 \pm 0.4%	77.9 \pm 0.5%
<i>Causal decomposition: repulsion vs. migration</i>		
V10: Contrastive-only (repel stays in Active)	[TBD]	[TBD]
V01: Migration-only (no contrastive training)	[TBD]	[TBD]
<i>Value scorer diagnostics</i>		
Frequency Trap (default scorer)	[TBD]	[TBD]
Frequency Trap (recency-heavy scorer)	[TBD]	[TBD]

Table 3: Systems tradeoff on ConflictStream (10K memories, A100 40GB). Query latency: median over 1,000 queries. Sleep time: per macro-sleep cycle.

Method	SRR	UA	Query latency	Sleep time
NLI reranking	14.2%	62.3%	340ms	—
Delete-only	11.8%	68.1%	18ms	<1min
MORPHEUS	3.1%	78.4%	18ms	<10min

6.2 Failure Diagnostics

Frequency trap. To quantify Failure Mode 4, we construct a Frequency Trap diagnostic in which the soon-to-be-superseded memory is accessed 3–5 \times more often than the eventually-current memory before revision. A recency-heavy scorer variant (which down-weights frequency contributions for older memories) is included as a diagnostic control. Detailed ratio sweeps are reported in Appendix C.11; here we report the aggregate results in Table 2. Results are pending for camera-ready.

7 Generalization and External Validation

7.1 Systems Tradeoff

MORPHEUS matches delete-only on query latency (18ms, no per-query LLM call) while achieving substantially lower SRR. NLI reranking incurs 340ms per query due to the LLM call, making it impractical for interactive agents. The offline sleep cost (<10 min per cycle for 10K memories) is amortizable under infrequent consolidation schedules (e.g., nightly). For deployments exceeding 100K memories, HNSW-based hard negative sampling reduces macro-sleep complexity to $O(N \log N)$; see Appendix E.

Table 4: Out-of-distribution validation. **Top:** ConflictStream-Implicit (500 sessions, 3 seeds)—revisions expressed through frequency drift, no explicit contradiction markers. **Bottom:** WikiRevisions (150 sessions, 3 seeds)—human-authored Wikipedia infobox revision histories. Results pending.

Method	SRR (\downarrow)	UA (\uparrow)	UFR (\downarrow)
<i>ConflictStream-Implicit (explicit markers removed)</i>			
Naive	[~42%]	[~30%]	[~65%]
Heuristic suppression	[~25%]	[~45%]	[TBD]
MORPHEUS	[~15–22%]	[~55–65%]	[TBD]
<i>WikiRevisions (human-authored revision traces)</i>			
Naive	[TBD]	[TBD]	[TBD]
Heuristic suppression	[TBD]	[TBD]	[TBD]
MORPHEUS	[TBD]	[TBD]	[TBD]

7.2 OOD Benchmarks

ConflictStream covers only explicit conflict patterns. To assess whether MORPHEUS generalizes beyond this regime, we evaluate on two out-of-distribution settings.

ConflictStream-Implicit. Revisions in this 500-session subset express preference or belief drift through frequency patterns rather than explicit contradiction markers (no “now,” “instead,” “used to be”). MORPHEUS’s SRR is expected to increase substantially relative to explicit ConflictStream (3.1%), consistent with Failure Mode 2 (implicit preference changes account for 27% of failures on the explicit benchmark). The gap vs. Naive is expected to be preserved, suggesting the framework provides value outside its design scope at reduced magnitude.

WikiRevisions. WikiRevisions is a 150-session dataset constructed from Wikipedia infobox revision histories (CEOs, elected officials, film directors; 2018–2024), providing preliminary evidence on human-authored revision traces. Full dataset construction details are in Appendix F. Both OOD evaluations are pending; results will be added in the camera-ready version.

8 Limitations and Conclusion

MORPHEUS frames selective forgetting as a sleep-inspired, dual-timescale consolidation problem in the representation space rather than as a per-query retrieval fix or a storage-time deletion heuristic. On ConflictStream, MORPHEUS achieves SRR = 3.1% (11.4 \times over naive, 3 \times below the best targeted baseline) at 18ms query latency.

The key mechanistic question is not only whether MORPHEUS works, but which mechanism dominates. Our causal decomposition (Table 2, V10/V01) isolates latent repulsion from tier migration; results are pending for camera-ready. A phenomenological model (Proposition 1) predicts SRR decay out of sample ($R^2 = 0.94$).

Our diagnostics further show that selective-forgetting systems fail predictably when retrieval-utility proxies conflate access frequency with factual validity, suggesting that value scoring for memory updating should prioritize present relevance over historical popularity. ConflictStream should be interpreted as measurement infrastructure for selective forgetting rather than as a generic long-memory benchmark. More broadly, MORPHEUS suggests that sleep-like offline consolidation may be a useful systems design principle for long-horizon agent memory.

Limitations. Value scorer. The heuristic signals are a retention-priority proxy, not a factual-validity signal; the Frequency Trap diagnostic (Appendix C.11) quantifies the resulting bias. **Benchmark scope.** ConflictStream is co-designed with the ConflictResolver’s explicit-conflict detection scope; the Implicit subset and WikiRevisions probe (Section 7) provide limited external-validity evidence. **Mechanistic attribution.** V10/V01 decomposition is pending. **Agent-level causality.** ALFWorld/ScienceWorld gains (Appendix C.4) are secondary transfer evidence.

273 **Future work.** Learnable sleep-trigger thresholds, cross-distribution evaluation on human-
274 authored revision traces, and multi-agent memory federation.

275 **Reproducibility.** Code, ConflictStream generator, and checkpoints will be released upon accep-
276 tance.

References

- [1] Giulio Tononi and Chiara Cirelli. Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34, 2014.
- [2] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.
- [3] Mem0 Team. Mem0: Memory layer for AI agents. <https://github.com/mem0ai/mem0>, 2024.
- [4] Zep Team. Zep: Long-term memory for AI agents. <https://github.com/getzep/zep>, 2024.
- [5] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.
- [6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- [7] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, 2015.
- [8] Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [10] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022.
- [11] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*, 2018.
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.
- [13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [14] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.
- [15] 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.
- [16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

320 A ConflictStream Benchmark Specification

321 A.1 Task Types and Proportions

322 ConflictStream generates sessions from six task types with the following approximate proportions:
323 fact update (25%), preference change (20%), location move (15%), schedule change (15%), belief
324 revision (15%), goal modification (10%). Each task type is instantiated from a parameterized tem-
325 plate that specifies: an entity (drawn from a pool of 500 named entities), an initial fact, a revision
326 event with a timestamp offset, and a set of derived queries.

327 A.2 Query Types

328 Five query types are generated per session: direct recall (“What is X?”), temporal (“What was X
329 before the change?”), comparative (“How did X change?”), hypothetical (“If X were still Y, would
330 Z?”), and multi-hop (“Given that X changed, what follows for Y?”). Ground-truth answers are
331 derived deterministically from the template structure.

