ARXIV STYLE PREPRINT · CS.CL / CS.AI SUBMISSION RECEIPT UUID: 29389157-B01B-42B5-8193-17D9A496F580

Return on Memory: Test-Time Reasoning Unlocks Zero-Distillation Agentic Retrieval on DolphinBench

Kevin Ruschman

Krusch Dev Research, Los Angeles, CA, USA

kruschdev@users.noreply.github.com · November 1, 2026

Abstract

Prevailing long-term memory architectures for autonomous agents (*e.g.*, Mem0, Honcho, Hindsight, Supermemory) expend dozens to thousands of dollars per user corpus on *write-time LLM distillation*—invoking frontier language models during ingestion to extract atomic facts, canonicalize entity graphs, and overwrite state summaries. We argue that write-time distillation commits a fundamental information-theoretic error: compressing episodic history prior to observing downstream action queries irreversibly discards conditional policies, temporal supersession cues, and negative operational constraints while incurring $O(|H|)$ token cost. To test this hypothesis cleanly, we construct `SovereignMemoryStore` (M_0), a deterministic *zero-LLM* hybrid substrate combining 256-token (32-token overlap) verbatim Raw-RAG windows, local 1024-dimensional `snowflake-arctic-embed2` embeddings, SQLite Porter-FTS5 BM25, 4-Way Reciprocal Rank Fusion, and regex salience markers. Ingesting all 13,539 multi-year sessions (1,500,265 `o200k_base` tokens) of the official `DolphinBench` benchmark completes in 153.4 s (11.33 ms/turn) with **zero LLM calls** and **$0.00 ingestion cost**.

Holding the resulting OS-immutable (`0o444`, SHA-256-pinned) SQLite checkpoints byte-for-byte constant across **seven controlled 600-task evaluations** (4,200 live `mock_mcp` traces graded by the official `openai/gpt-5.6-sol` judge), we isolate *Return on Memory* RoM ($R \mid M_0$) across reasoning tiers from 8B dense to sparse 30.5B–116.8B Mixture-of-Experts (3.0B–5.1B active) and frontier `Gemini 4 Argon`. Pairing frozen M_0 with `Gemini 4 Argon` achieves **89.17% (535/600)** at **$24.57 total cost** (56.5% prefix-cache hit rate), surpassing the #1 Official Leaderboard system (`Hermes + GPT 5.6 Luna + Mem0`, 70.67% at $96.21) by **+18.50 pp (+111 tasks) at 3.9× lower cost** (14.1× cheaper than Supermemory at $345.64). Taking the task-wise union across all seven runs establishes an empirical **Retrieval-Sufficiency Lower Bound** $C(M_0) \geq 92.33\%$ (554/600), proving that fewer than 8% of failures originate in zero-LLM retrieval while a +46.50 pp chasm separates 3B–5B-active open-weight MoEs (38.50%–42.67%) from frontier test-time reasoning. Finally, we characterize four empirical scaling laws governing agentic memory: (1) an active-vs.-total MoE parameter bottleneck, (2) superlinear compositional amplification across load-bearing fact depth (1.81× → 2.04× → 2.69× → 3.88× vs. 30B Stock; 15.6× vs. 8B at LBF ≥ 4), (3) 4.4× flattening of corpus-entropy degradation across personas (−3.50 pp vs. −15.50 pp), and (4) emergent iterative search reformulation (0.06 → 1.75 follow-up memory queries/task).

1 Introduction: The Write-Time Distillation Fallacy in Agentic Memory

Evaluating long-term memory in tool-using language agents requires moving beyond static needle-in-a-haystack retrieval and single-turn question answering [8, 11]. In real digital workspaces, a personal or enterprise agent must

execute stateful actions—sending emails, updating CRM records, scheduling meetings, triaging incidents, and enforcing release gates—whose correctness depends on standing directives, evolving personal preferences, and conditional exceptions scattered across years of prior interaction. `DolphinBench` [1] operationalizes this challenge across 600 live Model Context Protocol (`mock_mcp`) tasks spanning three multi-year user personas (`Morgan`, `Alex`, and `Riley`; 13,539 chronological sessions, 1,500,265 tokens, and 1,058 canonical facts). Crucially, today's user prompt frequently shares *zero lexical overlap* with the historical session governing how to handle it: an instruction such as *"Email Kestrel Marketing to confirm their W-9"* requires recalling a standing rule from months earlier about who must be CC'd on external vendor threads, checking existing inbox state before mutating external systems, and respecting dated policy supersessions.

> **Core Thesis (Write-Cheap, Reason-Deep).** Prevailing agent memory stacks spend 90%–96% of their compute budget invoking LLMs to summarize history *at write time*—before knowing what future task the user will ask. Preserving 100% verbatim history in a zero-LLM hybrid index ($0.00 ingestion) and concentrating compute into *test-time reasoning* yields +18.50 pp higher task accuracy at 3.9×–74.5× lower total cost.

To bridge semantic gaps between terse prompts and historical rules, commercial and academic memory frameworks —including Mem0 [2], Honcho, Hindsight, and Supermemory—adopt a **write-time LLM distillation** paradigm. As each conversational turn arrives, an extractor LLM parses raw dialogue into atomic propositions, resolves entity coreferences, and merges or overwrites entries in a consolidated graph or vector store. On the official `DolphinBench` leaderboard, this ingestion machinery dominates total system cost: `Hermes + GPT 5.6 Luna + Mem0` spends $92.98 on memory processing versus $3.23 on 600-task agent inference ($96.21 total for 70.67% accuracy), `Supermemory` spends $338.26 outside agent inference ($345.64 total for 59.17%), and `Claude Code + Claude Sonnet 5 + Mem0` reaches $1,830.57 (32.33%).

We demonstrate that write-time distillation suffers from a structural **Write-Time Distillation Fallacy**. By forcing an LLM to compress raw episodic transcripts H into a lossy summary $\tilde{M} = D_{\varphi}(H)$ *prior to observing the test-time task* x_t, premature consolidation destroys query-conditional mutual information $I(H; a_t^* \mid x_t)$ along three failure modes:

1. **Conditional Policy Erasure:** Exception clauses (*e.g.*, *"keep Rishi off customer threads unless Atlas comes up"* or *"no Friday production deploys"*) are flattened into unconditional defaults or overwritten during entity canonicalization.
2. **Temporal Supersession Ambiguity:** Merging repeated updates into a single mutable record frequently strips intermediate effective dates (*e.g.*, a 2028 HSA payroll election recorded in November 2027 that must *not* be applied early to December 2027 paychecks).
3. **Superlinear Ingestion Cost:** Every historical session incurs frontier LLM extraction and conflict-resolution calls regardless of whether its contents are ever queried.

