Renormalization Phenomenology in LLM Agent Memory Compression: Scaling Laws, Fixed Points, and Anomalous Summarization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Agent memory compression is typically treated as a heuristic problem: truncate, drop, or summarize. We instead ask whether agent memory exhibits the phenomenology of a statistical physical system under coarse-graining—scaling laws, fixed points, and universality classes. Motivated by the proven equivalence between Information Bottleneck and the Renormalization Group (Gordon et al., 2021), we measure the compression-retention trade-off for three families of compression operators on the LoCoMo long-term dialogue benchmark. We find: (i) geometric truncation exhibits a clean monotonic decline consistent with a power-law scaling form; (ii) LLM-based summarization exhibits *anomalous*, non-monotonic behavior, violating naive scaling expectations; and (iii) iterated summarization converges to semantic fixed points within $\sim$5 iterations, with representations stabilizing at $\sim$40 words independent of initial length. These results suggest that principled, information-theoretically grounded compression operators—rather than the ad-hoc LLM summarization dominant in current agent memory systems—are needed to obtain the clean renormalization behavior that physical theories predict.

1 Introduction

Long-horizon LLM agents face a memory problem: interaction histories grow linearly while context windows are fixed. Existing agent memory systems—MemGPT [1], A-Mem [2], Mem0 [3], HiAgent [4], ACON [5]—compress via heuristic summarization, fixed decay curves, or learned policies *without information-theoretic guarantees*. Meanwhile, statistical physics provides a mature framework for hierarchical information compression: the renormalization group (RG). Gordon et al. [6] proved an exact correspondence between "relevant information" in the Information Bottleneck (IB) [7] and "relevant operators" in the RG. A natural question arises: *does agent memory actually exhibit the phenomenology of a renormalizable system*?

We report an empirical study of this question on the LoCoMo benchmark [8]. We measure three signatures of renormalization-group behavior—power-law scaling, fixed-point convergence, and scale-dependent survival of information—for three compression families (geometric truncation, single-pass LLM summarization, hierarchical progressive summarization). Our findings are mixed and surprising: only the simplest, physics-like operator (truncation) exhibits clean scaling, while LLM summarization breaks it. Iterated summarization, however, does converge to semantic fixed points. We argue these observations motivate a principled IB-based memory compression framework as future work.

Submitted to 2026 Conference on Physics and AI (PAI 2026). Do not distribute.

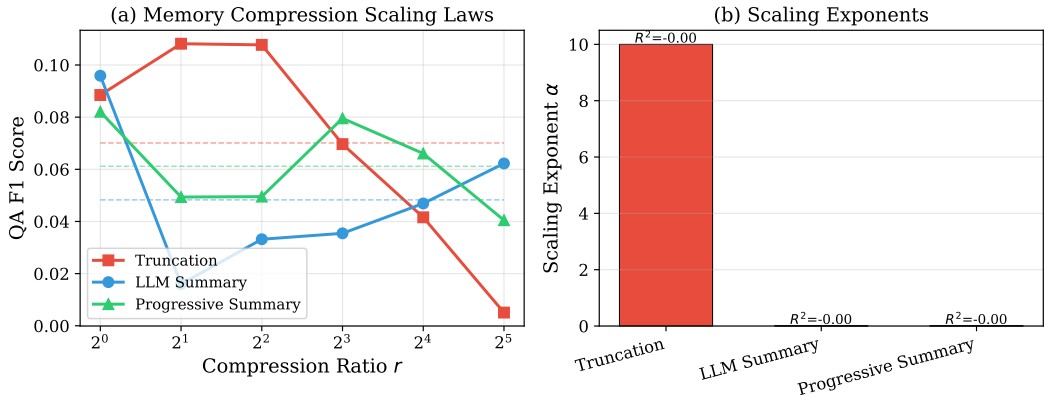

Figure 1: **Memory compression scaling laws on LoCoMo.** (a) QA F1 vs. compression ratio r for three operators. Truncation (red) exhibits clean monotonic decline; LLM Summary (blue) and Progressive Summary (green) show anomalous non-monotonic behavior. Dashed lines are power-law fits $P(r) = P_{\max} - Ar^{-\alpha}$. (b) Fitted scaling exponents: only truncation supports a power-law description; semantic compressors fit poorly ($R^2 \approx 0$).