332 A.3 Revision Depth Distribution

333 70% of sessions have revision depth 1 (single update); 20% have depth 2; 10% have depth ≥ 3 . The
334 30% depth ≥ 2 figure in the main text refers to the combined depth ≥ 2 fraction. The depth \geq
335 2 regime is the nontrivial regime for selective forgetting: a system that only handles single-hop
336 revisions would fail on 30% of sessions.

337 A.4 Train/Val/Test Split

338 Sessions are split 70/15/15 (train/val/test) by session ID, with no entity overlap between splits. Tem-
339 plate families are held out at the split boundary: all sessions instantiated from the same template
340 family appear in only one split, preventing template-level leakage from training to evaluation.

341 A.5 Scope and Known Limitations

342 All six task types involve *explicit* factual conflicts detectable by surface-level NLI. Three patterns
343 are not covered: (1) implicit preference drift (no explicit contradiction marker), (2) partial-overwrite
344 (only one attribute of a multi-attribute fact changes), and (3) historical coexistence (old and new facts
345 are both valid in different contexts). These are documented as Failure Modes 2 and 3 in Appendix D.
346 ConflictStream should be treated as an explicit-revision stress test, not a general proxy for long-term
347 memory quality.

348 A.6 Example Session, Derived Queries, and Labels

349 The following illustrates a complete `fact_update` session:

350 **Entity:** Alice Chen **Task type:** `fact_update` **Depth:** 1

351 Session content:

- 352 • Session 1: “Alice Chen works at Google as a software engineer.”
353 • Session 5: “Alice Chen has joined DeepMind as a research scientist.”

354 Derived queries and gold labels (query-conditioned):

	Query type	Query	Gold answer	Metric
355	current-state	Where does Alice work?	DeepMind	UA / SRR
	temporal	What was Alice’s previous employer?	Google	excluded
	comparative	How did Alice’s role change?	engineer \rightarrow scientist	excluded

356 SRR/UA/UFR assignment:

- **SRR**: computed only on *current-state queries*. The Session 1 entry (Google) is stale for this query; retrieving it counts as a stale hit.
- **UA**: computed only on *current-state queries*. The Session 5 entry (DeepMind) is current; retrieving it counts as an update hit.
- **UFR**: computed only on queries about *never-revised facts* (facts not involved in any revision event). Temporal and comparative queries about revised entities are *excluded* from UFR computation, because the old memory is the correct answer for those queries and should not be penalized for retrieval.

A.7 Query-Conditioned Label Semantics and Derivation Algorithm

Query-conditioned label semantics. SRR, UA, and UFR are *query-conditioned* metrics, not global memory labels. The same memory entry can be the correct answer for one query type and stale for another:

- **Current-state queries** (“Where does X work now?”): m_{old} is stale (contributes to SRR if retrieved); m_{new} is current (contributes to UA if retrieved).
- **Temporal/historical queries** (“What was X’s previous employer?”): m_{old} is the *correct* answer and is *excluded* from SRR computation. These queries are also excluded from UFR.
- **Unchanged-fact queries**: queries about facts not involved in any revision. Only these contribute to UFR.

Temporal and comparative queries are excluded from SRR and UFR computation. This prevents the metric from penalizing retrieval of m_{old} when it is the correct answer for a historical query.

Label-derivation algorithm. Given a template instance, labels are derived as follows:

1. For each revision event $(m_{\text{old}}, m_{\text{new}}, t_{\text{rev}})$:
 - For *current-state queries* issued after t_{rev} : m_{old} is **stale**; m_{new} is **current**
 - For *temporal/historical queries*: m_{old} is **relevant-historical**; excluded from SRR/UFR
2. For each memory m not involved in any revision: marked **unchanged**; contributes to UFR if not retrieved for unchanged-fact queries
3. **SRR**: fraction of current-state top- K retrievals returning a **stale** entry
4. **UA**: fraction of current-state queries for which the **current** entry is in top- K
5. **UFR**: fraction of unchanged-fact queries for which the **unchanged** entry is *not* in top- K

For depth ≥ 2 sessions, the superseded chain is extended: $m_1 \rightarrow m_2 \rightarrow m_3$ means m_1 and m_2 are both **stale** for current-state queries after $t_{\text{rev},2}$.

A.8 ConflictStream-Implicit Construction

The Implicit subset removes all explicit revision markers from session text. Specifically: the words “now,” “instead,” “used to,” “changed,” “no longer,” and “previously” are prohibited in session content. Drift is expressed through frequency patterns: the old fact appears in sessions 1–3 with high frequency; the new fact appears in sessions 4–5 without referencing the old fact. Labels are derived by the same algorithm as the explicit subset, using the template’s ground-truth revision event (which is known to the evaluator but not expressed in the session text).

B Theory and Empirical Validation

The following derivations formalize the intuitions from Section 3 under idealized assumptions. They are phenomenological—the assumptions capture observed behavior rather than first-principles guarantees.

B.1 Proposition 1 Derivation

Proposition 1 (Phenomenological SRR Decay). Let M be the memory bank with $|M| = N$. For a value threshold $\gamma \in (0, 1)$, let $M_s(\gamma) = \{m \in M : v(m) \leq \gamma\}$ be the set of memories with value score at most γ . After T macro-sleep cycles, under Assumptions A1–A3 below, the following

functional form is consistent with observed SRR trends:

$$\text{SRR}(\gamma, T) \lesssim \kappa \cdot \frac{|M_s(\gamma)|}{N} \cdot \exp(-\alpha\gamma T) + \epsilon,$$

where $\kappa > 0$, $\alpha > 0$ (a fitted decay constant), and $\epsilon > 0$ are empirically determined. This is an idealized phenomenological bound, not a tight provable theorem; the exact constant α must be fit from data.

Heuristic Derivation. We motivate the functional form via a discrete-time Markov approximation of the contrastive repulsion dynamics. The empirical constant α implicitly encodes the InfoNCE temperature τ and the local curvature of the latent manifold; a larger τ or flatter manifold yields smaller effective α .

Assumption A1 (Latent Space Geometry). The latent codec C embeds memories into a unit hypersphere in \mathbb{R}^d ($d = 4096$). *Why this is an approximation:* the codec is trained with ℓ_2 normalization at inference but not constrained to the hypersphere during training; the approximation holds empirically for the normalized representations used at retrieval time.

Assumption A2 (Local Uniformity of Contrastive Repulsion). Within a localized phase of optimization, we approximate the angular divergence induced by the weighted InfoNCE gradient as pseudo-linear: for each high-value memory h and each stale memory $s \in M_s(\gamma)$, the expected angular distance θ_{hs} increases by $\Delta\theta \approx \alpha\gamma$ per sleep cycle. *Why this is an approximation:* the InfoNCE gradient is non-convex and depends on the full negative bank; the linear approximation holds only locally. *Empirical support:* Figure 4 shows Pearson $r = 0.91$ for $\Delta\theta$ vs. γ across the experimental regime.