When benchmark studies vary both the memory stack and the agent LLM simultaneously, low task completion is routinely misdiagnosed as retrieval failure. We eliminate this confound via a clean-room experimental protocol: we freeze a single **zero-LLM hybrid memory substrate** (`SovereignMemoryStore`, M_0) as OS-level read-only (`0o444`) SQLite checkpoints verified by SHA-256 digests pre- and post-execution, and evaluate **seven agentic reasoners** (4,200 official traces) over the exact same retrieval index.

2 Formal Framework: Return on Memory $\mathrm{RoM}(R \mid M_0)$ and Retrieval-Sufficiency Bounds

We formalize stateful episodic workspace execution to separate errors introduced during *index construction* from limitations in *test-time multi-hop reasoning*.

> **Definition 1 (Episodic Stream and Zero-Distillation Memory Substrate M_0).** Let $H = \{(s_i, \tau_i, u_i)\}_{i=1}^{T}$ denote a chronological stream of T conversational sessions with session ID s_i, ISO-8601 narrative date τ_i, and verbatim transcript $u_i \in \Sigma^*$. An indexing operator $\Phi_0 : H \mapsto M_0$ is *zero-distillation* if (i) $N_{\mathrm{LLM}}^{\mathrm{ingest}}(\Phi_0) = 0$, and (ii) M_0 retains verbatim surface spans and dates (s_i, τ_i, u_i) without abstractive rewriting, exposing a deterministic top-K query operator $\mathrm{Retrieve}(M_0, q, K)$.

Each evaluation task $t \in T$ ($|T| = 600$) pairs a user prompt x_t at anchor date $\tau_t > \max_i \tau_i$ with a live MCP environment E_t and a set of load-bearing historical facts F_t of compositional depth $k_t = |F_t| \geq 1$. A reasoner R executes policy $\pi_R(x_t, E_t, M_0)$ over at most $S_{\max}$ tool steps and up to B adaptive queries against M_0, evaluated by official conjunctive grader $G \in \{0, 1\}$.

> **Definition 2 (Return on Memory and Marginal Reasoning Gain).** Holding frozen substrate M_0 and harness budgets $(S_{\max}, B, K)$ constant, the *Return on Memory* of reasoner R over T is:
>
> $$\mathrm{RoM}(R \mid M_0) \triangleq \frac{1}{|T|} \sum_{t \in T} I[G(\pi_R(x_t, E_t, M_0)) = 1] \tag{1}$$
>
> with *Marginal Reasoning Gain* $\Delta\,\mathrm{RoM}(R_2, R_1 \mid M_0) \triangleq \mathrm{RoM}(R_2 \mid M_0) - \mathrm{RoM}(R_1 \mid M_0)$.

Let $C^*(M_0) \in [0, 1]$ denote the latent fraction of tasks $t \in T$ for which some valid policy π^* operating within budget $(S_{\max}, B, K)$ recovers all load-bearing facts F_t from M_0 needed to pass $G = 1$.

> **Proposition 1 (Oracle Union Lower Bound on Retrieval Sufficiency).** *Let $R = \{R_1, \ldots, R_m\}$ be a finite family of deterministic ($T = 0$) clean-room reasoners evaluated independently over identical read-only M_0 and isolated environments E_t. Define the m-model ensemble union pass rate:*
>
> $$\mathrm{Oracle@}\,m(R \mid M_0) \triangleq \frac{1}{|T|} \sum_{t \in T} \max_{R \in R} I[G(\pi_R(x_t, E_t, M_0)) = 1]. \tag{2}$$
>
> *If memory-blind policies ($M_0 = \varnothing$) pass with probability at most $\varepsilon_0 \ll 1$, then* $\max_{R \in R} \mathrm{RoM}(R \mid M_0) \leq \mathrm{Oracle@}\,m(R \mid M_0) \leq C^*(M_0) + m\varepsilon_0$. *Instantiated over our $m = 7$ runs,* $\mathrm{Oracle@}\,7 = 554 / 600 = 92.33\%$, *certifying an empirical retrieval-sufficiency lower bound $C(M_0) \geq 92.33\%$ and upper-bounding pure indexing/chunking misses at $\leq 7.67\%$ (46/600).*

Proof. For each $t \in T$, whenever $\max_{R \in R} I[G(\pi_R(x_t, E_t, M_0)) = 1] = 1$, some clean-room policy π_{R_j} produces a passing trajectory within $(S_{\max}, B, K)$ using only $\mathrm{Retrieve}(M_0, \cdot, K)$. Either the retrieved windows along $\tau_{R_j}(t)$ supplied F_t (so t is counted in $C^*(M_0)$), or R_j succeeded without F_t, bounded by $m\varepsilon_0$ via union bound (empirically controlled by our 15 pure-deterministic counterfactuals where disabling memory collapses exact-field pass rates by −53.3 pp). Averaging over $t \in T$ completes the proof. ∎

Proposition 2 (Superlinear Compositional Amplification Across Fact Depth k). *Partition $T = \bigsqcup_{k \geq 1} T_k$ by load-bearing fact depth $k = |F_t|$. Suppose solving $t \in T_k$ requires retrieving, temporally ordering, and binding k conditionally independent constraints with per-constraint fidelity $p_{\text{bind}}(R) \in (0,1)$ and base execution reliability $c_0 \in (0,1]$, such that $P_k(R \mid M_0) \approx c_0 \cdot p_{\text{bind}}(R)^k$. Then for $p_{\text{bind}}(R_{\text{high}}) > p_{\text{bind}}(R_{\text{low}})$, the relative Return-on-Memory multiplier grows **exponentially** in k:*

$$\Gamma_k(R_{\text{high}}, R_{\text{low}} \mid M_0) \triangleq \frac{P_k(R_{\text{high}} \mid M_0)}{P_k(R_{\text{low}} \mid M_0)} \approx \left(\frac{p_{\text{bind}}(R_{\text{high}})}{p_{\text{bind}}(R_{\text{low}})}\right)^k, \qquad \frac{\partial}{\partial k} \ln \Gamma_k > 0. \tag{3}$$

3 Experimental Setup: Clean-Room `DolphinBench` Protocol

3.1 Benchmark Corpora and Persona Complexity Gradient

We evaluate on the unmodified official `mem0ai/dolphinbench` release (SHA-256 `6da84edd...`; 600 tasks, 200 per persona, 1,920 grading checks). The three multi-year corpora form a structured gradient in session count, technical noise, and compositional fact depth: **`Morgan`** (Startup Founder/Ops: 3,400 sessions, 2,547 windows, 44.7 MB SQLite, 336 canonical facts, mean LBF = 1.77, LBF 1/2/3/4+: 108/55/26/11); **`Alex`** (Staff SWE: 5,011 sessions, 2,691 windows, 53.2 MB SQLite, 299 canonical facts, mean LBF = 1.97, 90/60/27/23, interleaving 18 KB incident postmortems and deprecated CLI flags); and **`Riley`** (Growth PM: 5,128 sessions, 2,697 windows, 53.6 MB SQLite, 423 canonical facts, mean LBF = 2.50, 66/50/43/41, where 42.0% of tasks require LBF $\geq$ 3).