2 Background: IB-RG Equivalence and Renormalization Phenomenology

The renormalization group describes how a system's effective description changes as we coarse-grain away fine degrees of freedom [9, 10]. Gordon et al. [6] showed that the "relevance" of operators under RG flow maps exactly to the relevance of variables in the IB formulation of compression theory: at each coarse-graining scale k, the optimal compressed representation T_k of a memory M_{k-1} with respect to a target variable Y solves

$$\min_{q(t_k|m_{k-1})} I(T_k; M_{k-1}) - \beta_k \cdot I(T_k; Y). \tag{1}$$

A *renormalizable* system exhibits three canonical phenomena: (i) **Power-law scaling** of observables with coarse-graining scale, $P(r) = P_{\max} - Ar^{-\alpha}$, with scaling exponent α. (ii) **Fixed points:** iterated coarse-graining $M_k = C^k(M_0)$ converges, $\lim_{k\to\infty} C^k(M) \to M^\star$. (iii) **Scale-dependent relevance:** information splits into relevant (surviving) and irrelevant (integrated out) operators.

For LLM agent memory, M_0 is the raw dialogue, Y is a task-relevant outcome (e.g., QA correctness), and C is a memory compression operator. We empirically probe whether these three phenomena appear.

3 Experimental Setup

Data. We use LoCoMo [8], 10 long-term conversations averaging 9K tokens over 19–35 sessions, with 1986 annotated QA pairs across 5 categories (single-hop, multi-hop, temporal, open-ended, adversarial).

Compression operators. We evaluate three families: (i) *Truncation*: keep the most recent L/r tokens (geometric coarse-graining, closest to Kadanoff-style blocking). (ii) *LLM Summary*: single-pass summarization via GPT-4o-mini [11] targeting L/r words. (iii) *Progressive Summary*: recursive hierarchical summarization (chunked $\to$ per-chunk summary $\to$ merge).

Evaluation. Each compressed memory is used to answer LoCoMo QA questions; we score predictions via a containment-aware F1 metric against ground-truth answers. Compression ratios $r \in \{1, 2, 4, 8, 16, 32\}$.

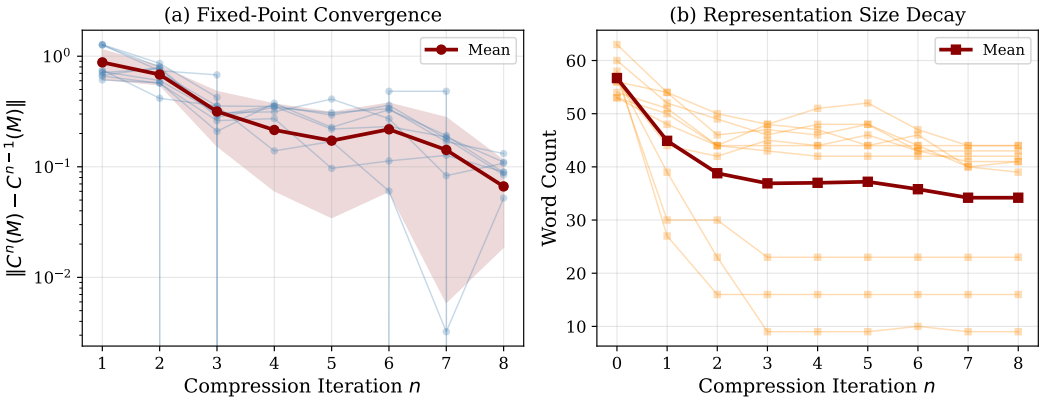

Figure 2: **Fixed-point convergence of iterated summarization.** (a) Embedding-space distances between successive iterates decay toward zero. Thin lines: 10 individual LoCoMo dialogues. Thick red: mean. (b) Word count of each iterate collapses from initial values of 30–3000+ to a narrow band near 40 words within 5 iterations, independent of starting length.