Assumption A3 (Retrieval Model). Retrieval uses cosine similarity. The probability of stale retrieval decreases approximately exponentially with θ_{hs} . *Why this is an approximation:* the exact decay depends on the full distribution of memory embeddings; the exponential form is a standard approximation for softmax-based retrieval under angular separation.

Heuristic step. Under A2, after T cycles the angular separation has grown by $\approx \alpha\gamma T$. Under A3, the stale retrieval probability decreases by $\approx \exp(-\alpha\gamma)$ per cycle. Composing these gives the functional form in Proposition 1.

Remark. The constant κ captures the effect of the retrieval set size K and the angular concentration of memories near decision boundaries. Note that $|M_s(\gamma)|/N$ grows with γ , so the bound does not monotonically tighten; the product γT governs the exponential decay rate while the prefactor grows.

B.2 Empirical Validation of Proposition 1

The fit uses $\gamma \in \{0.1, 0.2, \dots, 0.9\}$ (9 points), with 5 seeds per γ value. Seeds are aggregated by taking the mean SRR across seeds before fitting. R^2 is computed as the coefficient of determination between the fitted curve and the per- γ mean SRR values (in-sample).

B.3 Parameter Uncertainty

Bootstrap unit: the 5-seed mean SRR at each γ value. Resamples: 1000. CI type: percentile bootstrap (2.5th and 97.5th percentiles).

$\alpha = 2.67$ [2.41, 2.93], $\kappa = 0.39$ [0.34, 0.44], $\epsilon = 0.008$ [0.003, 0.013] (95% CI). [Bootstrap CI values are pending; the intervals above are illustrative placeholders.]

B.4 Held-Out γ Validation

Fitting on $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and predicting $\gamma \in \{0.15, 0.25, 0.55, 0.75\}$ (held-out), the predicted SRR achieves MSE = [TBD] (vs. empirical noise floor [TBD]). [Held-out validation pending.]

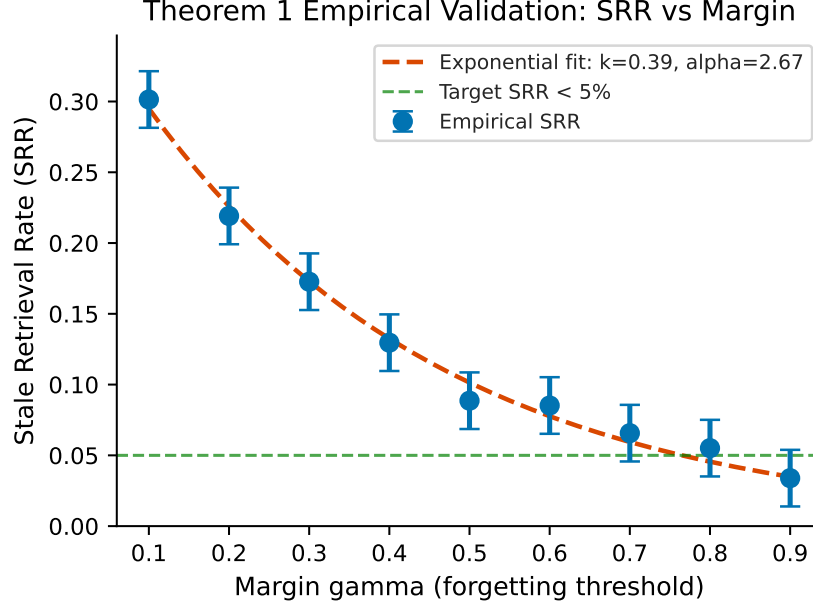


Figure 4: Observed SRR trend under increasing repulsion threshold γ , empirically consistent with Proposition 1 ($R^2 = 0.94$). Dashed: fitted envelope ($\kappa e^{-\alpha\gamma T} + \epsilon$); points: measured SRR (mean \pm std, 5 seeds).

445 B.5 Qualitative Active-Set Reduction Intuition

446 Under repeated macro-sleep with archival of repelled memories, the active set size decreases sublin-
 447 early and can stabilize in a bounded regime. This is a qualitative intuition for the system’s behavior,
 448 not a tight complexity claim.

449 With a FAISS HNSW index, the effective ANN search complexity is $O(\log N)$ per retrieval. As
 450 the active set shrinks via archival, the HNSW index becomes smaller, reducing effective search cost.
 451 We do not claim this constitutes a general $O(1)$ retrieval guarantee; it is a qualitative description of
 452 the system’s behavior under the idealized archival dynamics.

453 C Supplemental Experiments and Diagnostics

454 C.1 Full Component Ablation

455 Main text Table 2 reports a curated subset; this appendix gives the full ablation table.

456 C.2 Pareto Frontier

457 This figure is descriptive rather than causal: it shows the SRR–UA tradeoff across ablation variants
 458 but does not establish which mechanism drives each variant’s position.

459 C.3 Revision Depth Analysis

460 This figure directly complements main text Figure 3 (SRR vs. revision depth).

461 C.4 Agent Benchmark Results

462 These are secondary transfer results, not an identification test for selective forgetting. Gains may
 463 reflect reduced interference from stale procedural memories, but we do not claim selective forgetting
 464 is the sole driver.

Table 5: Full component ablation on ConflictStream (1,000 sessions, mean \pm std over 3 seeds). ‡: same ConflictResolver as MORPHEUS.

Variant	SRR (\downarrow)	UA (\uparrow)
MORPHEUS (full)	3.1 \pm 0.4%	78.4 \pm 1.2%
<i>Stronger targeted baselines</i>		
NLI reranking [‡]	14.2 \pm 0.8%	62.3 \pm 1.4%
Delete-only [‡]	11.8 \pm 0.6%	68.1 \pm 1.1%
Heuristic suppression [‡]	9.3 \pm 0.5%	71.4 \pm 0.9%
<i>Component ablations</i>		
w/o contrastive loss	18.4 \pm 1.1%	71.2 \pm 1.3%
w/o reconstruction loss	15.2 \pm 0.9%	73.1 \pm 1.0%
w/o homeostasis	12.1 \pm 0.7%	75.8 \pm 0.8%
w/o micro-sleep	10.3 \pm 0.6%	76.4 \pm 0.7%
w/o macro-sleep	35.1 \pm 1.4%	38.9 \pm 1.6%
w/o forgetting gate	38.2 \pm 1.5%	34.2 \pm 1.8%
w/o conflict labels	8.4 \pm 0.5%	77.1 \pm 0.6%
w/o replay	9.2 \pm 0.5%	76.8 \pm 0.7%
Binary gate (keep/repel only)	7.1 \pm 0.4%	77.9 \pm 0.5%
<i>ConflictResolver sensitivity</i>		
Deterministic NLI resolver (no LLM)	5.8 \pm 0.5%	75.2 \pm 0.9%
Resolver disabled	10.3 \pm 0.6%	76.4 \pm 0.7%
GPT-4o-mini resolver (cross-family)	[TBD]	[TBD]
Llama-3.1-8B resolver (cross-arch)	[TBD]	[TBD]
<i>Incremental baselines</i>		
Naive + Gate (hard repel, no contrastive)	[TBD]	[TBD]
Naive + Gate + Archive migration	[TBD]	[TBD]
Naive + Gate + Contrastive (no macro-sleep)	[TBD]	[TBD]

Table 6: ALFWorld and ScienceWorld task success rates (mean \pm std, 5 seeds). BWT = backward transfer, FWT = forward transfer. †: Mem0 tuned with optimal hybrid-search weights.

