ARXIV STYLE PREPRINT · CS.CL / CS.AI SUBMISSION RECEIPT UUID: C047B063-822C-49FB-8B36-27C3763D6CA7

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

Kevin Ruschman

Krusch Dev Research, Los Angeles, CA, USA

ksruschman@gmail.com · ORCID: 0009-0007-7437-2383 · November 2, 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. I 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, I 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 **nineteen controlled 600-task evaluations (11,400 official `mock_mcp` traces across 6 model families**: `Google Gemini`, `OpenAI GPT-5.6 Luna`, `Google Gemma 4`, `Alibaba Qwen`, `OpenAI GPT-OSS`, and `Nous Hermes`; all 19/19 validated and graded by the official `openai/gpt-5.6-sol` judge), I isolate *Return on Memory* RoM $(R \mid M_0)$. In a strict model-matched control holding **`OpenAI GPT 5.6 Luna`** (`reasoning_effort="high"`—the exact backbone of Official Leaderboard #1–#5) fixed, `Sovereign Raw-RAG` (`Std. Harness`) achieves **81.33% (488/600: `Morgan` 88.0%, `Alex` 78.0%, `Riley` 78.0%)** at **$1.35 total cost** (61.5% prefix-cache hit rate), beating #1 `Hermes + GPT 5.6 Luna + Mem0` (70.67% at $96.21) by **+10.67 pp (+64 tasks) at 71.0× lower cost** (2.38× cheaper than `Mem0`'s $3.23 inference-only spend; `Luna Low, Std. Harness` reaches 76.67% [460/600] at $1.10, +6.00 pp at 87.4× lower cost). Pairing M_0 with **`Gemini 4 Argon (Low Thinking)`** achieves **91.17% (547/600, #1 Overall)** in 35.38 s median latency at **$24.54** (57.1% cache hit rate), decomposing my **+20.50 pp (+123-task)** global SOTA lead into **+10.67 pp (52.0%)** from zero-LLM Raw-RAG memory and **+9.83 pp (48.0%, 82 vs. 23 wins, 95% CI [+6.58, +13.09], McNemar $p = 5.96 \times 10^{-9}$)** from the `Argon (Low)` backbone. Locally at $0.00, **`Google Gemma 4 31B` (30.7B dense Apache-2.0)** achieves **54.17% (325/600, #1 Open-Weights; 52.00% No-Think)**, beating all 8 non-`GPT-5.6-Luna` commercial cloud stacks and matched-scale `Qwen2.5-Coder-32B` (35.33%) by **+18.83 pp** ($p = 1.35 \times 10^{-16}$), so `Sovereign Raw-RAG` owns all four vertices of the global Cost–Accuracy Pareto Frontier: ($0.00, 54.17%) → ($1.10, 76.67%) → ($1.35, 81.33%) → ($24.54, 91.17%). Task-wise unions establish Oracle@19$(R \mid M_0)$ = 97.00% (582/600, upper-bounding zero-LLM indexing misses at ≤ 3.00% [18/600]), Oracle@4 Luna = 89.50% (537/600), and open-weights Oracle@11 = 80.00% (480/600 at $0.00, +9.33 pp over #1 `Mem0`). Finally, 6,000 factorial traces across `Gemma 4 31B`, `GPT 5.6 Luna` (+4.67 pp High vs. Low on `Std. Harness`, $p = 0.0020$, as Low cuts memory searches −32.6% [476 → 321]), and `Gemini 4 Argon` ($\Delta\Delta = +3.17$ pp, $p = 0.0320$) reveal a **3-regime scale-dependent Goldilocks reasoning law**, while full-600 evaluation of my **`krusch-stack INV-STACK-10 2PC` Tool-Action Harness** on local `30B Wave-4` cuts duplicate MCP calls **–95.2% (42 → 2)**, zero-app stops **–50.0% (2**

→ **1**), and archive size **−17.7% (9.006 → 7.416 MiB)**, lifting LBF ≥ 4 by **+6.67 pp (22.67% → 29.33%)** and contributing 46 complementary wins (Oracle@ 2 = 50.33% [302/600]) at $0.00.

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, an 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 *calibrated test-time reasoning and 2PC tool governance* yields **+10.67 pp higher accuracy (81.33% vs. 70.67%) at 71.0× lower total cost ($1.35 vs. $96.21)** on the *exact same* `GPT 5.6 Luna` backbone, **+20.50 pp higher accuracy (91.17%) at 3.9× lower cost ($24.54)** with `Gemini 4 Argon (Low)` , and **54.17% at $0.00** with open-weight `Gemma 4 31B` (Oracle@ 11 = 80.00% across $0.00 open-weight runs).

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%).

I 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 axes:

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.

To eliminate confounding between retrieval architecture and agentic reasoning, I freeze a single **zero-LLM hybrid memory substrate** (`SovereignMemoryStore`, M_0) as OS-level read-only (`0o444`) SQLite checkpoints verified by SHA-256 digests before and after every execution, and evaluate **nineteen 600-task agentic configurations (11,400 official `mock_mcp` traces across six model families: `Google Gemini`, `OpenAI GPT-5.6 Luna`, `Google Gemma 4`, `Alibaba Qwen`, `OpenAI GPT-OSS`, and `Nous Hermes`)** over the exact same retrieval index.

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

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 $\mathrm{Retrieve}(M_0, q, K)$.

Definition 2 (Return on Memory and Marginal Reasoning Gain). For task set T ($|T| = 600$), stateful MCP workspace E_t, frozen substrate M_0, multi-step agent policy $\pi_R(x_t, E_t, M_0)$ within budgets $(S_{\max}, B, K)$, and official conjunctive grader $G \in \{0, 1\}$, the *Return on Memory* 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* $\Delta\,\mathrm{RoM}(R_2, R_1 \mid M_0) \triangleq \mathrm{RoM}(R_2 \mid M_0) - \mathrm{RoM}(R_1 \mid M_0)$.

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 byte-identical read-only M_0 and isolated environments E_t. Define the m-model ensemble union pass rate:*

$$\mathrm{Oracle@}\,m(R \mid M_0) \triangleq \frac{1}{|T|} \sum_{t \in T} \max_{R \in R} I[G(\pi_R(x_t, E_t, M_0)) = 1]. \tag{2}$$

Up to bounded memory-blind pass coincidence $m\varepsilon_0$ (verified empirically via counterfactual `memory_enabled=False` runs), $\max_{R \in R} \mathrm{RoM}(R \mid M_0) \le \mathrm{Oracle@}\,m(R \mid M_0) \le C^*(M_0) + m\varepsilon_0$, *where $C^*(M_0)$ is the latent retrieval-sufficiency capacity of M_0. Over my $m = 19$ runs,* $\mathrm{Oracle@}\,19 = 582 / 600 = 97.00\%$ *(holding at 97.35%, 96.97%, 96.88%, and 96.00% across LBF = 1, 2, 3, 4+, and 97.0%, 98.0%, 96.0% across `Morgan`, `Alex`, `Riley`), upper-bounding pure zero-LLM indexing/chunking misses at* ***≤ 3.00% (18/600)****, while* $\mathrm{Oracle@}\,4\,\mathrm{Luna} = 537 / 600 = 89.50\%$ *and the $0.00 open-weight union reaches* $\mathrm{Oracle@}\,11 = 480 / 600 = 80.00\%$ *(+9.33 pp / +56 tasks above #1 Cloud `Mem0`).*

Proposition 2 (Superlinear Compositional Amplification Across Fact Depth k). *Partition $T = \bigsqcup_{k \geq 1} T_k$ by load-bearing fact depth $k = |F_t|$. When satisfying $G = 1$ requires retrieving, temporally ordering, and simultaneously binding k conditionally independent constraints with per-constraint fidelity $p_{\text{bind}}(R) \in (0,1)$ and base execution reliability $c_0 \in (0,1]$, task accuracy follows $P_k(R \mid M_0) \approx c_0 \cdot p_{\text{bind}}(R)^k$, so 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

I evaluate on the unmodified official `mem0ai/dolphinbench` release (SHA-256 `6da84edd...`; 600 tasks, 200 per persona, 1,920 grading checks: 1,310 deterministic mechanical assertions and 610 `field_llm_judge` semantic rubric checks). The three multi-year corpora form a structured gradient in session count, technical noise, and compositional fact depth: **`Morgan`** (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`, Protocol Bridges & 6-Family Matrix