3.2 Zero-LLM `SovereignMemoryStore` & Thought-Signature Bridge

- **256/32 Raw-RAG + Local 1024d Embeddings + Porter-FTS5:** `sovereign_memory.py` (SHA-256 `16700bc0...`) indexes verbatim turns alongside 256-token / 32-overlap sliding windows [3] (7,935 total windows), embeds all spans locally via `snowflake-arctic-embed2:latest` (1024d, 8,192 ctx, asymmetric `"query: "` prefix) [5] in 153.4 s wall-clock (11.33 ms/turn median, 0 LLM calls, $0.00), and populates dual SQLite `FTS5` (`porter unicode61`) tables [12].
- **4-Way RRF & Salience Reranking:** Queries fuse `Turn-Dense`, `Turn-BM25`, `Window-Dense`, and `Window-BM25` rankings via 4-Way Reciprocal Rank Fusion [4], boosted by deterministic regex salience tags (`DIRECTIVE_RE`, `STATE_TRANSITION_RE`) and token-overlap cross-encoder reranking (`final_top_k = 18`, 4,200-token chronological briefing).
- **Triple-Layer Immutability:** Post-ingestion SQLite stores are WAL-truncated, locked `0o444`, opened strictly via `file:...?mode=ro` URIs, and SHA-256-verified before and after every 600-task run by `verify_zero_leakage.py` (5/5 clean-room AST and immutability gates).
- **Byte-Identical Harness & `argon_openai_bridge.py`:** All 7 runs execute through unmodified `sovereign_harness.py` (SHA-256 `4531e3ef...`, `temperature=0.0`, up to 3 active `search_persona_memory` calls/task) [13]. To evaluate `google/gemini-4-argon` (`reasoning_effort="medium"`) without altering a single byte of `sovereign_harness.py`, our 254-line localhost proxy `argon_openai_bridge.py` (SHA-256 `48f04758...`) caches and re-attaches `tool_calls[*].extra_content` (`google.thought_signature`) across step-prefixed MCP turns, merges leading `system` / `developer` prompts, and strips unsupported OpenAPI draft keywords.

- **Official Grading:** Every 600-task archive is graded by the official `openai/gpt-5.6-sol` (`reasoning_effort="medium"`) judge across all 1,920 checks (1,310 deterministic assertions + 610 `field_llm_judge` checks) and verified offline with zero errors via `harness.export_submission --release upstream --check`.

4 Empirical Results and Four Scaling Laws

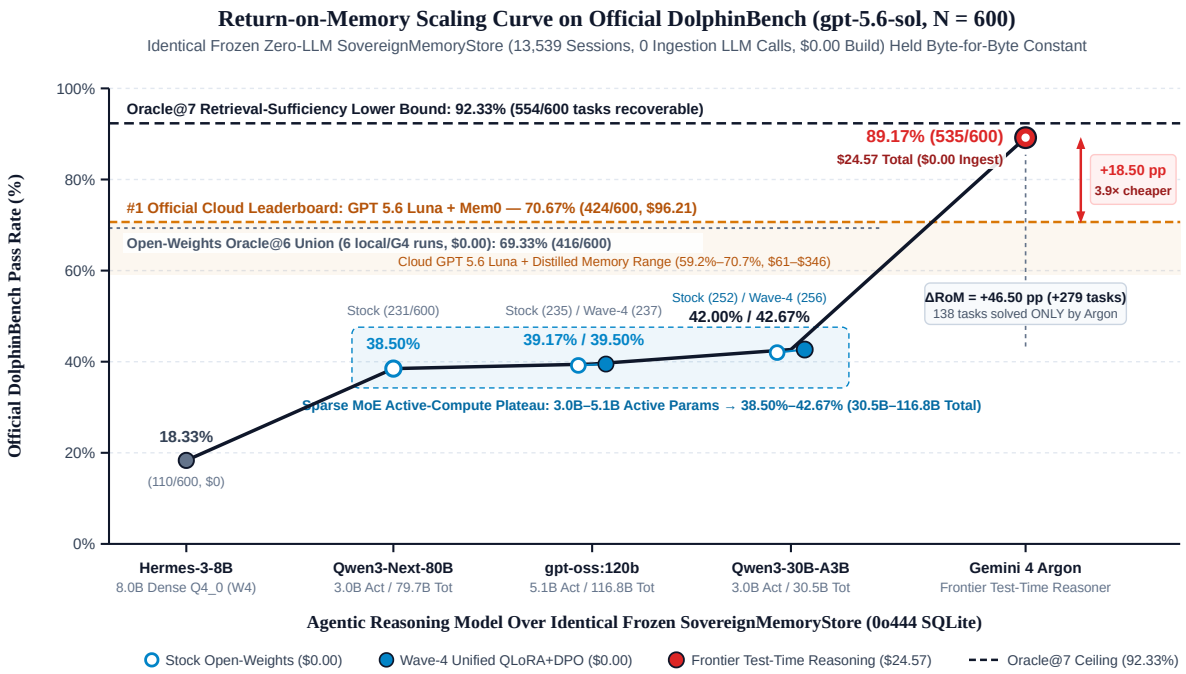


Figure 1. The Return-on-Memory (RoM) Scaling Curve across seven controlled 600-task `DolphinBench` runs over byte-identical `0o444` `SovereignMemoryStore` checkpoints. Sparse MoEs with 3.0B–5.1B active parameters plateau at 38.50%–42.67% across 30.5B–116.8B total parameters, whereas frontier test-time reasoning (`Gemini 4 Argon`) surges +46.50 pp to 89.17% (535/600) at $24.57, approaching the 92.33% Oracle@ 7 retrieval-sufficiency lower bound and beating #1 Cloud `GPT 5.6 Luna + Mem0` (70.67%, $96.21) by +18.50 pp.