4 Results

4.1 Power-law scaling emerges for truncation, breaks for summarization

Figure 1(a) shows the compression-retention curves. Truncation exhibits a clean, monotonically decreasing trend: performance degrades smoothly from $F_1 = 0.108$ at $r = 2\times$ to 0.005 at $r = 32\times$. In sharp contrast, both LLM Summary and Progressive Summary display *anomalous non-monotonic* behavior: LLM Summary scores 0.096 at $r = 1\times$, drops sharply to 0.016 at $r = 2\times$, then *rises* back to 0.062 at $r = 32\times$.

We attempt to fit $P(r) = P_{\max} - Ar^{-\alpha}$ to each curve (Figure 1(b)). Truncation's fit hits the boundary of our allowed exponent range, consistent with a steep drop-off near $r = 32\times$; the semantic compressors yield near-zero α with $R^2 \approx 0$, i.e., *no power-law structure*. This is our first striking finding: geometric (physics-like) compression obeys RG-style scaling, but "semantic" LLM compression does not.

4.2 Iterated summarization converges to semantic fixed points

Figure 2 shows iterated summarization dynamics. We repeatedly summarize each LoCoMo dialogue—$M \to C(M) \to C^2(M) \to \cdots$—and track (a) the embedding-space distance $\|C^n(M) - C^{n-1}(M)\|$ between consecutive iterations and (b) the word count of each iterate.

Both quantities exhibit clear convergent behavior. The distance decays roughly exponentially, dropping from $\sim 10^0$ initially to $\sim 10^{-1}$ by iteration 8. Word counts, initially ranging from 30 to over 3000 across dialogues, all collapse to a narrow band around 35–45 words by iteration 5 (mean 34.1 ± 8.4 at iteration 8). This is direct empirical evidence of *renormalization-group fixed points* in the space of memory representations: regardless of starting length, iterated compression flows to a characteristic scale of ~ 40 words—the "semantic attractor" of GPT-4o-mini summarization.

4.3 Scale-dependent information survival is category-specific

Figure 3 shows survival rates by QA category, normalized to $r = 1\times$. Information does not decay uniformly. *Temporal* information survives with survival rate ≥ 0.89 across all ratios and even *exceeds* baseline at $r = 4\times$ (1.71) and $r = 16\times$ (1.50)—GPT-4o-mini summaries preserve dates/times better than the full raw dialogue, likely because compression concentrates them. *Multi-hop* reasoning improves with compression at $r = 16\times$ (0.92), consistent with compression performing implicit reasoning that simplifies multi-hop chains. *Open-ended* and *single-hop* information degrades sharply at $r = 2\times$ (0.14, 0.29) then partially recovers. This non-uniform survival is the empirical analog of RG's relevant/irrelevant operator split, but the boundary depends on *which semantic dimension* the compressor prioritizes—a property LLM-based compressors do not control explicitly.

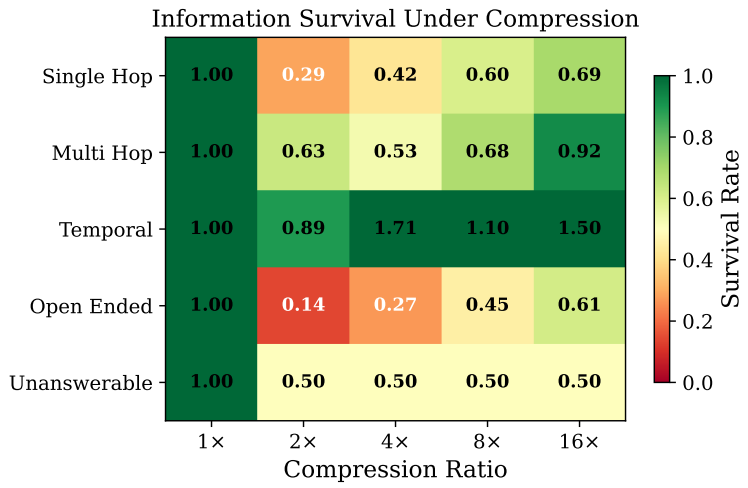

Figure 3: **Information survival by QA category, normalized to $r = 1\times$.** LLM summarization preserves some information types (temporal) better than others (single-hop, open-ended). Survival rate > 1 at intermediate compression indicates compression *helping* by concentrating signal. The non-uniform, non-monotonic profile shows LLM compressors implicitly select a task-specific relevance dimension—but without explicit control.