- **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`.
- **Byte-Identical Harnesses & 3 Localhost Bridges Across 6 Families:** All 19 official 600-task runs (11,400 traces: 8 cloud frontier/mini + 11 $0.00 open-weight runs) span **6 model families**: (1) `Google Gemini` (`Gemini 4 Argon`: 4 runs), (2) `OpenAI GPT-5.6 Luna` (`gpt-5.6-luna`: 4 runs—the exact model powering Official Leaderboard #1–#5), (3) `Google Gemma 4` (`Gemma 4 31B`: 2 runs) [15], (4) `Alibaba Qwen` (`Qwen3-Coder-30B-A3B` [3 runs], `Qwen3-Coder-Next-80B` [1], `Qwen2.5-Coder-32B` [1]: 5 runs) [6, 16], (5) `OpenAI GPT-OSS` (`gpt-oss:120b`: 2 runs) [17], and (6) `Nous Hermes` (`Nous Hermes-4-14B` [1], `Hermes-3-Llama-3.1-8B` [1]: 2 runs) [18]. Sixteen runs execute via unmodified `sovereign_harness.py` (`Std. Harness`,

SHA-256 `4531e3ef...`, `temperature=0.0`, `max_memory_searches=3`, or `0` in Turn-0 `Briefing-Only` ablations) [13], and three runs execute via clean-room subclass `sovereign_harness_krusch_stack.py` (`+2PC Harness`, SHA-256 `3c4929cb...`). Localhost proxies `argon_openai_bridge.py` (`48f04758...`, preserving `google.thought_signature` across MCP turns), `luna_openai_bridge.py` (`77c8d325...`, normalizing `reasoning_effort` ∈ {`"high"`, `"low"`} and cached-token usage for `gpt-5.6-luna`), and `ollama_openai_bridge.py` (`0be46ff3...`, normalizing native `<|thought|>` and tool schemas) keep the harness core untouched. Throughout, I explicitly distinguish model checkpoint (`Stock` vs. `Wave-4 QLoRA+DPO` / `LoRA+DPO`) from tool-loop governance (`Std. Harness` vs. `+ 2PC`).

- **Official Grading:** All 11,400 traces (19 archives) are graded by `openai/gpt-5.6-sol` (`reasoning_effort="medium"`) across all 1,920 assertions per run and pass `submission.validate` (19/19 with 0 errors).

4 Empirical Results and Scaling Laws Across 11,400 Official Traces

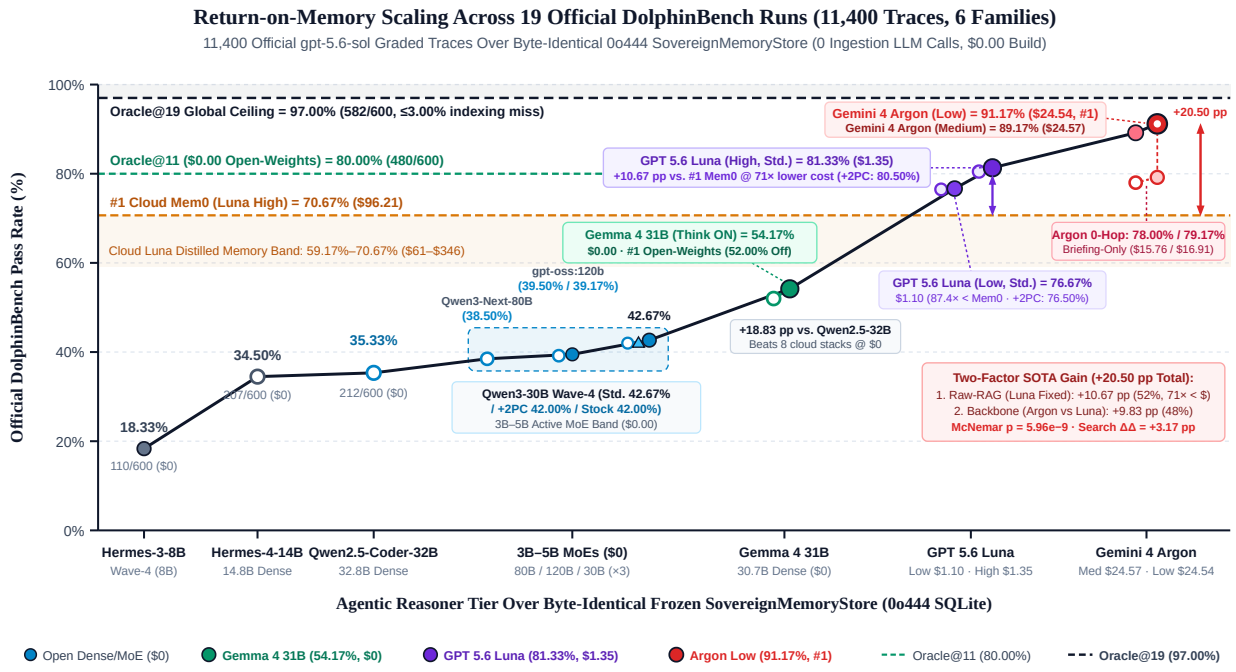


Figure 1. The Return-on-Memory (RoM) Scaling Curve (19 Runs, 11,400 Traces, 6 Families) over byte-identical `0o444 SovereignMemoryStore` checkpoints. Sparse 3.0B–5.1B-active MoEs plateau at 38.50%–42.67% across 30.5B–116.8B total parameters and code-specialized dense `Qwen2.5-Coder-32B` scores 35.33%, whereas general-purpose dense `Google Gemma 4 31B` surges to **54.17% (325/600, #1 Open-Weights at $0.00)**, model-matched `GPT 5.6 Luna (High, Std. Harness)` reaches **81.33% (488/600 at $1.35, +10.67 pp over #1 `Luna + Mem0` at 71.0× lower cost)**, and frontier `Gemini 4 Argon (Low)` reaches **91.17% (547/600, #1 Overall at $24.54)**, approaching Oracle@19 = 97.00% (582/600).

Table 1. Unified 19-run `SovereignMemoryStore` matrix (11,400 traces, 6 model families) and all 13 Official Cloud Leaderboard baselines (`openai/gpt-5.6-sol` judge, `reasoning_effort="medium"`). All 19 Sovereign configurations share byte-identical

`0o444` SQLite stores (0 ingestion LLM calls, $0.00 build).