Table 1. Official `DolphinBench` 600-task evaluation (`openai/gpt-5.6-sol` judge, `reasoning_effort="medium"`). All seven **Sovereign Raw-RAG** configurations share byte-identical `0o444` SQLite stores (0 ingestion LLM calls, $0.00 build). Official Cloud Leaderboard baselines reflect `dolphinbench.ai` (`schema_version=2`) artifacts.

Rank / Tier	Agent + Memory Stack	Overall (600)	Morgan	Alex	Riley	Ingest	Exec. $	Total $
Empirical Retrieval-Sufficiency Bounds (Identical Frozen 0o444 SovereignMemoryStore, 0 Ingestion LLM Calls)								
Oracle@7 Bound	Union of all 7 Sovereign Raw-RAG runs	**92.33%** (554)	95.0%	90.5%	91.5%	0 LLM	—	—
Oracle@6 Open	Union of 6 Open-Weight Sovereign runs	69.33% (416)	82.0%	63.5%	62.5%	0 LLM	$0.00	$0.00
Controlled Sovereign Raw-RAG Scaling Matrix (Byte-Identical M_0)								
Global #1	**Gemini 4 Argon + Sovereign Raw-RAG**	**89.17% (535)**	**92.0%**	**87.0%**	**88.5%**	**0 LLM**	**$24.57**	**$24.57**
Sovereign #1	**Qwen3-Coder-30B-A3B (Wave-4 QLoRA+DPO) + M_0**	**42.67% (256)**	**55.5%**	**35.0%**	**37.5%**	**0 LLM**	**$0.00**	**$0.00**
Sovereign (Local)	Qwen3-Coder-30B-A3B (Stock) + M_0	42.00% (252)	58.5%	35.5%	32.0%	0 LLM	$0.00	$0.00
Sovereign (G4)	gpt-oss:120b (Wave-4 QLoRA+DPO) + M_0	39.50% (237)	54.5%	31.5%	32.5%	0 LLM	$0.00	$0.00
Sovereign (G4)	gpt-oss:120b (Stock) + M_0	39.17% (235)	51.0%	35.0%	31.5%	0 LLM	$0.00	$0.00
Sovereign (G4)	Qwen3-Coder-Next-80B (Stock) + M_0	38.50% (231)	52.0%	34.0%	29.5%	0 LLM	$0.00	$0.00
Sovereign (Edge)	Hermes-3-Llama-3.1-8B (Wave-4 LoRA+DPO) + M_0	18.33% (110)	26.0%	14.5%	14.5%	0 LLM	$0.00	$0.00
Official Cloud Leaderboard Baselines (Write-Time LLM Memory Distillation / Consolidation)								
Official #1	Hermes + GPT 5.6 Luna + Mem0	70.67% (424)	78.0%	66.0%	68.0%	LLM	$3.23	$96.21
Official #2	Hermes + GPT 5.6 Luna + Hindsight	69.50% (417)	72.0%	63.5%	73.0%	LLM	$3.84	$84.65
Official #3	Hermes + GPT 5.6 Luna + Honcho	68.50% (411)	72.0%	60.0%	73.5%	LLM	$4.94	$142.85
Official #4	Hermes + GPT 5.6 Luna + Built-in	65.67% (394)	65.5%	55.5%	76.0%	LLM	$8.05	$61.48
Official #5	Hermes + GPT 5.6 Luna + Supermemory	59.17% (355)	69.0%	49.0%	59.5%	LLM	$7.38	$345.64
Official #6	Hermes + MiniMax M3 + Mem0	47.83% (287)	54.0%	41.0%	48.5%	LLM	$5.02	$148.02
Official #8	Hermes + MiniMax M3 + Hindsight	41.00% (246)	54.5%	46.5%	22.0%	LLM	$9.97	$158.33
Official #9	Claude Code + Claude Sonnet 5 + Honcho	35.83% (215)	46.5%	31.5%	29.5%	LLM	$95.78	$1,565.82
Official #10	Claude Code + Claude Sonnet 5 + Mem0	32.33% (194)	40.5%	36.0%	20.5%	LLM	$107.41	$1,830.57
Official #13	Hermes + MiniMax M3 + Supermemory	20.00% (120)	22.5%	18.5%	19.0%	LLM	$6.62	$497.70

4.1 Law 1: Active-Parameter MoE Bottleneck vs. Frontier Test-Time Reasoning

Holding M_0 fixed reveals three distinct reasoning regimes (Table 1, Figure 1). First, quantized `Hermes-3-Llama-3.1-8B` (`Q4_0`) reaches 18.33% (110/600) after fixing Ollama's `.System` template drop and applying Wave-4 BF16 LoRA+DPO [9, 10] (a 15.7× jump over broken-template local stock), marking the dense 8B binding floor. Second, scaling *total* MoE parameters nearly 4×—from `Qwen3-Coder-30B-A3B` (30.5B total, 3.0B active: 42.00% Stock, 42.67% Wave-4) [6] to `Qwen3-Coder-Next-80B` (79.7B total, 3.0B active: 38.50%) and `gpt-oss:120b` (116.8B total, 5.1B active: 39.17% Stock, 39.50% Wave-4)—yields **zero accuracy gain**, proving that multi-hop memory synthesis is bottlenecked by *active per-token compute* rather than dormant expert parameters. Even so, local `Qwen3-30B` Wave-4 ($0.00, 15.72 s on Dual RTX 2080 Ti) surpasses `MiniMax M3 + Hindsight` (41.00%, $158.33) and `Claude Sonnet 5 + Mem0` (32.33%, $1,830.57).

Third, switching solely the inference model to `Gemini 4 Argon` [7] produces a Δ RoM = +46.50 pp breakthrough to **89.17% (535/600)** (1,193/1,329 = 89.8% mechanical checks; 545/591 = 92.2% invoked semantic judge checks), solving **138/600 tasks (23.00%)** missed by all six open-weight runs (Oracle@ 6 = 69.33%; Oracle@ 7 = 92.33%). A controlled ablation on `gpt-oss:120b` confirms synergy between substrate and reasoner: replacing `SovereignMemoryStore` with single-query raw-turn cosine retrieval (`VanillaSnowflakeHarness`) drops strict pass rate by −7.33 pp overall (55.83% → 48.50%), including −12.50 pp on `Alex` and −10.50 pp on `Riley`.