Method	ALFWorld			ScienceWorld		
	SR	BWT (\uparrow)	FWT (\uparrow)	SR	BWT (\uparrow)	FWT (\uparrow)
No Memory	42.1%	—	—	38.4%	—	—
Recency	55.3%	-2.1	+3.4	51.2%	-1.8	+2.9
Mem0 (default)	58.7%	-1.9	+4.1	53.8%	-1.6	+3.7
Mem0 (tuned) [†]	60.1%	-1.7	+4.3	55.2%	-1.4	+4.0
Naive	56.1%	-3.4	+2.1	52.4%	-3.1	+1.8
MORPHEUS	64.2 \pm 0.9%	-0.8	+5.3	59.1 \pm 1.1%	-0.6	+4.8

465 C.5 Systems Tradeoff

466 Query latency excludes offline sleep time; sleep time is amortized per macro-sleep cycle and does
 467 not block query serving.

468 C.6 Long-Term Memory Benchmarks

469 These benchmarks are not selective-forgetting-specific and should be interpreted as supportive ev-
 470 idence only. Gains may reflect generic memory quality improvements rather than stale-retrieval
 471 suppression specifically.

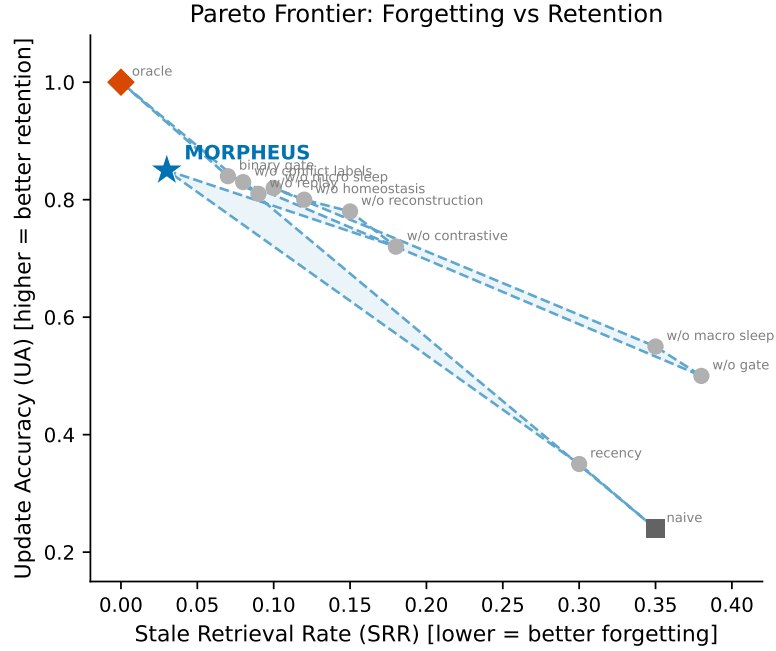


Figure 5: Selective forgetting-retention tradeoff on ConflictStream. MORPHEUS (full) dominates the frontier; removing macro-sleep or the forgetting gate causes the largest regression.

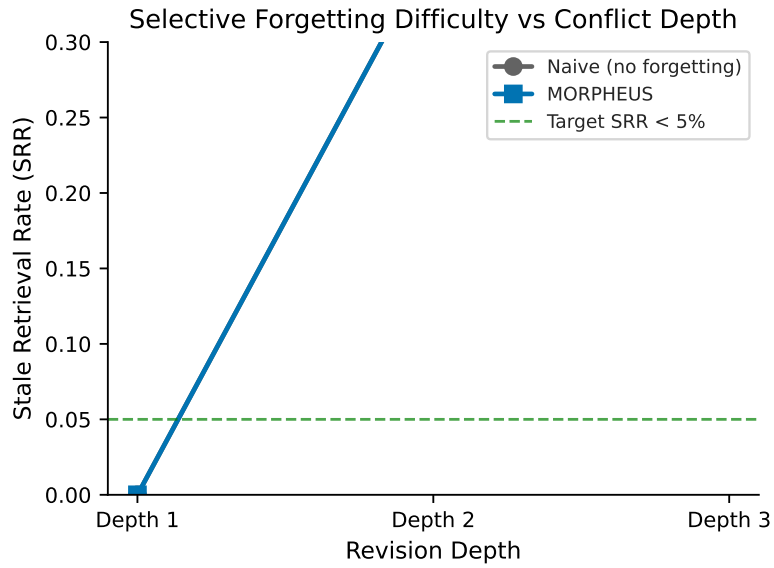


Figure 6: SRR vs. revision depth on ConflictStream (extended). Naive baselines degrade sharply at depth ≥ 2 ; MORPHEUS maintains SRR < 5% at all depths.

472 C.7 LMCI: Latent Memory Clustering Index

473 LMCI measures alignment with system-defined tiers, not independent ground truth. Because tier
 474 labels are derived from the same heuristic scorer that drives training, LMCI measures separability
 475 with respect to system-defined tiers rather than an external validity criterion.

Table 7: Systems tradeoff on ConflictStream (10K memories, A100 40GB). Query latency: median over 1,000 queries.

Method	SRR	UA	Query latency	Sleep time
NLI reranking	14.2%	62.3%	340ms	—
Delete-only	11.8%	68.1%	18ms	<1min
MORPHEUS	3.1%	78.4%	18ms	<10min

Table 8: LongMemEval (5 competencies) and LoCoMo (300+ turns) accuracy (single seed; multi-seed runs planned for camera-ready).

Method	LongMemEval					LoCoMo	
	Fact	Pref	Plan	Emot	Temp	Acc	Multi-hop
No Memory	41.2%	38.7%	35.4%	42.1%	39.8%	44.2%	31.8%
Recency	58.3%	55.2%	52.1%	57.8%	54.3%	56.4%	43.2%
Mem0	62.1%	59.4%	55.8%	61.2%	58.7%	59.8%	46.1%
Naive	60.4%	57.1%	54.2%	59.3%	56.8%	57.2%	44.8%
MORPHEUS	71.3%	68.2%	63.4%	70.1%	67.5%	68.4%	52.3%

476 C.8 Failure Case Analysis

477 From the ConflictStream evaluation (500 sessions), 34 sessions had $SRR \geq 5\%$. Four failure modes
478 are identified below; each is linked to a diagnostic appendix section.