Rank / Tier	Agent + Memory Stack	Overall (600)	Morgan	Alex	Riley	Ingest	Exec. $	Total $
Empirical Retrieval-Sufficiency Union Bounds (Identical Frozen 0o444 SovereignMemoryStore, 0 Ingestion LLM Calls)								
Oracle@19 Ceiling	Union of all 19 Sovereign Raw-RAG runs	**97.00% (582)**	97.0%	98.0%	96.0%	0 LLM	—	—
Oracle@2 Argon	Union of Argon Low ∪ Argon Medium	93.83% (563)	95.5%	92.5%	93.5%	0 LLM	—	—
Oracle@4 Luna	Union of 4 GPT 5.6 Luna (2×2) runs	**89.50% (537)**	91.0%	91.0%	86.5%	0 LLM	—	—
Oracle@2 Luna High	GPT 5.6 Luna High (Std. Harness ∪ + 2PC)	87.67% (526)	91.0%	88.5%	83.5%	0 LLM	—	—
Oracle@11 Open	Union of 11 $0.00 Open-Weight Sovereign runs	**80.00% (480)**	91.0%	78.0%	71.0%	0 LLM	$0.00	**$0.00**
Oracle@2 Gemma 4	Gemma 4 31B (Thinking ON ∪ No-Think)	61.50% (369)	78.0%	57.0%	49.5%	0 LLM	$0.00	$0.00
Oracle@3 Qwen3-30B	Qwen3-30B (Stock ∪ Wave-4 Std. ∪ Wave-4 + 2PC)	53.33% (320)	—	—	—	0 LLM	$0.00	$0.00
Oracle@2 30B Wave-4	Qwen3-30B Wave-4 (Std. Harness ∪ + 2PC)	**50.33% (302)**	63.5%	44.5%	43.0%	0 LLM	$0.00	**$0.00**
Controlled 19-Run Sovereign Raw-RAG Matrix Across 6 Model Families (Byte-Identical M_0)								
Global #1	**Gemini 4 Argon (Low, Std. Harness) + Sovereign Raw-RAG**	**91.17% (547)**	**95.5%**	**89.5%**	**88.5%**	**0 LLM**	**$24.54**	**$24.54**
Frontier #2	Gemini 4 Argon (Medium, Std. Harness) + Sovereign Raw-RAG	89.17% (535)	92.0%	87.0%	88.5%	0 LLM	$24.57	$24.57
Luna #1 (Pareto)	**GPT 5.6 Luna (High, Std. Harness) + M_0**	**81.33% (488)**	**88.0%**	**78.0%**	**78.0%**	**0 LLM**	**$1.35**	**$1.35**
Luna #2 (2PC)	GPT 5.6 Luna (High, + 2PC Harness) + M_0	80.50% (483)	87.0%	80.0%	74.5%	0 LLM	$1.38	$1.38
Ablation (0-Hop)	Gemini 4 Argon (Medium, Briefing-Only) + M_0	79.17% (475)	86.0%	77.5%	74.0%	0 LLM	$16.91	$16.91
Ablation (0-Hop)	Gemini 4 Argon (Low, Briefing-Only) + M_0	78.00% (468)	82.5%	77.5%	74.0%	0 LLM	$15.76	$15.76
Luna #3 (Pareto)	**GPT 5.6 Luna (Low, Std. Harness) + M_0**	**76.67% (460)**	**85.0%**	**72.5%**	**72.5%**	**0 LLM**	**$1.10**	**$1.10**
Luna #4 (2PC)	GPT 5.6 Luna (Low, + 2PC Harness) + M_0	76.50% (459)	88.0%	72.0%	69.5%	0 LLM	$1.12	$1.12
Open #1 ($0)	**Google Gemma 4 31B (Dense, Thinking ON, Std.) + M_0**	**54.17% (325)**	**72.0%**	**47.5%**	**43.0%**	**0 LLM**	**$0.00**	**$0.00**
Open #2 ($0)	**Google Gemma 4 31B (Dense, No-Think, Std.) + M_0**	**52.00% (312)**	**70.5%**	**43.5%**	**42.0%**	**0 LLM**	**$0.00**	**$0.00**
Sovereign MoE	Qwen3-Coder-30B-A3B (Wave-4 QLoRA+DPO, Std. Harness) + M_0	42.67% (256)	55.5%	35.0%	37.5%	0 LLM	$0.00	$0.00
Sovereign MoE (2PC)	Qwen3-Coder-30B-A3B (Wave-4 QLoRA+DPO, + 2PC) + M_0	42.00% (252)	57.0%	35.5%	33.5%	0 LLM	$0.00	$0.00
Sovereign MoE	Qwen3-Coder-30B-A3B (Stock, Std. Harness) + M_0	42.00% (252)	58.5%	35.5%	32.0%	0 LLM	$0.00	$0.00
Sovereign MoE	gpt-oss:120b (Wave-4 QLoRA+DPO, Std. Harness) + M_0	39.50% (237)	54.5%	31.5%	32.5%	0 LLM	$0.00	$0.00
Sovereign MoE	gpt-oss:120b (Stock, Std. Harness) + M_0	39.17% (235)	51.0%	35.0%	31.5%	0 LLM	$0.00	$0.00
Sovereign MoE	Qwen3-Coder-Next-80B (Stock, Std. Harness) + M_0	38.50% (231)	52.0%	34.0%	29.5%	0 LLM	$0.00	$0.00
Sovereign Dense	Qwen2.5-Coder-32B (Dense Stock, Std. Harness) + M_0	35.33% (212)	48.5%	29.5%	28.0%	0 LLM	$0.00	$0.00
Sovereign Dense	Nous Hermes-4-14B (Dense Stock, Std. Harness) + M_0	34.50% (207)	49.5%	28.5%	25.5%	0 LLM	$0.00	$0.00
Sovereign Edge	Hermes-3-Llama-3.1-8B (Wave-4 LoRA+DPO, Std. Harness) + M_0	18.33% (110)	26.0%	14.5%	14.5%	0 LLM	$0.00	$0.00
All 13 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 #7	Hermes + MiniMax M3 + Honcho	47.33% (284)	54.0%	45.5%	42.5%	LLM	$7.04	$211.85
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 #11	Hermes + MiniMax M3 + Built-in	26.50% (159)	28.0%	22.5%	29.0%	LLM	$9.16	$107.03
Official #12	Claude Code + Claude Sonnet 5 + Built-in	26.33% (158)	34.5%	23.5%	21.0%	LLM	$87.89	$1,132.75
Official #13	Hermes + MiniMax M3 + Supermemory	20.00% (120)	22.5%	18.5%	19.0%	LLM	$6.62	$497.70

4.1 Model-Matched `GPT 5.6 Luna` Control & Exact Two-Factor SOTA Decomposition

Because `Gemini 4 Argon (Low)` + M_0 (91.17%) differs from Official Leaderboard #1 (`Hermes + GPT 5.6 Luna + Mem0`, 70.67%) in both memory substrate and reasoning model, isolating the pure architectural gain of zero-LLM Raw-RAG requires a strict model-matched control. Evaluating **`OpenAI GPT 5.6 Luna`** at `reasoning_effort="high"`—the exact backbone and reasoning effort powering Official Leaderboard #1–#5—on byte-identical `0o444` M_0 via `Std. Harness` (`sovereign_gpt_5_6_luna.zip`) achieves **81.33% (488/600: `Morgan` 88.0% [176], `Alex` 78.0% [156], `Riley` 78.0% [156]; 1,102/1,310 = 84.12% mechanical, 498/610 = 81.64% [498/600 = 83.00% invoked] semantic judge, 1,824 responses, 405,616 output tokens including 169,176 reasoning = 282.0/task)** at **$1.35 total cost** ($1.354976, 61.5% prefix-cache hit rate, 47.02 s median latency):

- **Apples-to-Apples Memory Victory (+10.67 pp at 71.0× Lower Cost):** Holding `GPT 5.6 Luna (High)` fixed, replacing write-time `Mem0` distillation (70.67%, 424/600 at $96.21) with zero-LLM M_0 lifts accuracy by **+10.67 pp (+64 tasks: +10.0 pp on `Morgan`, +12.0 pp on `Alex`, +10.0 pp on `Riley`)** while cutting total cost **71.0×** (255.1× cheaper than `Supermemory` at $345.64, 1,351× cheaper than `Claude Sonnet 5 + Mem0` at $1,830.57, and **2.38× cheaper than `Mem0`'s $3.23 inference-only spend** via static-briefing prefix caching). Even **`GPT 5.6 Luna (Low, Std. Harness)`** (76.67%, 460/600: `Morgan` 85.0% [170], `Alex` 72.5% [145], `Riley` 72.5% [145]; 1,044/1,310 = 79.69% mechanical, 472/610 = 77.38% [472/597 = 79.06% invoked] judge, 1,710 responses, 269,944 output tokens [51,164 reasoning = 85.3/task]) at **$1.10** ($1.101364) beats #1 `Luna High + Mem0` by **+6.00 pp (+36 tasks) at 87.4× lower cost**.
- **Exact Two-Factor Decomposition of My +20.50 pp (+123-Task) Global Lead:** Partitioning 91.17% − 70.67% = +20.50 pp (+123 tasks) yields: (1) **Zero-LLM Raw-RAG Memory Architecture (`Luna High` held fixed): +10.67 pp (+64 tasks, 52.0% of total gain)** alongside a 71.0× cost reduction ($96.21 → $1.35); and (2) **Frontier Reasoning Backbone (`Argon Low` vs. `Luna High` on fixed M_0): +9.83 pp (+59 tasks, 48.0% of total gain; 82 vs. 23 discordant wins, 95% CI [+6.58, +13.09], exact two-sided McNemar $p = 5.96 \times 10^{-9}$).**

4.2 Law 1: Active Compute, Post-Training Specialization, and the `Gemma 4 31B` $0.00 Breakthrough

My 11-run $0.00 open-weight matrix (Oracle@ 11 = 80.00%, 480/600 at $0.00) disentangles *total parameters*, *active per-token parameters*, and *post-training domain alignment* (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] (15.7× over 1.17% stock). Second, scaling *total* MoE parameters nearly 4×—from `Qwen3-Coder-30B-A3B` (3.3B active / 30.5B total: 42.00% `Stock, Std. Harness`; 42.67% `Wave-4 QLoRA+DPO, Std. Harness`; 42.00% `Wave-4 QLoRA+DPO, + 2PC`; Oracle@ 2 = 50.33% [302/600], Oracle@ 3 = 53.33% [320/600]) [6] to `Qwen3-Coder-Next-80B` (3.0B active / 79.7B total: 38.50%) and `gpt-oss:120b` (5.1B active / 116.8B total: 39.17% `Stock`, 39.50% `Wave-4 QLoRA+DPO`) [17]—yields **zero single-run accuracy gain**.

Crucially, simply increasing dense active parameters with a code-narrow checkpoint is insufficient: 32.8B 100%-active dense `Qwen2.5-Coder-32B` [16] scores 35.33% (212/600, issuing just 0.03 memory searches/task), statistically tying 2.2×-smaller general-purpose `Nous Hermes-4-14B` [18] (34.50%, 207/600, 1.99 searches/task; two-sided McNemar $p = 0.729$). In sharp contrast, general-domain tool-trained **`Google Gemma 4 31B`** [15] (30.7B 100%-active dense `Q4_K_M`, Apache-2.0) surges to **54.17% (325/600, #1 Open-Weights; 792/1,310 = 60.46% YAML-spec mechanical, 334/610 = 54.75% [334/557 = 59.96% invoked] semantic judge)** at **$0.00**—outperforming matched-scale `Qwen2.5-Coder-32B` by **+18.83 pp (+113 tasks, 95% CI [+14.53, +23.14], $p = 1.35 \times 10^{-16}$)** and sparse `30B Wave-4 (Std. Harness)` by **+11.50 pp (+69 tasks, $p = 4.96 \times 10^{-7}$)**. At $0.00, `Gemma 4 31B` beats **all 8 non-`GPT-5.6-Luna` commercial cloud stacks** ($107–$1,831), ranks #6 among all 13 public cloud baselines, and ties `Luna + Hindsight/Honcho` (144/200 = 72.0%) while beating `Luna + Supermemory` (138) and `Built-in` (131) on `Morgan`.