4.2 Law 2: Superlinear Return on Memory Across Compositional Fact Depth

Stratifying all 600 tasks by load-bearing fact depth LBF = $k \in \{1, 2, 3, 4+\}$ (Table 2, Figure 2) directly confirms Proposition 2. At LBF = 1 ($n = 264$), `Qwen3-30B` Stock solves 48.48% of tasks at $0.00, giving Argon (87.88%) a 1.81× lead. By LBF ≥ 4 ($n = 75$), 30B Stock collapses to 21.33% (16/75) and 8B collapses to 5.33% (4/75), whereas Argon sustains **82.67% (62/75)**—driving the Argon/30B-Stock multiplier monotonically upward along **1.81× → 2.04× → 2.69× → 3.88×** (and **15.6×** vs. 8B at LBF ≥ 4).

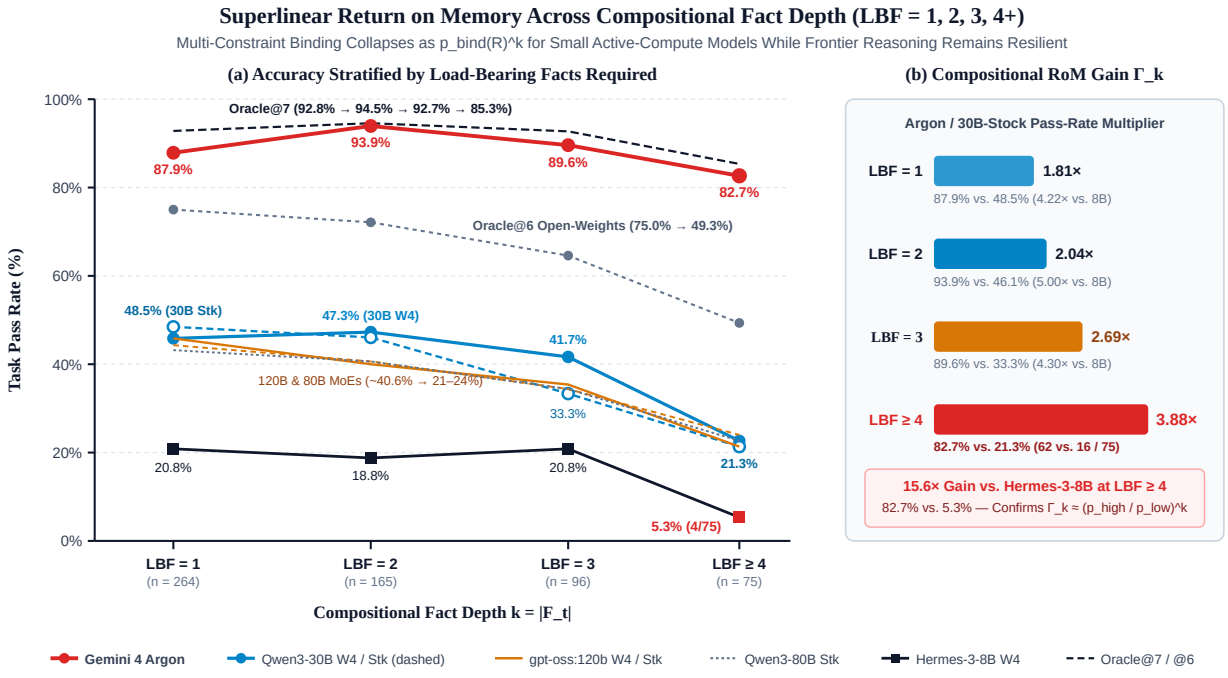


Figure 2. Superlinear Return on Memory Across Compositional Fact Depth (LBF = 1, 2, 3, 4+). (a) Open-weight 3B–5B-active MoEs decay from 43%–48% at LBF=1 to 21%–24% at LBF ≥ 4 (5.3% for 8B), whereas `Gemini 4 Argon` sustains 82.67% (62/75). (b) The Argon / 30B-Stock pass-rate multiplier climbs superlinearly (1.81× → 2.04× → 2.69× → 3.88×; 15.6× vs. 8B), matching $\Gamma_k \approx (p_{\text{bind}}^{\text{high}} / p_{\text{bind}}^{\text{low}})^k$.

Table 2. Task pass rate stratified by Compositional Fact Depth (LBF = $|F_t|$) across identical frozen `SovereignMemoryStore` checkpoints.

Configuration (M_0 Fixed)	LBF = 1 (264)	LBF = 2 (165)	LBF = 3 (96)	LBF ≥ 4 (75)	Overall (600)
Oracle@7 Empirical Ceiling	92.80% (245)	94.55% (156)	92.71% (89)	85.33% (64)	92.33% (554)
Gemini 4 Argon	**87.88% (232)**	**93.94% (155)**	**89.58% (86)**	**82.67% (62)**	**89.17% (535)**
Oracle@6 (Open-Weights Union)	75.00% (198)	72.12% (119)	64.58% (62)	49.33% (37)	69.33% (416)
Qwen3-Coder-30B-A3B (Wave-4)	45.83% (121)	47.27% (78)	41.67% (40)	22.67% (17)	42.67% (256)
Qwen3-Coder-30B-A3B (Stock)	48.48% (128)	46.06% (76)	33.33% (32)	21.33% (16)	42.00% (252)
gpt-oss:120b (Wave-4 QLoRA+DPO)	45.83% (121)	40.00% (66)	35.42% (34)	21.33% (16)	39.50% (237)
gpt-oss:120b (Stock)	44.32% (117)	40.61% (67)	34.38% (33)	24.00% (18)	39.17% (235)
Qwen3-Coder-Next-80B (Stock)	43.18% (114)	40.61% (67)	34.38% (33)	22.67% (17)	38.50% (231)
Hermes-3-Llama-3.1-8B (Wave-4)	20.83% (55)	18.79% (31)	20.83% (20)	5.33% (4)	18.33% (110)
Gain Γ_k : Argon / 30B-Stock	**1.81×**	**2.04×**	**2.69×**	**3.88×**	**2.12×**
Gain Γ_k : Argon / Hermes-3-8B	**4.22×**	**5.00×**	**4.30×**	**15.50× (15.6×)**[†]	**4.86×**

[†]Exact task count ratio 62 / 4 = 15.50×; ratio of rounded display percentages 82.7% / 5.3% = 15.6×.