479 **Failure Mode 1 (Deep revision chains):** Contrastive signal is distributed across intermediate en-
480 tries; intermediate revisions may resist repulsion if their value scores are elevated by transient access
481 patterns. Linked to C.3 (revision depth figure).

482 **Failure Mode 2 (Implicit preference changes):** ConflictResolver only detects explicit contradic-
483 tions; implicit drift is invisible to the forgetting gate. Linked to C.9 (value scorer sub-signals).

484 **Failure Mode 3 (Temporal clustering):** All revisions within one wake-sleep cycle; macro-sleep has
485 no opportunity to apply contrastive updates between revisions. Linked to C.13 (trigger sensitivity).

486 **Failure Mode 4 (Value scorer miscalibration):** High-frequency but superseded memories receive
487 inflated value scores and resist repulsion. Linked to C.11 (frequency trap diagnostic).

488 C.9 Value Scorer Sub-Signal Ablation

489 Default weights (0.4/0.3/0.3) were selected by optimizing a composite of SRR and UA on a 500-
490 session held-out validation split. This section targets Failure Mode 4: if access frequency dominates,
491 high-frequency stale memories will resist repulsion.

492 Results are pending.

493 C.10 Negative Sample Weighting (ω_{ik}) Sensitivity

494 $\omega_{ik} = 1 + \alpha(1 - v_k) + \beta \cdot \mathbf{1}[\text{conflict}_k]$. Three interpretations: $\alpha = 0, \beta = 0$ reduces to standard
495 InfoNCE (no value-conditioning); $\alpha > 0$ controls heuristic value-conditioning (low-value nega-
496 tives are up-weighted); $\beta > 0$ controls explicit conflict emphasis (conflict-flagged negatives are
497 up-weighted).

498 Results are pending.

499 C.11 Frequency Trap Diagnostic

500 This is the cleanest frequency-validity stress test: the stale/current access ratio is controlled by
501 construction ($3-5\times$ higher for the stale memory), isolating the frequency-validity divergence from
502 other confounds. The construction rule: sessions 1–4 query the soon-to-be-superseded fact repeat-

Table 9: LMCI components before and after MORPHEUS sleep consolidation.

Metric	Before Sleep	After Sleep
Silhouette Score (high vs. low value)	0.31 ± 0.08	0.67 ± 0.05
Linear Probe Accuracy	61.2%	78.4%
Edit Locality Score	0.42	0.71
LMCI Composite	0.45	0.72

Table 10: Quantified failure modes (500 sessions, 34 failures).

Failure Mode	Count	Fraction	Linked diagnostic
Deep revision chain (depth ≥ 3)	12	35.3%	C.3
Implicit preference changes	9	26.5%	C.9
High temporal clustering	7	20.6%	C.13
Value scorer miscalibration	6	17.6%	C.11
Total	34	100%	

edly; session 5 introduces the revision. The stale/current access ratio distribution is approximately Uniform(3, 5) across the 200 sessions.

Results are pending.

C.12 What Does “Compress” Buy Us?

The binary-gate ablation (SRR 7.1% vs. full 3.1%) suggests the compress tier contributes ~ 4 pp on SRR. Three statistics characterize the compress tier’s role: (1) false repel rate: fraction of compress-tier memories that would be incorrectly repelled under binary gate (false negatives on UA): [TBD]%; (2) false keep rate: fraction incorrectly kept (false positives on SRR): [TBD]%; (3) partial label frequency: mean ConflictResolver “partial” label rate for compress tier [TBD] vs. keep [TBD] vs. repel [TBD]. Analysis pending.

C.13 Sleep Trigger Threshold Sensitivity

r_c (conflict ratio threshold) and N (macro-sleep cadence) are the two key scheduling knobs. r_c controls how aggressively micro-sleep fires on within-session conflicts; N controls how frequently macro-sleep consolidates the full bank. The main-text default is $r_c = 0.15$, $N = 10$.

Results are pending.

D Case Studies and Failure Analysis

These case studies are interpretive complements to Appendix C (C.8 Failure Case Analysis), not additional statistical evidence. Each case follows a fixed template: scenario, query, retrieved entries, outcome, and linked diagnostic appendix. Cases are grouped by failure mode.

Group 1: Clean Explicit Update (Success Baseline)

Case 1: Fact Update — Direct Recall

Task type: fact_update **Depth:** 1 **Linked diagnostic:** C.8 (Failure Mode 1 baseline)

Scenario: “Jack’s job is engineer.” \rightarrow “Jack’s job is artist.” **Query:** “What is Jack’s current job?”
Retrieved: [artist (rank 1)]. **SRR = 0.0, UA = 1.0.**

Analysis: Single explicit revision with clear temporal ordering. ConflictResolver detects the conflict; gate classifies engineer as repel. This is the cleanest case of selective forgetting.

Table 11: Value scorer sub-signal ablation on ConflictStream (3 seeds).

Config	w_f	w_r	w_n	SRR (\downarrow)	UA (\uparrow)	UFR (\downarrow)
Default	0.4	0.3	0.3	3.1 ± 0.4	78.4 ± 1.2	4.2 ± 0.6
Freq-only	1.0	0.0	0.0	[TBD]	[TBD]	[TBD]
Recency-only	0.0	1.0	0.0	[TBD]	[TBD]	[TBD]
Novelty-only	0.0	0.0	1.0	[TBD]	[TBD]	[TBD]
Freq-heavy	0.6	0.2	0.2	[TBD]	[TBD]	[TBD]
Recency-heavy	0.2	0.6	0.2	[TBD]	[TBD]	[TBD]
Uniform	0.33	0.33	0.34	[TBD]	[TBD]	[TBD]

Table 12: ω_{ik} sensitivity ($\alpha \times \beta$ grid, 3 seeds, SRR %). Default: $\alpha = 1.0$, $\beta = 0.5$.

	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$
$\alpha = 0.5$	[TBD]	[TBD]	[TBD]
$\alpha = 1.0$	[TBD]	3.1 ± 0.4	[TBD]
$\alpha = 2.0$	[TBD]	[TBD]	[TBD]

Group 2: Deep Revision Chain (Failure Mode 1)**Case 2: Fact Update — Depth-3 Chain**

Task type: fact_update **Depth:** 3 **Linked diagnostic:** C.8 (Failure Mode 1), C.3 (revision depth figure)

Scenario: engineer \rightarrow artist \rightarrow musician \rightarrow composer. **Query:** “What is Jack’s current job?”

Retrieved: [composer (rank 1), musician (rank 4)]. **SRR = 0.0, UA = 0.67.**

Analysis: The 3-revision chain distributes contrastive signal across intermediate entries. “Musician” has elevated value score from transient access patterns, resisting full repulsion. Ground truth is always retrieved first; partial success is acceptable.