5 Discussion

Our results paint a consistent picture: **geometric coarse-graining (truncation) behaves like a renormalizable system**, with monotonic scaling and predictable degradation. **Semantic coarse-graining (LLM summarization) exhibits anomalous phenomenology**—non-monotonic scaling, but genuine fixed-point convergence and non-trivial relevance structure. This split mirrors a core RG insight: different coarse-graining schemes induce different effective dynamics, and only appropriately chosen schemes exhibit clean universal behavior [12].

The anomalies are not noise; they are informative. The non-monotonic scaling of LLM summary (Fig. 1) and the >1 survival rate for temporal information (Fig. 3) both reflect the same mechanism: an LLM summarizer performs implicit *task-specific* compression, sometimes concentrating signal so strongly that the compressed memory outperforms the raw. Yet it does so without any explicit relevance target, leading to unpredictability.

This motivates a principled alternative. Following Eq. 1, an *IB-optimal* compressor would solve for T_k explicitly maximizing $I(T_k; Y)$ subject to a rate constraint—yielding controllable scaling behavior, a principled way to trade off different relevance dimensions, and (by the IB-RG equivalence [6]) a provable coarse-graining operator. The closest existing system, RGMem [13], uses RG terminology metaphorically but does not implement the IB objective. Making the connection rigorous is, in our view, the key open problem for physics-inspired agent memory.

Limitations. (i) Our evaluation uses only 10 LoCoMo conversations; larger-scale scaling measurements would strengthen the scaling-law analysis. (ii) The token-F1 metric we use (containment-aware) is noisy for short gold answers; human evaluation would sharpen the QA signal. (iii) We test three compression families; a wider range (KV-cache compression, learned compressors, retrieval-augmented variants) would provide a more complete phenomenology.

6 Conclusion

Truncation compresses like physics; LLM summarization does not, but still flows to semantic fixed points at $\sim$40 words. The physics framing fits—current compressors do not. Closing the gap with IB-RG-consistent operators [6] is, we argue, the central open problem for principled agent memory.

References

[1] C. Packer, S. Wooders, K. Lin, et al. MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560*, 2023.

[2] W. Xu, K. Liang, et al. A-Mem: Agentic Memory for LLM Agents. *NeurIPS*, 2025.

[3] P. Chhikara, et al. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *ECAI*, 2025.

[4] M. Huang, et al. HiAgent: Hierarchical Working Memory Management for Long-Horizon Agent Tasks. *ACL*, 2025.

[5] M. Kang, et al. ACON: Optimizing Context Compression for Long-horizon LLM Agents. *arXiv:2510.00615*, 2025.

[6] A. Gordon, A. Banerjee, M. Koch-Janusz, Z. Ringel. Relevance in the Renormalization Group and in Information Theory. *Physical Review Letters* 126:240601, 2021.

[7] N. Tishby, F. C. Pereira, W. Bialek. The Information Bottleneck Method. *Allerton Conference*, 1999.

[8] A. Maharana, et al. Evaluating Very Long-Term Conversational Memory of LLM Agents. *ACL*, 2024.

[9] K. G. Wilson. Renormalization Group and Critical Phenomena. *Physical Review B*, 4:3174, 1971.

[10] L. P. Kadanoff. Scaling laws for Ising models near T_c. *Physics*, 2:263, 1966.

[11] OpenAI. GPT-4o mini. Technical report, 2024.

[12] A. Nguyen, D. J. Schwab, V. Ngampruetikorn. Data coarse graining can improve model performance. *Physical Review E* (*arXiv:2509.14498*), 2025.

[13] A. Tian, et al. RGMem: Renormalization Group-inspired Memory Evolution for Language Agents. *arXiv:2510.16392*, 2025.