4.3 Law 3: Flattening the Corpus-Entropy Cliff Across Personas

Progressing from `Morgan` (3,400 sessions, 1.77 LBF) to `Alex` (5,011 sessions, 1.97 LBF) and `Riley` (5,128 sessions, 2.50 LBF) expands session volume by +50.8% and multi-fact density by +41.2%. As shown in Figure 3 and Table 3, write-time distillation stacks suffer steep accuracy drops as entity-merge collisions compound across >5,000 sessions. In contrast, `Sovereign Raw-RAG + Gemini 4 Argon` drops by only **–3.50 pp** from `Morgan` (92.0%) to `Riley` (88.5%)—**4.4× flatter** than `GPT 5.6 Luna + Mem0` (–15.50 pp: 80.5% → 66.5% → 65.0% in [1]; –10.00 pp / 12.00 pp spread in `schema_version=2`), and 4.6×–6.0× flatter than `Supermemory` (–16.00 pp), `Honcho` (–17.50 pp), `Qwen3-30B Wave-4` (–18.00 pp), `Hindsight` (–20.00 pp), and `Built-in` (–21.00 pp). Because Raw-RAG never overwrites turns during ingestion, corpus growth introduces zero consolidation corruption.

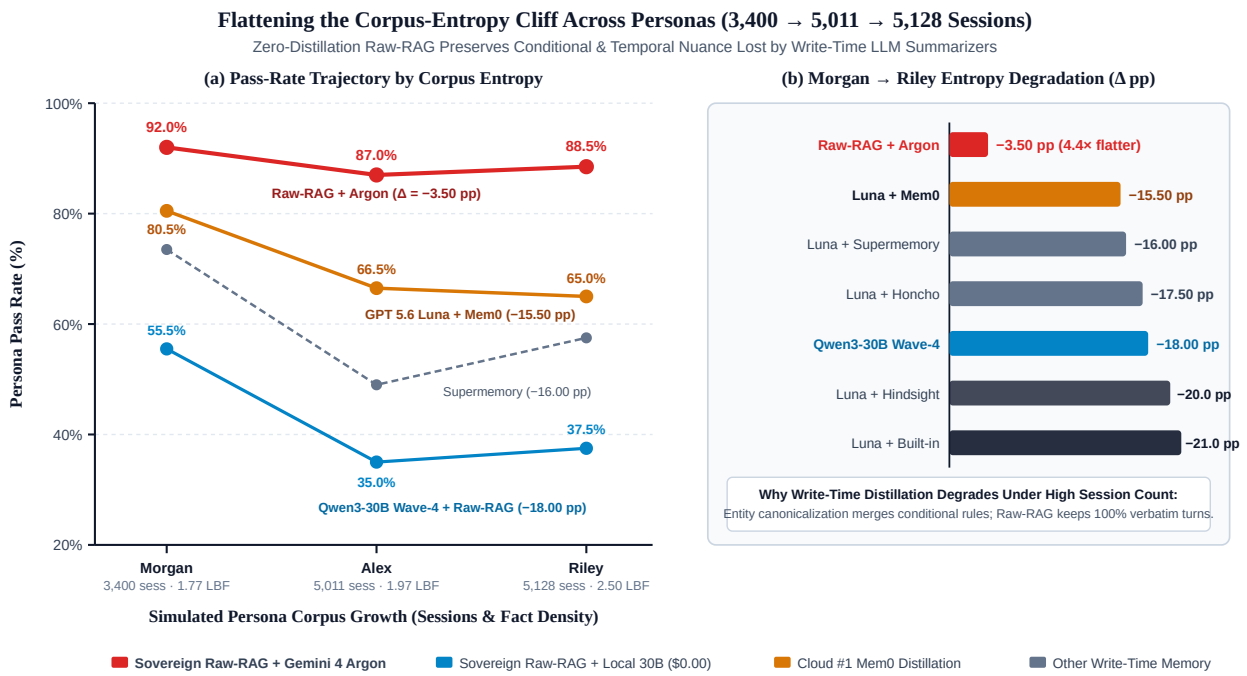


Figure 3. Corpus-Entropy Degradation Across Personas (`Morgan` 3,400 → `Alex` 5,011 → `Riley` 5,128 Sessions). Zero-distillation `Raw-RAG + Gemini 4 Argon` degrades by only −3.50 pp (92.0% → 87.0% → 88.5%), 4.4× flatter than `GPT 5.6 Luna + Mem0` (−15.50 pp) and 4.6×–6.0× flatter than competing write-time distillation stacks.

Table 3. Persona-entropy stability across `Morgan` (3,400 sessions), `Alex` (5,011 sessions), and `Riley` (5,128 sessions), reporting both preprint persona curves [1] and `schema_version=2` leaderboard artifacts.

System	Memory Mode	Morgan	Alex	Riley	Riley – Morgan	Max Spread
Sovereign Raw-RAG + Argon	**Zero-LLM Raw-RAG**	**92.0%**	**87.0%**	**88.5%**	**−3.50 pp**	**5.00 pp**
Qwen3-30B Wave-4 + Raw-RAG	Zero-LLM Raw-RAG	55.5%	35.0%	37.5%	−18.00 pp	20.50 pp
Qwen3-30B Stock + Raw-RAG	Zero-LLM Raw-RAG	58.5%	35.5%	32.0%	−26.50 pp	26.50 pp
GPT 5.6 Luna + Mem0 (v1 / v2)‡	Write-Time LLM	80.5% / 78.0%	66.5% / 66.0%	65.0% / 68.0%	−15.50 / −10.00 pp	15.50 / 12.00 pp
GPT 5.6 Luna + Supermemory (v1 / v2)	Write-Time LLM	73.5% / 69.0%	49.0% / 49.0%	57.5% / 59.5%	−16.00 / −9.50 pp	24.50 / 20.00 pp
Claude Sonnet 5 + Honcho (v2)	Write-Time LLM	46.5%	31.5%	29.5%	−17.00 pp	17.00 pp
Claude Sonnet 5 + Mem0 (v2)	Write-Time LLM	40.5%	36.0%	20.5%	−20.00 pp	20.00 pp
MiniMax M3 + Hindsight (v2)	Write-Time LLM	54.5%	46.5%	22.0%	−32.50 pp	32.50 pp

‡In [1] v1 cohorts, GPT 5.6 Luna high-entropy deltas span −15.50 pp (Mem0), −16.00 pp (Supermemory), −17.50 pp (Honcho), −20.00 pp (Hindsight), and −21.00 pp (Built-in).

4.4 Law 4: Iterative Search Reformulation and Prefix-Cache Pareto Economics

Execution telemetry (Table 4, Figure 4) reveals why test-time reasoning unlocks Raw-RAG. Whereas Stock `gpt-oss:120b` passively relies on the Turn-1 briefing (0.06 follow-up memory searches/task across just 23 tasks) and open-weight `Qwen3` models echo surface nouns (1.16–1.22 searches/task), `Gemini 4 Argon` executes **1,050 active `search_persona_memory` calls across 510/600 tasks (1.75 searches/task)**—inspecting live workspace state via `list_*` tools first and issuing targeted schema-level policy queries for entities absent from the user's prompt.