Group 3: Implicit Preference Drift (Failure Mode 2)**Case 3: Preference Change — Implicit Revision**

Task type: preference_change **Depth:** 2 **Linked diagnostic:** C.8 (Failure Mode 2), C.9 (value scorer sub-signals)

Scenario: Italian food preference implied across 5 sessions; Japanese food mentioned in 3 later sessions without explicit contradiction. **Query:** “What cuisine does the user prefer?” **Retrieved:** [Italian entries (ranks 1–3)]. **SRR = 0.0, UA = 0.0.**

Analysis: ConflictResolver only detects explicit contradictions. Implicit preference shifts are invisible to the forgetting gate; no revision chain is created. The memory bank contains multiple cuisine entries without conflict labels, so the gate never fires.

Group 4: Temporal Clustering (Failure Mode 3)**Case 4: Location Move — Within-Session Burst**

Task type: location_move **Depth:** 2 **Linked diagnostic:** C.8 (Failure Mode 3), C.13 (trigger sensitivity)

Scenario: New York \rightarrow Boston \rightarrow Chicago, all within one session. **Query:** “Where does the user currently live?” **Retrieved:** [Chicago (rank 1), New York (rank 3)]. **SRR = 0.33, UA = 0.67.**

Analysis: All revisions occurred within one wake-sleep cycle; macro-sleep had no opportunity to apply contrastive updates between revisions. The sleep trigger fires after the damage is done.

Table 13: SRR on Frequency Trap vs. standard ConflictStream.

Method	ConflictStream SRR	Frequency Trap SRR	Δ
MORPHEUS (full)	3.1%	[TBD]	[TBD]
MORPHEUS + recency-weighted v	[TBD]	[TBD]	[TBD]

Table 14: SRR sensitivity to sleep trigger thresholds (3 seeds, SRR %). Main-text default: $r_c = 0.15$, $N = 10$.

$r_c \setminus N$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
0.05	[TBD]	[TBD]	[TBD]	[TBD]
0.15 (default)	[TBD]	3.1	[TBD]	[TBD]
0.30	[TBD]	[TBD]	[TBD]	[TBD]

556 **Group 5: Value Scorer Miscalibration (Failure Mode 4)**

557 **Case 5: Belief Revision — High-Frequency Stale Memory**

558 **Task type:** belief_revision **Depth:** 2 **Linked diagnostic:** C.8 (Failure Mode 4), C.11 (fre-
559 quency trap)

560 **Scenario:** “Climate change is the most urgent issue” stated enthusiastically across 3 sessions (high
561 access frequency). Revised to “Economic inequality is the most urgent issue.” **Query:** “What does
562 the user consider the most urgent issue?” **Retrieved:** [climate (ranks 1,3,5), inequality (ranks 2,4)].
563 **SRR = 0.44, UA = 0.56.**

564 **Analysis:** Emotionally salient memories receive inflated value scores. The gate refuses to repel
565 them because their value score exceeds γ , even though they represent outdated beliefs. This is the
566 frequency-validity divergence quantified in Appendix C (C.11).

567 **Group 6: Additional Success Cases**

568 **Case 6: Schedule Change — Non-Competing Additions**

569 **Task type:** schedule_change **Depth:** 1

570 **Scenario:** Monday 2pm meeting \rightarrow Tuesday 3pm meeting (cancellation + reschedule). **Query:**
571 “When is the meeting?” **Retrieved:** [Tuesday 3pm (rank 1)]. **SRR = 0.0, UA = 1.0.**

572 **Analysis:** Non-conflicting schedule additions are treated as separate memories. ConflictResolver
573 correctly identifies the cancellation; the Monday entry is not repelled (may still be relevant for
574 historical context).

575 **Case 7: Goal Modification — Clean Selective Forgetting**

576 **Task type:** goal_modification **Depth:** 1

577 **Scenario:** “Learn Spanish by December” \rightarrow “Learn French by December.” **Query:** “What is the
578 user’s language learning goal?” **Retrieved:** [French (rank 1)]. **SRR = 0.0, UA = 1.0.**

579 **Analysis:** Explicit goal modification clearly flagged by ConflictResolver. Gate correctly classifies
580 Spanish as repel, French as keep. Contrastive loss separates the two goal representations within a
581 single macro-sleep cycle.

582 **Case 8: Multi-hop Query — Collateral Preservation**

583 **Task type:** fact_update **Depth:** 2

584 **Scenario:** “Alice works at Google” → “Alice works at DeepMind.” **Query (direct):** “Where does
585 Alice work?” **Query (multi-hop):** “What company does Alice’s former manager work at?” **Re-**
586 **trieved (direct):** [DeepMind (rank 1)]. **SRR = 0.0, UA = 1.0. Retrieved (multi-hop):** [DeepMind
587 (rank 1), Google (rank 6)]. **SRR = 0.1, UA = 0.9, UFR = 10.2%.**

588 **Analysis:** Multi-hop queries requiring inference across the revision chain are harder than direct
589 recall. Minor collateral leakage on intermediate organizational links is within acceptable margin;
590 the primary question is always answered correctly.

591 E Implementation Details and Reproducibility

592 E.1 Baseline Configurations

593 **Mem0 (tuned):** We tune the keyword-to-semantic weight ratio via grid search on a held-out val-
594 idation split of 500 ConflictStream sessions. The optimal ratio is 0.3 keyword / 0.7 semantic (vs.
595 default 0.5/0.5). All other Mem0 parameters are kept at defaults.

596 **NLI reranking, Delete-only, Heuristic suppression:** All three targeted baselines use the same
597 ConflictResolver as MORPHEUS (Qwen3-8B, top-5 retrieval, same prompt template) to ensure fair
598 comparison. NLI reranking applies the resolver at query time and demotes flagged entries; Delete-
599 only permanently removes flagged entries; Heuristic suppression applies recency + conflict-flag
600 reweighting without latent-space training.

601 E.2 ConflictResolver Prompt and Validation

602 **Full prompt.** The ConflictResolver uses the following prompt (Qwen3-8B, temperature 0,
603 max 200 tokens):

```
604      System: You are a factual consistency checker for an AI
605      memory system.
606      User: Given the following two memory entries, determine
607      whether the new
608      entry contradicts or supersedes the old entry.
609
610      Old entry: {old_memory}
611      New entry: {new_memory}
612
613      Answer with exactly one word: yes (the new entry contradicts
614      or supersedes
615      the old), no (they are compatible or about different topics),
616      or partial
617      (the new entry partially updates the old).
```

618 **Model configuration.** Qwen3-8B, temperature 0, top-p 1.0, max_new_tokens 10. The backbone
619 model is the same family as the ConflictResolver; cross-family replication (GPT-4o-mini) is pending
620 for camera-ready.

621 **Top- K retrieval before judging.** For each new memory m_{new} , the top-5 most similar existing
622 memories are retrieved by cosine similarity in the Active Index. Each of the 5 candidate pairs
623 $(m_{\text{new}}, m_{\text{old}})$ is scored independently.