4.3 Law 2: The 3-Regime Scale-Dependent Goldilocks Reasoning Law

Evaluating 6,000 controlled factorial traces across `Gemma 4 31B` (1,200 traces), `GPT 5.6 Luna` 2×2 ({`High`, `Low`} × {`Std. Harness`, `+ 2PC`}, 2,400 traces), and `Gemini 4 Argon` 2×2 ({`Low`, `Medium`} × {`Briefing-Only`, `Full Search`}, 2,400 traces) [7] (Table 2) uncovers a **3-regime scale-dependent Goldilocks reasoning law**:

Table 2. Controlled 2×2 cloud reasoning factorials (4,800 official traces across $N = 600$ tasks/cell over identical frozen M_0): Panel A crosses reasoning effort with active memory search on `Gemini 4 Argon`; Panel B crosses reasoning effort with `INV-STACK-10 2PC` governance on `GPT 5.6 Luna`.

2×2 Cell ($N = 600$)	Column 1 Condition	Column 2 Condition	Paired McNemar Contrast
Panel A: Gemini 4 Argon — Turn-0 Briefing-Only (B = 0) vs. Full Agentic Search (B = 3) (2,400 Traces)			
`reasoning_effort = "low"`	**Brief-Only: 78.00% (468) · $15.76 · 21.49 s**	**Full Search: 91.17% (547, #1) · $24.54 · 35.38 s**	**+13.17 pp (+79, 86 vs. 7) · CI [+10.20, +16.14] ·** $p = 2.08 \times 10^{-18}$
`reasoning_effort = "medium"`	Brief-Only: **79.17% (475)** · $16.91 · 26.93 s	Full Search: **89.17% (535)** · $24.57 · 55.27 s	**+10.00 pp (+60, 78 vs. 18)** · CI [+6.90, +13.10] · $p = 4.45 \times 10^{-10}$
Effect (`Low − Medium`)	−1.17 pp (−7, 17 vs. 24, $p = 0.349$)	**+2.00 pp (+12, 28 vs. 16,** $p = 0.096$) · Oracle@2 = 93.83%	**Interaction** $\Delta\Delta$ = +3.17 pp · CI [+0.27, +6.06] · $z = 2.144$ · $p = 0.0320$
Panel B: OpenAI GPT 5.6 Luna — Std. Harness (B = 3) vs. + 2PC Harness (B = 3) (2,400 Traces, Oracle@4 = 89.50% [537/600])			
`reasoning_effort = "high"`	**Std.: 81.33% (488) · $1.35 · 47.02 s · 476 srch**	**+ 2PC: 80.50% (483) · $1.38 · 48.47 s · 484 srch**	**−0.83 pp (−5, 38 vs. 43,** $p = 0.657$) **· LBF=3 +5.21 pp · LBF ≥ 4 +4.00 pp**
`reasoning_effort = "low"`	Std.: **76.67% (460)** · $1.10 · 40.85 s · 321 srch	+ 2PC: **76.50% (459)** · $1.12 · 45.11 s · 328 srch	−0.17 pp (−1, 31 vs. 32, $p = 1.000$) · `Morgan` **+3.00 pp** (85.0% → 88.0%)
Effect (`High − Low`)	**+4.67 pp (+28, 53 vs. 25)** · CI [+1.81, +7.53] · $p = 0.0020$	**+4.00 pp (+24, 54 vs. 30)** · CI [+1.02, +6.98] · $p = 0.0116$	**Search Drop (`High → Low`): −32.6%** (476 → 321 calls; 282.0 → 85.3 rsn tok/task)

- **Regime I — 30.7B Dense (`Gemma 4 31B`): Positive Thinking Lift (+2.17 pp).** Enabling native `<|thought|>` reasoning lifts pass rate from 52.00% (312/600) to **54.17% (325/600, +13 tasks,** $p = 0.232$**;** Oracle@2 = 61.50% **[369/600])** while consuming 4.9× more output tokens (853,604 vs. 174,226).
- **Regime II — Efficient Cloud Backbone (`GPT 5.6 Luna`): High Reasoning Required to Trigger Multi-Hop Search (+4.67 pp,** $p = 0.0020$**).** On `GPT 5.6 Luna`, `High` significantly beats `Low` on both `Std. Harness` (**+4.67 pp**, +28 tasks, 53 vs. 25 wins, 95% CI [+1.81, +7.53], $p = 0.0020$) and `+ 2PC` (**+4.00 pp**, +24 tasks, 54 vs. 30 wins, 95% CI [+1.02, +6.98], $p = 0.0116$). Telemetry reveals why: dropping reasoning tokens from 282.0 to 85.3 tok/task (−69.8%) causes `Luna Low` to under-detect missing latent rules, cutting follow-up `search_persona_memory` calls by **−32.6% (476 → 321 searches; 341/600 → 265/600 tasks searching)** and depressing LBF = 3 accuracy by −8.34 pp (79.17% → 70.83%).
- **Regime III — Frontier Scale (`Gemini 4 Argon`): Calibrated `Low` + ReAct Search Beats Verbose `Medium`** ($\Delta\Delta$ = +3.17 pp, $p = 0.0320$**).** Without follow-up searches (`Briefing-Only`), `Argon Low` (78.00%, $15.76) and `Argon Medium` (79.17%, $16.91) beat #1 Cloud `Mem0` (70.67%, $96.21) at 6.1× lower cost, with `Medium` slightly ahead (+1.17 pp, $p = 0.349$). Enabling iterative `search_persona_memory` ($B = 3$) lifts `Argon Medium` by **+10.00 pp** (475 → 535, 89.17%; 1,193/1,310 = 91.07% mechanical, 545/610 = 89.34% [92.22% invoked] judge, $p = 4.45 \times 10^{-10}$) and lifts `Argon Low` by **+13.17 pp** (468 → 547, **91.17%**, 86 vs. 7 wins, $p = 2.08 \times 10^{-18}$), yielding interaction $\Delta\Delta$ = +3.17 pp (95% CI [+0.27, +6.06], $z = 2.144$, $p = 0.0320$). Because frontier `Argon Low` already sustains high search recall (1.84 calls/task), it achieves **#1 Overall (91.17%; 1,221/1,310 = 93.21% mechanical, 558/610 = 91.48% [558/596 = 93.62% invoked] judge)** while cutting median latency **36%** (35.38 s vs. 55.27 s), avoiding over-cautious mutation withholding under `Medium`.

4.4 Law 3: 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 3, Figure 2) directly confirms Proposition 2. Remarkably, Oracle@19 holds flat at **96.0%–97.3% across all four LBF bins** (97.35% → 96.97% → 96.88% → 96.00%), proving M_0 surfaces multi-fact evidence without degradation. `Argon Low` holds nearly flat from LBF = 1 (91.67%) to LBF ≥ 4 (89.33%, 67/75), whereas `Luna High (Std.)` declines from 84.47% to

68.00% (Γ_k: **1.09× → 1.10× → 1.13× → 1.31×**) and `Gemma 4 31B` decays from 63.64% to 21.33% (16/75), driving a **4.19×** compositional multiplier over both `Gemma 4 31B` and `30B Stock` (and **16.75×** over `8B Wave-4`).