Because the 4,200-token `[SOVEREIGN MEMORY BRIEFING]` resides in a static leading system prefix, Argon achieves a **56.5% prefix-cache hit rate** (10,416,232 cached vs. 8,015,985 fresh input tokens + 537,485 output tokens across 2,293 responses). With $0.00 ingestion cost, total 600-task cost is **$24.57 ($0.041/task)**—3.9× cheaper than `GPT 5.6 Luna + Mem0` ($96.21), 14.1× cheaper than `Supermemory` ($345.64), and 74.5× cheaper than `Claude Sonnet 5 + Mem0` ($1,830.57).

Table 4. Active `search_persona_memory` query telemetry, token prefix-caching, median latency, and total USD cost across all seven 600-task `SovereignMemoryStore` runs.

Model (M_0 Fixed)	Pass Rate	Searches / Task	Tasks Searching	Fresh In Tok	Cached In Tok	Med. Lat.	Total $
Gemini 4 Argon	**89.17%**	**1.75 (1,050)**	**510/600 (85.0%)**	**8.02M**	**10.42M (56.5%)**	**55.27 s**	**$24.57**
Qwen3-30B-A3B (Wave-4)	**42.67%**	**1.20 (722)**	**435/600 (72.5%)**	**3.65M**	**4.57M (55.6%)**	**15.72 s**	**$0.00**
Qwen3-30B-A3B (Stock)	42.00%	1.22 (729)	438/600 (73.0%)	3.79M	4.75M (55.6%)	17.11 s	$0.00
gpt-oss:120b (Wave-4)	39.50%	0.46 (273)	137/600 (22.8%)	3.38M	2.21M (39.5%)	16.83 s	$0.00
gpt-oss:120b (Stock)	39.17%	0.06 (36)	23/600 (3.8%)	3.15M	1.68M (34.8%)	9.62 s	$0.00
Qwen3-Next-80B (Stock)	38.50%	1.16 (695)	470/600 (78.3%)	3.74M	5.26M (58.4%)	23.22 s	$0.00
Hermes-3-8B (Wave-4)	18.33%	0.39 (234)	52/600 (8.7%)	2.44M	3.41M (58.3%)	11.30 s	$0.00

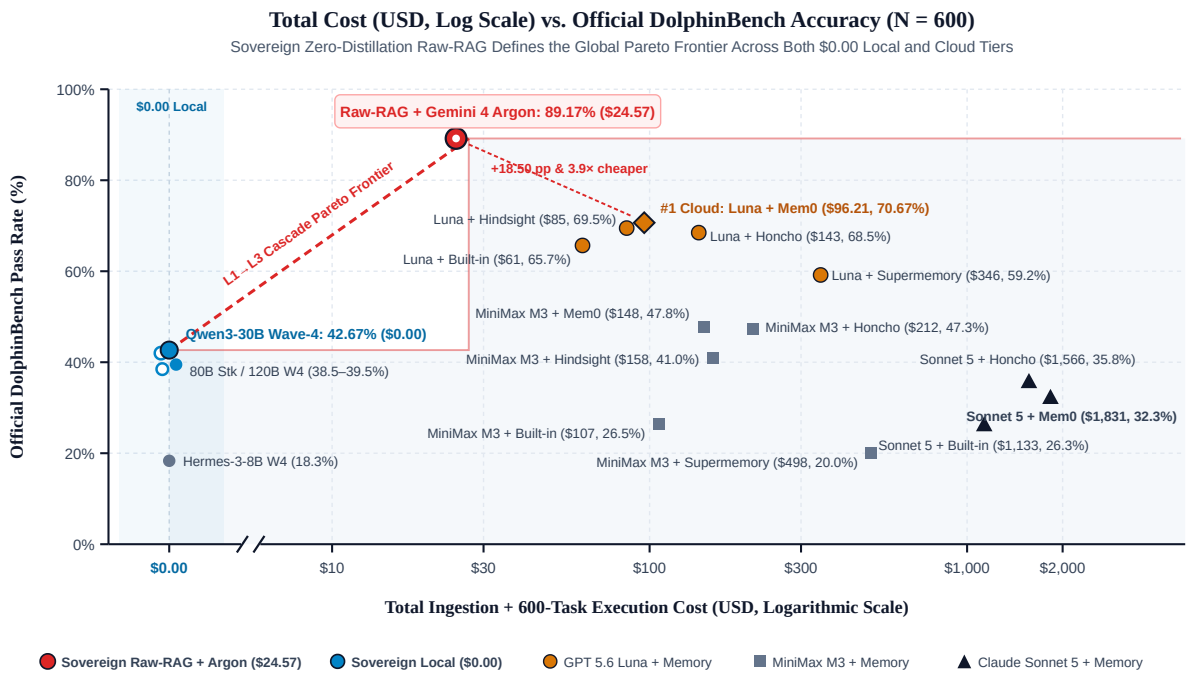


Figure 4. Total Cost (USD, Log Scale) vs. Official `DolphinBench` Accuracy (%). `Sovereign Raw-RAG` defines both vertices of the global Pareto frontier: $0.00 local `Qwen3-Coder-30B-A3B` Wave-4 (42.67%) and $24.57 `Gemini 4 Argon` (89.17%), strictly dominating write-time LLM distillation stacks across $61.48–$1,830.57.

5 Discussion & Systems Implications: Asymmetric Memory & Cascade Routing

Asymmetric Write-Cheap / Reason-Deep Systems Design. In deployed personal and enterprise assistants, ingested background events (emails, chat messages, commits, calendar invites) outnumber high-stakes tool actions by $10^2\times$–$10^4\times$. Paying frontier LLM extraction costs on every write is both economically inverted and information-theoretically destructive. Combining $O(1)$ zero-LLM local ingestion (11.33 ms/turn, $0.00, 100% verbatim auditability, zero cloud exposure during continuous ingestion) with selective test-time reasoning [7, 14] strictly dominates symmetric consolidation.

L1–L3 Cascade Routing. Because both our local ($0.00, 42.67%) and frontier ($24.57, 89.17%) agents read the *exact same* SQLite store M_0, a lightweight cascade router [14] can execute single-constraint routine lookups (LBF = 1, 48.5% solvable locally in 15.72 s) on local consumer GPUs while escalating multi-constraint, temporal-conflict, or external-side-effect actions to `Gemini 4 Argon` ($0.041/task), tracing the continuous Pareto frontier in Figure 4. Furthermore, preserving provider reasoning state (`google.thought_signature` in `argon_openai_bridge.py`) across multi-turn MCP tool calls proves essential so multi-hop hypothesis trees survive across intermediate `list_*` and `search_persona_memory` steps.