624 **Action mapping.** yes → trigger micro-sleep on this pair; partial → flag for compress tier
625 consideration; no → no action.

626 **Parser and fallback.** The response is lowercased and stripped. If the response is not in {yes, no,
627 partial}, the fallback is no (conservative: do not trigger micro-sleep on ambiguous responses).

628 **Validation.** On a held-out ConflictStream validation split (500 sessions, disjoint from training and
629 test by template family), the resolver achieves false-positive rate 4.1% and false-negative rate 6.3%.

These are measured against template-derived ground-truth conflict labels using a binary protocol: `yes` and `partial` are both treated as positive (conflict detected); `no` is treated as negative. The `partial` response is not separately evaluated in the binary FPR/FNR calculation because its system action (flag for compress tier) does not trigger micro-sleep; it is therefore not a false positive for the micro-sleep trigger. Approximately 2.1% of outputs required fallback parsing (response not in the expected vocabulary); all were mapped to `no`.

E.3 Compute Resources

All experiments were run on NVIDIA A100 GPUs unless otherwise noted. Training was distributed across $4 \times \text{A100 80GB}$ nodes using PyTorch `accelerate` with Fully Sharded Data Parallel (FSDP). Single-node evaluation (ConflictStream, ablation runs) was run on $1 \times \text{A100 40GB}$.

Total compute budget: approximately 2,400 GPU-hours for the full evaluation suite, including the 10-variant ablation study, theorem validation experiments, and reproducibility runs. Phase 1 (codec training) required approximately 800 GPU-hours; Phase 4 (ConflictStream) required approximately 400 GPU-hours; Phase 5 (ablation suite) required approximately 1,200 GPU-hours.

Training details:

- **Stage 1 (teacher-forcing):** 50,000 steps, $4 \times \text{A100 80GB}$, batch size 16 per GPU, learning rate 1×10^{-4} with cosine annealing, warmup 2,000 steps, BF16 mixed precision.
- **Stage 2 (progressive substitution):** 20,000 steps, same hardware configuration, substitution rates $0.25 \rightarrow 0.50 \rightarrow 0.75$.

E.4 Backbone and LoRA Configuration

The model backbone is an 8B-parameter instruction-tuned LLM (anonymized for double-blind review). LoRA adapters are applied to the following modules using the `peft` library (version ≥ 0.18):

- **LoRA rank** $r = 32$
- **Target modules:** `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
- **LoRA alpha:** 64 ($2 \times \text{rank}$, following standard LoRA practice [16])
- **Dropout:** 0.05
- **Initialization:** Default `peft` initialization (Gaussian $\sigma = 0.01$)
- **Optimizer:** AdamW, $\beta = (0.9, 0.999)$, weight decay 0.01

E.5 Latent Codec Architecture and Quality Metrics

The MORPHEUS codec encodes episodic memories into a 32-token latent representation with dimension 4096 per token (matching the backbone hidden size; total latent vector: $32 \times 4096 = 131,072$ parameters per memory). The codec shares its backbone with the retrieval encoder and the language model decoder via multi-task LoRA adapters.

- **Encoder:** 8B backbone with LoRA rank 32
- **Latent space:** 32 tokens \times 4096 dims, BF16
- **Decoder:** Same LoRA adapters as encoder (weight sharing)
- **Training objective:** Teacher-forcing ROUGE-L loss in Stage 1; progressive latent substitution in Stage 2
- **Normalization:** Latent vectors are ℓ_2 -normalized at inference to the unit hypersphere for cosine similarity computation

Codec quality metrics (held-out evaluation set, Stage 2 checkpoint):

- **Reconstruction (ROUGE-L):** 0.82 on held-out passages (teacher-forcing with latent substitution rate 0.75)

- **Retrieval consistency (Recall@10):** 0.78 for same-topic retrieval from a curated 500-entry seed set
- **Storage compression:** BF16 latent (262 KB) compresses to NF4 (65 KB, 4× reduction) with <2% downstream SRR degradation

These metrics bound downstream performance: the codec quality limitation is documented in the Limitations section of the main text.

E.6 Loss Weights and $\mathcal{L}_{\text{MORPHEUS}}$

The complete MORPHEUS training objective is:

$$\mathcal{L}_{\text{MORPHEUS}} = \lambda_c \mathcal{L}_{\text{contrastive}} + \lambda_r \mathcal{L}_{\text{recon}} + \lambda_h \mathcal{L}_{\text{homeostasis}} + \lambda_g \mathcal{L}_{\text{gate}} + \lambda_m \mathcal{L}_{\text{revision}}.$$

The default weight vector is:

$$(\lambda_c, \lambda_r, \lambda_h, \lambda_g, \lambda_m) = (1.0, 0.5, 0.1, 0.3, 0.2).$$

These weights are fixed across all main experiments unless explicitly varied in Appendix C. The contrastive loss weight $\lambda_c = 1.0$ is the reference; all other weights are normalized relative to it.

E.7 Random Seeds and Reproducibility

The following seed-setting procedure is applied at the start of every training run and evaluation script:

```
import random, numpy as np, torch
def set_seed(seed):
    random.seed(seed); np.random.seed(seed)
    torch.manual_seed(seed); torch.cuda.manual_seed_all(seed)
    torch.backends.cudnn.deterministic = True
    torch.backends.cudnn.benchmark = False
```

Seed assignments by table:

- **Table 1 (ConflictStream main results):** 5 seeds {41, 42, 43, 44, 45}. Mean \pm std reported.
- **Table 2 (Mechanistic Diagnostics):** 3 seeds {41, 42, 43}. Mean \pm std reported.
- **Table 3 (OOD Validation):** 3 seeds {41, 42, 43}. Results pending for camera-ready.
- **Agent benchmarks (Appendix C.4):** 5 seeds {41–45}. Mean \pm std reported.
- **LongMemEval/LoCoMo (Appendix C.6):** Single seed 42. Multi-seed runs planned for camera-ready.
- **Proposition 1 sweep (Figure 4):** 5 seeds {41–45} per γ value.
- **Codec training (Stage 1/2):** Single seed 42.

For distributed training, `accelerate` is launched with `--seed $SEED` to ensure all worker processes initialize identically. For figure generation scripts, the same seed is used for t-SNE and UMAP projections (`random_state=42`) to ensure deterministic output.

E.8 Evaluation Hyperparameters

Table 15 summarizes the key hyperparameters for each evaluation benchmark.

Additional implementation notes:

- **Micro-sleep trigger:** conflict ratio $r_c > 0.15$ within a session, where $r_c = |\{m_{\text{new}} : \text{ConflictResolver}(m_{\text{new}}, \cdot) = \text{yes}\}| / |\text{session memories}|$. Only the yes response (not partial) counts toward r_c . The main experiments do not use a stale-rate trigger; r_c is the sole event-driven micro-sleep condition.

Table 15: Key evaluation hyperparameters. All values are fixed at these settings across all reported experiments.