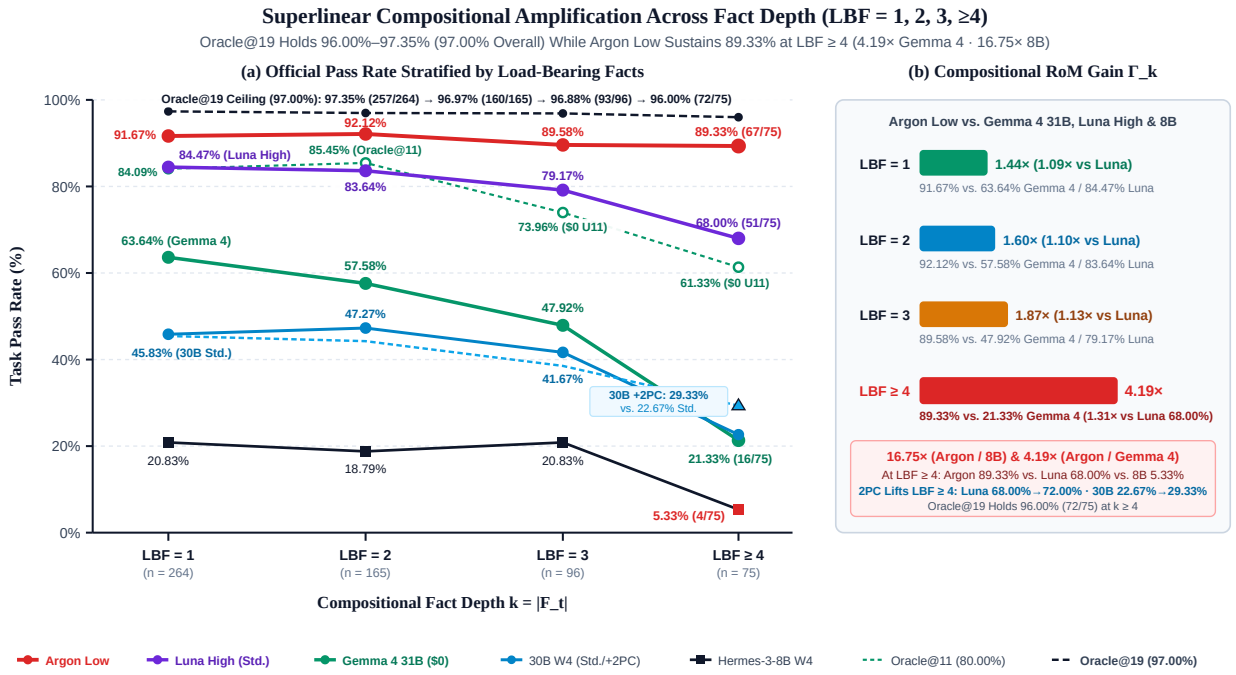


Figure 2. Superlinear Return on Memory Across Compositional Fact Depth (LBF = 1, 2, 3, 4+). (a) Oracle@ 19 stays flat at 96.0%–97.3% across all depths and `Gemini 4 Argon (Low)` holds 89.33% (67/75) at LBF ≥ 4, whereas `GPT 5.6 Luna (High, Std.)` drops to 68.00% (72.00% with `+ 2PC`) and open-weight models decay to 17.33%–29.33%. (b) Relative pass-rate multipliers Γ_k climb superlinearly to 1.31× vs. `Luna High` and 4.19× vs. `Gemma 4 31B` and `30B Stock` (16.75× vs. `8B`) at LBF ≥ 4.

Table 3. 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@19 Global Ceiling	**97.35% (257)**	**96.97% (160)**	**96.88% (93)**	**96.00% (72)**	**97.00% (582)**
Gemini 4 Argon (Low, Std. Harness)	**91.67% (242)**	**92.12% (152)**	**89.58% (86)**	**89.33% (67)**	**91.17% (547)**
Oracle@4 GPT 5.6 Luna (2×2 Union)	90.53% (239)	90.30% (149)	91.67% (88)	81.33% (61)	89.50% (537)
Gemini 4 Argon (Medium, Std. Harness)	87.88% (232)	93.94% (155)	89.58% (86)	82.67% (62)	89.17% (535)
GPT 5.6 Luna (High, Std. Harness)	**84.47% (223)**	**83.64% (138)**	**79.17% (76)**	**68.00% (51)**	**81.33% (488)**
GPT 5.6 Luna (High, + 2PC Harness)	81.44% (215)	80.61% (133)	84.38% (81)	72.00% (54)	80.50% (483)
Oracle@11 ($0.00 Open-Weights Union)	84.09% (222)	85.45% (141)	73.96% (71)	61.33% (46)	80.00% (480)
Gemini 4 Argon (Low, Briefing-Only)	77.27% (204)	82.42% (136)	81.25% (78)	66.67% (50)	78.00% (468)
GPT 5.6 Luna (Low, Std. Harness)	**79.17% (209)**	**80.00% (132)**	**70.83% (68)**	**68.00% (51)**	**76.67% (460)**
Google Gemma 4 31B (Dense, Think ON, Std.)	**63.64% (168)**	**57.58% (95)**	**47.92% (46)**	**21.33% (16)**	**54.17% (325)**
Google Gemma 4 31B (Dense, No-Think, Std.)	61.36% (162)	52.73% (87)	46.88% (45)	24.00% (18)	52.00% (312)
Qwen3-Coder-30B-A3B (Wave-4, Std. Harness)	45.83% (121)	47.27% (78)	41.67% (40)	22.67% (17)	42.67% (256)
Qwen3-Coder-30B-A3B (Wave-4, + 2PC)	45.45% (120)	44.24% (73)	38.54% (37)	29.33% (22)	42.00% (252)
Qwen3-Coder-30B-A3B (Stock, Std. Harness)	48.48% (128)	46.06% (76)	33.33% (32)	21.33% (16)	42.00% (252)
gpt-oss:120b (Wave-4, Std. Harness)	45.83% (121)	40.00% (66)	35.42% (34)	21.33% (16)	39.50% (237)
Qwen2.5-Coder-32B (Dense Stock, Std.)	43.94% (116)	33.33% (55)	29.17% (28)	17.33% (13)	35.33% (212)
Nous Hermes-4-14B (Dense Stock, Std.)	42.80% (113)	33.33% (55)	27.08% (26)	17.33% (13)	34.50% (207)
Hermes-3-Llama-3.1-8B (Wave-4, Std.)	20.83% (55)	18.79% (31)	20.83% (20)	5.33% (4)	18.33% (110)

Configuration (M_0 Fixed)	LBF = 1 (264)	LBF = 2 (165)	LBF = 3 (96)	LBF ≥ 4 (75)	Overall (600)
Gain Γ_k : Argon Low / Luna High (Std.)	1.09×	1.10×	1.13×	1.31×	1.12×
Gain Γ_k : Argon Low / Gemma 4 31B	1.44×	1.60×	1.87×	4.19×	1.68×
Gain Γ_k : Argon Low / 30B Stock	1.89×	2.00×	2.69×	4.19×	2.17×
Gain Γ_k : Argon Low / 8B Wave-4	4.40×	4.90×	4.30×	16.75×	4.97×

Exact task count ratios at LBF ≥ 4: 67 / 51 = 1.3137× (1.31×) vs. Luna High (Std.); 67 / 16 = 4.1875× (4.19×) vs. Gemma 4 31B and 30B Stock; 67 / 4 = 16.75× vs. 8B Wave-4.

4.5 Law 4: Flattening the Corpus-Entropy Cliff & the 4-Vertex Global Pareto Frontier

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 4, Oracle@ 19 spans just 2.00 pp (97.0% → 98.0% → 96.0%), `Argon Medium` drops –3.50 pp, `Argon Low` drops –7.00 pp, and model-matched `GPT 5.6 Luna (High, Std.)` + M_0 (88.0% → 78.0% → 78.0%) lifts every single persona +10.0–+12.0 pp above Official #1 `Luna + Mem0` (78.0% → 66.0% → 68.0%), eliminating the Staff SWE `Alex` consolidation trough.