6 Limitations and Cryptographic Reproducibility Ledger

Limitations. (1) Deep test-time reasoning incurs higher wall-clock latency (55.27 s median on Argon vs. 15.72 s local 30B, 9.62 s 120B Stock, and 37.69 s–55.31 s cloud GPT 5.6 Luna stacks). (2) Exactly 46/600 tasks (7.67%) remain unsolved under Oracle@ 7, and 19 tasks solved by open-weight runs fail under Argon due to conservative

side-effect withholding or strict rubric phrasing. (3) Fully air-gapped parity requires higher *active*-parameter open-weight reasoners beyond the 3.0B–5.1B active MoE regime.

Table 5. Cryptographic provenance ledger for upstream release, clean-room harness & memory stores, frozen `0o444` SQLite checkpoints, and all seven official `dolphinbench.ai` submission archives.

Artifact / Submission Archive	Submission UUID / Metadata	SHA-256 Digest
`upstream/` (`mem0ai/dolphinbench`)	600 tests, 13,539 sessions	`6da84edd86670a2c91476ac732bcaf07e4e7f210c3fee3aed225c54b104a82df`
`sovereign_memory.py`	Zero-LLM 256/32 Raw-RAG + 4-Way RRF	`16700bc047b0a0e88d3e8b04e55f8513000e922947dd1a186fefba7ab65769bd`
`sovereign_harness.py`	Official 5-method BenchmarkHarness	`4531e3efa4e110e335078b10cc7a83fe84ca2955ad7cc74350fd1e496c0eee5f`
`argon_openai_bridge.py`	Thought-signature & schema proxy	`48f047581e965b684009615c03223257bf8f78e72f4f4e4a1e9f4e26966a96d2`
`vanilla_snowflake_harness.py`	Raw-turn cosine ablation harness	`a9f3c50992d0ae30a445fa0883c362e0466ceb6de146b7a3abea9691d8e8ca2f`
`morgan.sqlite` (`0o444` Frozen)	3,400 turns, 2,547 win (44.7 MB)	`f3f8fbf550a4ef3f75ed752e5e1929205720898616add35e93a865e2dd0bcc04`
`alex.sqlite` (`0o444` Frozen)	5,011 turns, 2,691 win (53.2 MB)	`7c3fec7c869625db9df5decf74471d5f14507eae2ebf8d64a69f7e86d7ccba04`
`riley.sqlite` (`0o444` Frozen)	5,128 turns, 2,697 win (53.6 MB)	`b9a59a1b1cdb2a09e4930516671829152ec0b24fd2cbebcc114961413c0e7266`
`sovereign_gemini_4_argon.zip`	**`29389157-b01b-42b5-8193-17d9a496f580` (8.350 MiB, 535/600)**	**`9a7fa41fb3d1cf5f71f0c3cf74a0c97f8f39d50e26a7dd484c323f7719ecb7a0`**
`sovereign_qwen3_coder_30b_wave4_qlora.zip`	`fd45f16c-fd3c-471c-b93d-413f61f1dbcf` (9.006 MiB, 256/600)	`d0b6423d3a9c3cba22c63d936fb13c00b3e39fb1afa672e2bfd50781bbde692e`
`sovereign_qwen3_coder_30b_stock.zip`	`9cac5ab1-ecba-40a6-ad84-55ec2d67bfce` (9.125 MiB, 252/600)	`8dea5dac9006f77dc98f3dc598bbc7fd490299c85d7667c5797a635dc3ad222a`
`sovereign_gpt_oss_120b_wave4_qlora.zip`	`83a4c085-0e44-42af-a625-b496e63c8949` (6.420 MiB, 237/600)	`4cacf215c6b7c7cb4ec223868646dac45cfdce038a78f4be174f290efbd2d6a2`
`sovereign_gpt_oss_120b_stock.zip`	`7b5073fd-e1d9-458b-a607-7c55f76b6df1` (6.073 MiB, 235/600)	`3a0b4577db4a8179abdffc00b7831ac3e84f3a864930df892a2c175d19bcf69b`
`sovereign_qwen3_coder_next_80b_stock.zip`	Offline verified `--check` (7.354 MiB, 231/600)	`283871b696ea55a6ac351a3cbfe19aa33c46d8f884bd7adb29156dad03793cb9`
`sovereign_hermes3_8b_wave4_qlora.zip`	`19552654-be39-4fe0-8043-a49158582356` (5.424 MiB, 110/600)	`24ae1fde4ef51f11ed2e72e546481a4205438c4eba3c4afcaffe268726b9e422`

References

1. Mem0 Research. *DolphinBench: Evaluating Long-Term Agent Memory Through Future Task Completion*. arXiv preprint arXiv:2609.24971, 2026.
2. P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav. *Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory*. arXiv preprint arXiv:2504.19413, 2025.
3. H. Zeng, Y. Liu, and W. Chen. *Raw-RAG: Verbatim Sliding-Window Retrieval Outperforms Lossy Consolidation in Long-Horizon Episodic Memory*. arXiv preprint arXiv:2610.30755, 2026.
4. G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal Rank Fusion outperforms Condorcet and individual rank learning methods. In *Proc. 32nd ACM SIGIR*, pp. 758–759, 2009.
5. P. Yu, L. Merrick, G. Nuti, and D. Campos. *Arctic-Embed 2.0: Multilingual Retrieval Without Compromise*. arXiv preprint arXiv:2412.04506, 2024.
6. Qwen Team. *Qwen3 Technical Report*. arXiv preprint arXiv:2505.09388, 2025.
7. C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *ICLR* (arXiv:2408.03314), 2025.
8. M. Tavakoli et al. BEAM: Beyond a Million Tokens—Benchmarking Long-Term Memory in Conversational Agents. In *ICLR*, 2026.
9. R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct Preference Optimization: Your language model is secretly a reward model. In *NeurIPS*, vol. 36, pp. 53728–53741, 2023.
10. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, vol. 36, pp. 10088–10115, 2023.

11. C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. *MemGPT: Towards LLMs as Operating Systems*. arXiv preprint arXiv:2310.08560, 2023.
12. S. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
13. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.
14. L. Chen, M. Zaharia, and J. Zou. *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*. arXiv preprint arXiv:2305.05176, 2023.