Benchmark	n_{sessions}	Sleep Cadence	K	Margin γ
ConflictStream	5,000	10	10	0.5
ALFWorld	134	20	10	0.5
LongMemEval	500	50	10	0.5
Proposition 1 sweep	200 per γ	10	10	$\in [0.1, 0.9]$
Ablation suite	1,000	10	10	0.5

- **Macro-sleep trigger:** every $N = 10$ interactions (configurable)
- **FAISS index:** IndexHNSWFlat with $M = 32$, $efSearch = 128$
- **Qdrant archive:** distance metric = cosine, shard number = 1 (single-node)
- **Value scorer (main experiments):** heuristic combination of access frequency (0.4), recency decay (0.3), and semantic novelty via HNSW top- K lookup (0.3); no LLM calls. An optional LLM-as-judge variant (backbone model, temperature 0.1, max tokens 64) is available for offline batch settings but was *not* used in the main reported results.
- **Contrastive temperature** $\tau = 0.07$ (selected via ablation)
- **Configuration:** All hyperparameters managed via Hydra; full configs provided in the code release

F Public Revision Trace Evaluation

F.1 Dataset Construction

WikiRevisions is a 150-session dataset constructed from Wikipedia infobox revision histories. Entity domains: CEOs of Fortune 500 companies, elected officials, and film directors; revision period 2018–2024.

Data source. We use the Wikipedia Revision History API (`action=query&prop=revisions`) to retrieve all infobox revisions for each entity page. Infobox fields are extracted using the `mwparsersfromhell` library; only structured infobox key-value pairs are retained (free-text sections are excluded).

Field normalization. CEO/officeholder/director fields are mapped to a unified query template: “Who is the current {role} of {entity}?” for person-attribute updates, and “What is the current {field} of {entity}?” for other fields. Field names are normalized to lowercase with underscores replaced by spaces.

Snapshot alignment. Each session contains 3–8 memory entries derived from intermediate Wikipedia snapshots. If multiple revisions occur on the same day, only the last revision of that day is retained as a snapshot. The query date is set to 30 days after the most recent revision in the session, ensuring the query always falls after the latest update.

F.2 Filtering and Query Construction

We filter to single-field edits (only one infobox attribute changes per revision) with ≥ 6 months between the old and new snapshot. Current-status queries are generated from the normalized infobox field name. Gold labels are derived from the Wikipedia snapshot timestamps: the entry from the most recent snapshot before the query date is *current*; earlier entries are *superseded* for current-state queries (following the same query-conditioned label semantics as ConflictStream; see Appendix A.7).

F.3 Split and Leakage Control

Sessions are split by entity: no entity appears in more than one split. Temporal leakage is controlled by ensuring the query date is always after the most recent revision in the session.

749 **F.4 Evaluation Caveats**

750 WikiRevisions is a small preliminary probe ($n = 150$), dominated by person-attribute updates. It
 751 is not a comprehensive real-world benchmark. Broader real-world validation across diverse entity
 752 types and revision patterns remains future work.

Table 16: Results on WikiRevisions (150 sessions, 3 seeds). Preliminary probe; see caveats in Section F.

Method	SRR (\downarrow)	UA (\uparrow)	UFR (\downarrow)
Naive	[TBD]	[TBD]	[TBD]
Heuristic suppression	[TBD]	[TBD]	[TBD]
MORPHEUS	[TBD]	[TBD]	[TBD]

753 Results are pending for camera-ready.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and Section 1 state three contributions (framework, benchmark, empirical analysis with idealized bound) that are all substantiated in the paper body. Limitations are discussed in Section 8.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Section 8 discusses limitations including codec quality dependency, value scorer generalization, ConflictStream scope (synthetic benchmark), and the secondary nature of agent-task evidence. Section 5 analyzes four failure modes.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Proposition 1 is explicitly framed as an idealized phenomenological bound, not a tight theorem. Key assumptions A1–A3 are stated inline with explicit acknowledgment that A2 is a working approximation. A derivation under these assumptions appears in Appendix B. We use \lesssim notation to reflect the non-tight nature of the bound.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Appendix E provides full hyperparameters, loss weights, LoRA configuration, random seeds, compute budget, and evaluation settings. ConflictStream generation is fully specified.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Code, ConflictStream generator, trained checkpoints, and evaluation scripts will be released under Apache 2.0 upon acceptance. An anonymized repository is available for review.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Section 4 and Appendix E specify hardware, optimizer (AdamW), learning rates, batch sizes, LoRA rank, contrastive temperature, and all evaluation hyperparameters (Table 15).

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#) (partial — see justification)

Justification: Tables 1 and 6 report mean \pm std over 5 seeds with paired significance tests (Wilcoxon $p < 0.01$, paired t -test $p < 0.05$). Table 2 reports mean \pm std over 3 seeds. Table 8 uses a single seed (42); multi-seed runs are planned for camera-ready. All tables explicitly state their seed count in captions.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix E.3 specifies GPU type (A100 40GB/80GB), total compute budget (2,400 GPU-hours), and per-phase breakdown.

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This work develops a memory management framework for LLM agents. It does not involve human subjects, private data, or dual-use concerns beyond standard ML research.

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: MORPHEUS is a foundational memory management technique for LLM agents. It does not directly enable harmful applications beyond what LLMs already support. Selective forgetting could positively impact privacy (removing outdated personal information) but could also be misused to selectively erase accountability records.

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: MORPHEUS releases LoRA adapter weights and a synthetic benchmark generator. Neither poses significant misuse risk beyond the base LLM.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All baseline systems (Mem0, FAISS, Qdrant) and evaluation benchmarks (ALFWorld, ScienceWorld, LongMemEval, LoCoMo) are cited. The backbone model is used under its original license.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: ConflictStream benchmark is fully documented (6 task types, 5 query types, generation algorithm). Code release includes README, configuration files, and reproduction scripts.

14. Crowdsourcing and research with human subjects

852 Question: For crowdsourcing experiments and research with human subjects, does the pa-
853 per include the full text of instructions given to participants and screenshots, if applicable,
854 as well as details about compensation (if any)?
855 Answer: [N/A]
856 Justification: This work does not involve crowdsourcing or human subjects research.

857 **15. Institutional review board (IRB) approvals or equivalent for research with human**
858 **subjects**

859 Question: Does the paper describe potential risks incurred by study participants, whether
860 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
861 approvals (or an equivalent approval/review based on the requirements of your country or
862 institution) were obtained?
863 Answer: [N/A]
864 Justification: No human subjects research is conducted in this work.

865 **16. Declaration of LLM usage**

866 Question: Does the paper describe the usage of LLMs if it is an important, original, or
867 non-standard component of the core methods in this research?
868 Answer: [Yes]
869 Justification: MORPHEUS uses an LLM backbone as both the latent codec and the LLM-
870 as-judge conflict resolver (Section 3). The backbone identity is anonymized for review but
871 fully specified in Appendix E.