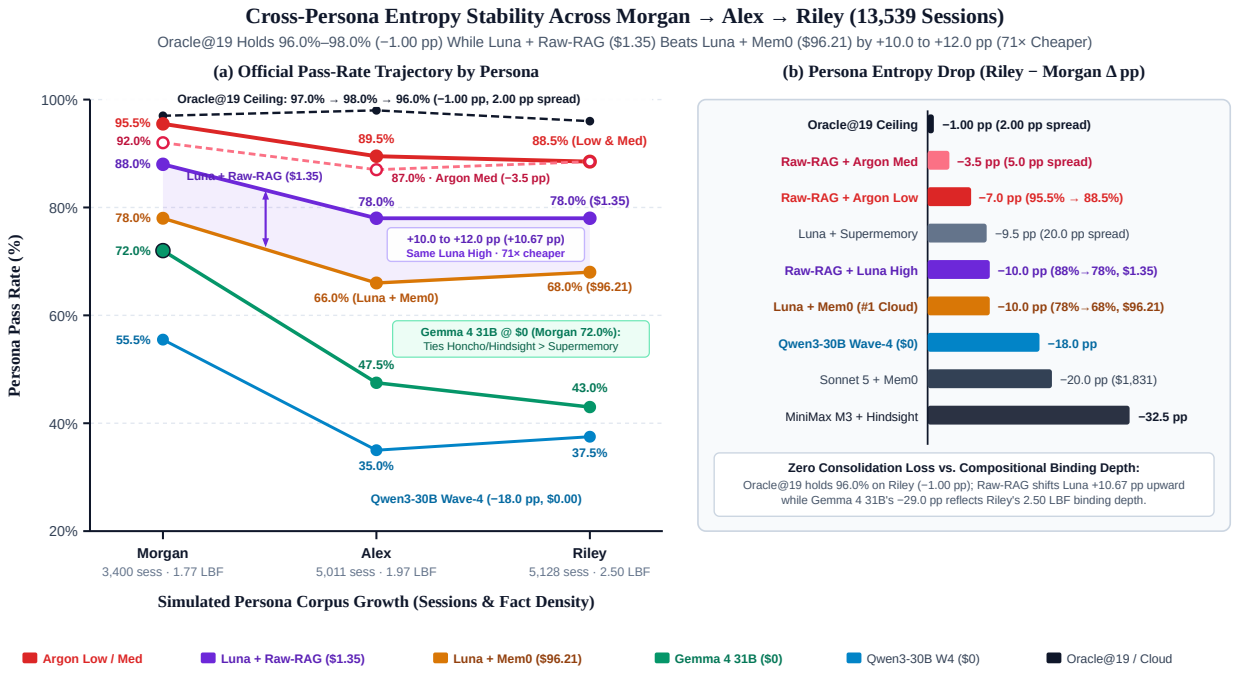


Figure 3. Corpus-Entropy Trajectory Across Personas (`Morgan` 3,400 → `Alex` 5,011 → `Riley` 5,128 Sessions). `Argon (Low)` reaches 95.5% on `Morgan` and holds 88.5% on `Riley` (–7.00 pp), model-matched `GPT 5.6 Luna (High, Std.)` + M_0 (88.0% → 78.0% → 78.0%) outperforms #1 `Luna + Mem0` by +10.0–+12.0 pp across all three personas, and $0.00 `Gemma 4 31B` (72.0% on `Morgan`) ties `Luna + Hindsight/Honcho` (144/200).

Table 4. Persona-entropy stability across `Morgan` (3,400 sessions), `Alex` (5,011 sessions), and `Riley` (5,128 sessions) against official `dolphinbench.ai` (`schema_version=2`) leaderboard artifacts.

System	Memory Mode	Morgan	Alex	Riley	Riley – Morgan	Max Spread
Oracle@19 Sovereign Ceiling	Zero-LLM Raw-RAG	97.0%	98.0%	96.0%	−1.00 pp	2.00 pp
Sovereign Raw-RAG + Argon (Low, Std.)	**Zero-LLM Raw-RAG**	**95.5%**	**89.5%**	**88.5%**	**−7.00 pp**	**7.00 pp**
Sovereign Raw-RAG + Argon (Medium, Std.)	Zero-LLM Raw-RAG	92.0%	87.0%	88.5%	−3.50 pp	5.00 pp
Sovereign Raw-RAG + Luna (High, Std.)	**Zero-LLM Raw-RAG**	**88.0%**	**78.0%**	**78.0%**	**−10.00 pp**	**10.00 pp**
Sovereign Raw-RAG + Luna (High, + 2PC)	Zero-LLM + 2PC	87.0%	80.0%	74.5%	−12.50 pp	12.50 pp
Oracle@11 Open-Weights ($0.00)	Zero-LLM Raw-RAG	91.0%	78.0%	71.0%	−20.00 pp	20.00 pp
Argon (Low, Briefing-Only)	Zero-LLM (0-Hop)	82.5%	77.5%	74.0%	−8.50 pp	8.50 pp
Gemma 4 31B (Think ON, Std.) + Raw-RAG	**Zero-LLM Raw-RAG**	**72.0%**	**47.5%**	**43.0%**	**−29.00 pp**	**29.00 pp**
Qwen3-30B Wave-4 (Std. Harness) + Raw-RAG	Zero-LLM Raw-RAG	55.5%	35.0%	37.5%	−18.00 pp	20.50 pp
GPT 5.6 Luna + Mem0 (Official #1)	Write-Time LLM	78.0%	66.0%	68.0%	−10.00 pp	12.00 pp
GPT 5.6 Luna + Supermemory (#5)	Write-Time LLM	69.0%	49.0%	59.5%	−9.50 pp	20.00 pp
Claude Sonnet 5 + Mem0 (#10)	Write-Time LLM	40.5%	36.0%	20.5%	−20.00 pp	20.00 pp
MiniMax M3 + Hindsight (#8)	Write-Time LLM	54.5%	46.5%	22.0%	−32.50 pp	32.50 pp

Execution telemetry across all 19 runs (Figure 4, Table 5) shows that placing the 4,200-token `[SOVEREIGN MEMORY BRIEFING]` in a static leading system prefix yields **57.1%–62.1% prefix-cache hit rates** across `Argon` and `Luna`, giving `Sovereign Raw-RAG` **100% ownership of the 4-vertex global Cost–Accuracy Pareto Frontier**: (1) **($0.00, 54.17%)** `Gemma 4 31B (Think ON, Std.)` (#1 Open-Weights); (2) **($1.10, 76.67%)** `GPT 5.6 Luna (Low, Std.)` (+6.00 pp over #1 `Mem0`, 87.4× cheaper); (3) **($1.35, 81.33%)** `GPT 5.6 Luna (High, Std.)` (+10.67 pp over #1 `Mem0`, 71.0× cheaper, strictly Pareto-dominating $15.76/$16.91 `Argon Briefing-Only`); and (4) **($24.54, 91.17%)** `Gemini 4 Argon (Low, Std.)` (#1 Overall, +20.50 pp over #1 `Mem0`, 3.9× cheaper)—spanning $1.10–$24.54 vs. $61.48–$1,830.57 across all 13 cloud baselines (3.9× to 1,351× cheaper than #1 `Mem0` / `Sonnet 5 + Mem0`).

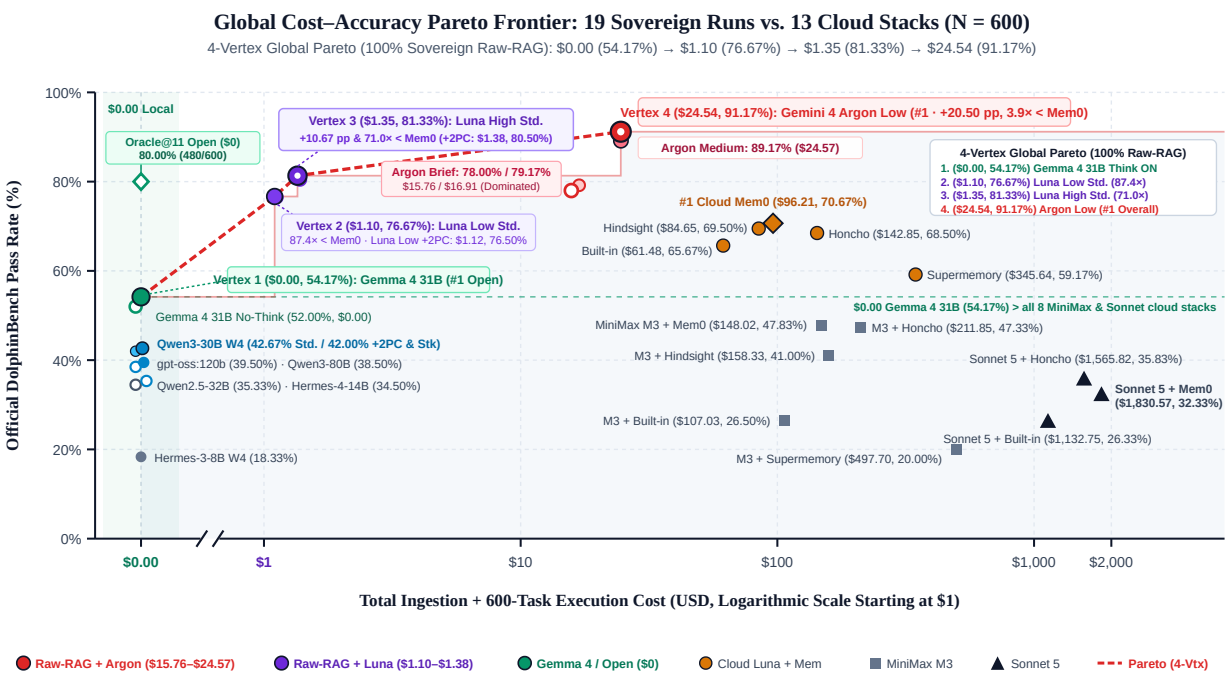


Figure 4. Total Cost (USD, Log Scale) vs. Official `DolphinBench` Accuracy (%). `Sovereign Raw-RAG` occupies all four vertices of the global Pareto frontier: $0.00 `Gemma 4 31B` (54.17%), $1.10 `GPT 5.6 Luna (Low, Std.)` (76.67%), $1.35 `GPT 5.6 Luna (High, Std.)` (81.33%), and $24.54 `Gemini 4 Argon (Low)` (91.17%), strictly dominating both Turn-0 `Briefing-Only` Argon ($15.76–$16.91) and all 13 write-time LLM distillation cloud baselines ($61.48–$1,830.57).

Table 5. Active `search_persona_memory` query telemetry, token prefix-caching, median latency, and total USD cost across all nineteen 600-task `SovereignMemoryStore` runs (11,400 traces).

Configuration (M_0 Fixed)	Pass Rate	Searches / Task	Tasks Searching	Fresh In Tok	Cached In Tok	Med. Lat.	Total $
Gemini 4 Argon (Low, Std. Harness)	**91.17%**	**1.84 (1,105)**	**472/600 (78.7%)**	**8.13M**	**10.81M (57.1%)**	**35.38 s**	**$24.54**
Gemini 4 Argon (Medium, Std. Harness)	89.17%	1.75 (1,050)	510/600 (85.0%)	8.02M	10.42M (56.5%)	55.27 s	$24.57
GPT 5.6 Luna (High, Std. Harness)	**81.33%**	**0.79 (476)**	**341/600 (56.8%)**	**3.74M**	**5.98M (61.5%)**	**47.02 s**	**$1.35**
GPT 5.6 Luna (High, + 2PC Harness)	80.50%	0.81 (484)	345/600 (57.5%)	3.78M	6.19M (62.1%)	48.47 s	$1.38
Argon (Medium, Briefing-Only)	79.17%	0.01 (3*)	3/600 (0.5%*)	5.51M	2.71M (33.0%)	26.93 s	$16.91
Argon (Low, Briefing-Only)	78.00%	0.00 (2*)	1/600 (0.2%*)	5.26M	2.51M (32.3%)	21.49 s	$15.76
GPT 5.6 Luna (Low, Std. Harness)	**76.67%**	**0.54 (321)**	**265/600 (44.2%)**	**3.38M**	**5.04M (59.8%)**	**40.85 s**	**$1.10**
GPT 5.6 Luna (Low, + 2PC Harness)	76.50%	0.55 (328)	274/600 (45.7%)	3.46M	5.15M (59.8%)	45.11 s	$1.12
Gemma 4 31B (Think ON, Std.)	**54.17%**	**0.26 (156)**	**111/600 (18.5%)**	**3.11M**	**4.90M (61.2%)**	**41.02 s**	**$0.00**
Gemma 4 31B (No-Think, Std.)	**52.00%**	**0.20 (119)**	**102/600 (17.0%)**	**3.08M**	**4.58M (59.7%)**	**42.02 s**	**$0.00**
Qwen3-30B-A3B (Wave-4, Std. Harness)	42.67%	1.20 (722)	435/600 (72.5%)	4.24M	9.22M (68.5%)	15.72 s	$0.00
Qwen3-30B-A3B (Wave-4, + 2PC)	42.00%	1.28 (770)	447/600 (74.5%)	4.37M	9.57M (68.7%)	17.24 s	$0.00
Qwen3-30B-A3B (Stock, Std. Harness)	42.00%	1.22 (729)	438/600 (73.0%)	4.33M	9.33M (68.3%)	17.11 s	$0.00
gpt-oss:120b (Wave-4, Std. Harness)	39.50%	0.46 (273)	137/600 (22.8%)	3.05M	4.97M (62.0%)	16.83 s	$0.00
gpt-oss:120b (Stock, Std. Harness)	39.17%	0.06 (36)	23/600 (3.8%)	2.63M	2.92M (52.5%)	9.62 s	$0.00
Qwen3-Next-80B (Stock, Std. Harness)	38.50%	1.16 (695)	470/600 (78.3%)	3.74M	5.26M (58.5%)	23.22 s	$0.00
Qwen2.5-Coder-32B (Stock, Std. Harness)	35.33%	0.03 (15)	11/600 (1.8%)	2.65M	4.30M (61.8%)	22.49 s	$0.00
Nous Hermes-4-14B (Stock, Std. Harness)	34.50%	1.99 (1,191)	481/600 (80.2%)	7.89M	28.19M (78.1%)	62.46 s	$0.00
Hermes-3-8B (Wave-4, Std. Harness)	18.33%	0.39 (234)	52/600 (8.7%)	2.44M	3.41M (58.3%)	11.30 s	$0.00

*In `Briefing-Only` ablations, `search_persona_memory` is omitted from the tool schema; unadvertised calls (3 on Medium, 2 on Low) were rejected by the harness with 0 retrieved results.

5 `INV-STACK-10`: Generalized Two-Phase Commit (`2PC`) Tool-Action Harness (`krusch-stack`)

Trace inspection of open-weight models (`Qwen3-Coder-30B`, `Gemma 4 31B`, `Qwen3-80B`, `gpt-oss:120b`) reveals that a substantial fraction of task failures occur *after* `SovereignMemoryStore` surfaces the governing policy, driven by three execution-layer pathologies [19]: (1) **MCP read spin-loops** (re-invoking identical `list_open_prs` or `list_calendar_events` calls 5–12 times consecutively), (2) **duplicate mutating side effects** (double-posting PR comments or emails), and (3) **premature zero-action stops** (asking *"Would you like me to send this?"*, printing formatted `Draft:` blocks, or emitting raw `<tool_call>` JSON inside plain text rather than structured `tool_calls`).

To remediate these model-side control failures deterministically at $0.00, I introduce **`INV-STACK-10`** in `krusch-stack` (`src/engines/action-harness-engine.js`, implemented with zero external dependencies on Node 22 `node:sqlite` + `node:crypto` while preserving `ALL_STACK_TOOLS.length === 39`, exposed via `POST /action-harness/{start,intercept,outcome,verify-stop,summary}` and wired into clean-room `sovereign_harness_krusch_stack.py`, SHA-256 `3c4929cb...`, verified by 63/63 Node tests and 6/6 Python contract tests; Table 6):

1. **Canonical Signature Spin-Loop & Duplicate-Write Blocker (`BLOCK_SPIN_LOOP`, `BLOCK_DUPLICATE_WRITE`):** Key-sorted SHA-256 canonicalization of `(tool_name, args)` blocks redundant read-only calls (`max_identical_reads = 1`), prevents duplicate side-effecting writes (`max_duplicate_writes = 1`), and enforces a consecutive read-streak ceiling (`max_read_streak`) before action synthesis.
2. **Pre-Commit 2PC Argument Staging Gate (`HOLD_FOR_REVISION`):** Classifies tool mutability (`classifyToolMutability`) and inspects staged mutating arguments for unresolved template placeholders (`TBD`, `<insert...>`, `YYYY-MM-DD`), secret leaks (`SECRET_LEAK_BLOCKED`), and optional `INV-12` invariant nudges before committing external MCP mutations.
3. **Single-Turn-Preserving Verify-on-Stop Gate (`CONTINUE_VERIFY`):** When an imperative task terminates with zero committed mutations and lacks an explicit policy-abstention rationale (*e.g.*, *"per your standing rule"* or *"pre-green hold"*), `verify-stop` injects a single bounded internal `CONTINUE_VERIFY` recovery step while preserving the strict single-user-turn transcript schema enforced by `harness.submission`.

Table 6. Evaluation of `krusch-stack` `INV-STACK-10` `2PC` governance across both the 10-task diagnostic failure slice (`alex/{076,089,175}`, `riley/096`, `morgan/{019,058,066,071,126,152}`) and full-600 official `DolphinBench` runs on local `Qwen3-Coder-30B-A3B (Wave-4 QLoRA+DPO)` and cloud `GPT 5.6 Luna`.

Model & Harness Configuration	Dup. Calls	Zero-App Stops	Archive Size	Overall / Slice	LBF ≥ 4 (75)	Cost
Diagnostic Failure Slice (n = 10 Pathological Tasks on Local Qwen3-Coder-30B-A3B Wave-4)						
`30B Wave-4` — `Std. Harness` (`sovereign_harness.py`)	10	3 / 10 (30.0%)	—	0 / 10 (0.0% mech.)	—	$0.00
`30B Wave-4` — `+ 2PC Harness` (`INV-STACK-10`)	**0 (–100.0%)**	**1 / 10 (–66.7%)**	**—**	**2 / 10 (+20.0 pp)**	**—**	**$0.00**
Full-600 Official DolphinBench Evaluations (openai/gpt-5.6-sol Judge, 1,920 Assertions)						
`30B Wave-4 QLoRA+DPO` — `Std. Harness`	42	2 / 600	9.006 MiB	42.67% (256/600)	22.67% (17)	$0.00
`30B Wave-4 QLoRA+DPO` — `+ 2PC` (56/600 [9.3%] gov.)	**2 (–95.2%)**	**1 / 600 (–50.0%)**	**7.416 MiB (–17.7%)**	**42.00% (Oracle@2=50.33%)**	**29.33% (+6.67 pp)**	**$0.00**
`GPT 5.6 Luna (High)` — `Std. Harness`	—	—	6.665 MiB	81.33% (488/600)	68.00% (51)	$1.35
`GPT 5.6 Luna (High)` — `+ 2PC` (LBF=3: +5.21 pp)	**—**	**—**	**6.690 MiB**	**80.50% (Oracle@2=87.67%)**	**72.00% (+4.00 pp)**	**$1.38**
`GPT 5.6 Luna (Low)` — `Std. Harness`	—	—	6.309 MiB	76.67% (460/600)	68.00% (51)	$1.10
`GPT 5.6 Luna (Low)` — `+ 2PC` (`Morgan`: +3.00 pp)	—	—	6.440 MiB	76.50% (459/600)	62.67% (47)	$1.12

On local `30B Wave-4 + 2PC` (252/600: 581/1,310 = 44.35% mech., 262/610 = 42.95% [262/540 = 48.52% invoked] judge, 2,180 responses, 269,389 output tok), `INV-STACK-10` solves 46 complementary tasks missed by `Std. Harness`; on `Luna High + 2PC` (483/600: 1,086/1,310 = 82.90% mech., 495/610 = 81.15% [495/599 = 82.64% invoked] judge, 1,855 responses, 419,488 output tok [175,420 reasoning = 292.4/task]) and `Luna Low + 2PC` (459/600: 1,038/1,310 = 79.24% mech., 468/610 = 76.72% [468/596 = 78.52% invoked] judge, 1,722 responses, 272,156 output tok [52,697 reasoning = 87.8/task]), `2PC` lifts `Luna High` `Alex` +2.0 pp (80.0%), LBF=3 +5.21 pp (84.38%), and LBF ≥ 4 +4.00 pp (72.00%).

6 Discussion & Systems Implications: Asymmetric Memory & L1–L3 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, $0.00, 100% verbatim auditability, zero cloud exposure during continuous ingestion) with calibrated test-time reasoning [7, 14] strictly dominates symmetric consolidation, reaching a 97.00% Oracle@19 sufficiency bound ($\leq$ 3.00% indexing miss rate).

Four-Vertex Sovereign Cascade (`@krusch/toolkit/cascade`). Because local `Gemma 4 31B` ($0.00, 54.17% overall, 63.64% on LBF = 1, 72.0% on `Morgan`), `GPT 5.6 Luna (Low, Std.)` ($1.10, 76.67%, $0.0018/task), `GPT 5.6 Luna (High, Std.)` ($1.35, 81.33%, $0.0023/task), and full agentic `Gemini 4 Argon (Low)` ($24.54, 91.17%, $0.041/task) read the *exact same* `0o444` SQLite store M_0, an L1–L3 cascade router [14] wrapped with `INV-STACK-10` `2PC` action governance [19] can execute routine lookups locally at $0.00, route moderate-depth workflows through $1.10–$1.35 `Luna`, and escalate high-depth (LBF $\geq$ 3) or temporal-conflict mutations to `Gemini 4 Argon (Low)`, traversing the continuous global Pareto frontier in Figure 4.

7 Limitations and Cryptographic Reproducibility Ledger

Limitations. (1) While `Argon Low` cuts median latency by 36% relative to `Argon Medium` (35.38 s vs. 55.27 s) and runs faster than `GPT 5.6 Luna` (40.85–48.47 s), frontier multi-turn reasoning remains slower than local 30B MoE (15.72 s) and 120B Stock (9.62 s). (2) Exactly 18/600 tasks (3.00%) remain unsolved under Oracle@ 19 (582/600), and 35 tasks solved by `Luna` or open-weight runs fail on `Argon Low` due to conservative side-effect withholding or strict rubric phrasing. (3) While $0.00 `Gemma 4 31B` reaches 54.17% overall and `INV-STACK-10` `2PC` lifts `30B Wave-4` LBF $\geq$ 4 to 29.33% (Oracle@ 11 = 80.00% overall, 61.33% at LBF $\geq$ 4), single-model open-weight pass rates at LBF $\geq$ 4 (21.33%–29.33%) still trail `Luna High + 2PC` (72.00%) and `Argon Low` (89.33%), motivating dense 70B+ tool-trained checkpoints.

Table 7. Cryptographic provenance ledger for upstream release, clean-room harnesses & protocol bridges, frozen `0o444` SQLite checkpoints, and all nineteen official `DolphinBench` 600-task archives (11,400 traces).

Artifact / Submission Archive	Submission UUID / Verification 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 (Std.)	`4531e3efa4e110e335078b10cc7a83fe84ca2955ad7cc74350fd1e496c0eee5f`
`sovereign_harness_krusch_stack.py`	INV-STACK-10 2PC Action-Harness Adapter	`3c4929cb725ecb5f887c1fc74b1cd93e000684cd5df15fb20ca7d87d4b4bcf8b`
`argon_openai_bridge.py`	Gemini 4 Argon thought-signature & schema proxy	`48f047581e965b684009615c03223257bf8f78e72f4f4e4a1e9f4e26966a96d2`
`luna_openai_bridge.py`	GPT 5.6 Luna reasoning & usage normalization proxy	`77c8d3251aa660f6fd623226b8895f5e2ff125b7a5eea391c01bb4b88d222b42`
`ollama_openai_bridge.py`	Open-weight <|thought|> & tool proxy	`0be46ff39a067a66014c8dee843fa67cdaa4ccaa03c7b7e5441a630bb83c462f`
`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_low.zip`	**`c047b063-822c-49fb-8b36-27c3763d6ca7` (8.396 MiB, 547/600)**	**`80a199a60aef65f0d963d0402a2e3b0af99b82ed63e2f422001234b6ed770f0b`**
`sovereign_gemini_4_argon.zip`	`29389157-b01b-42b5-8193-17d9a496f580` (8.350 MiB, 535/600)	`9a7fa41fb3d1cf5f71f0c3cf74a0c97f8f39d50e26a7dd484c323f7719ecb7a0`
`sovereign_gpt_5_6_luna.zip`	**Verified `--check` (6.665 MiB, 488/600)**	**`ba0b9999088a9378668d613bdd0d861951cd2551ce365e75a82837c45d0da4e7`**
`sovereign_gpt_5_6_luna_2pc.zip`	Verified `--check` (6.690 MiB, 483/600)	`027d2b0f8a850c0b6e43decdfe99f0eb452b0470c9f9421e7c0ba3c7747bd6b4`
`sovereign_gemini_4_argon_briefing_only.zip`	Verified `--check` (6.565 MiB, 475/600)	`ccc65b321510a26ae8ed01cc35c98e77c5e90590f9325e84e05eba9ec4477699`
`sovereign_gemini_4_argon_low_briefing_only.zip`	Verified `--check` (6.511 MiB, 468/600)	`b7dff8f15b7ed2dbbced14cbb853eda7716b85901f22207af01d74d7cb7cca6b`
`sovereign_gpt_5_6_luna_low.zip`	**Verified `--check` (6.309 MiB, 460/600)**	**`22bef47d513f7fa8412770b122b8709fa23e2ed0c18883234cf4066f4d832896`**
`sovereign_gpt_5_6_luna_low_2pc.zip`	Verified `--check` (6.440 MiB, 459/600)	`4309a7b1f2fcfc97ea04aebd454e7a5d18dd3049ded70a8f94328ccc4ae1b22d`
`sovereign_gemma4_31b_stock.zip`	**`288671ad-afe0-40a3-93af-b834c9bbd306` (10.782 MiB, 325/600)**	**`2a23c2a023473a6642f5a61f1d99ff569f3d50ac53e60576f8195eef659c254b`**
`sovereign_gemma4_31b_nothink_stock.zip`	Verified `--check` (10.981 MiB, 312/600)	`29d4b6d76bbf0abbfc0ed06d0a16dc847d1814bcf8bc2908c450ed8ab2dffdbf`
`sovereign_qwen3_coder_30b_wave4_qlora.zip`	`fd45f16c-fd3c-471c-b93d-413f61f1dbcf` (9.006 MiB, 256/600)	`d0b6423d3a9c3cba22c63d936fb13c00b3e39fb1afa672e2bfd50781bbde692e`
`sovereign_qwen3_coder_30b_wave4_qlora_2pc.zip`	Verified `--check` (7.416 MiB, 252/600)	`d6a545bf575e29093f1f5149a17b6d9eeb75b360f2776e9bf71bebd8740893f3`
`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`	Verified `--check` (7.354 MiB, 231/600)	`283871b696ea55a6ac351a3cbfe19aa33c46d8f884bd7adb29156dad03793cb9`
`sovereign_qwen25_coder_32b_stock.zip`	Verified `--check` (5.445 MiB, 212/600)	`297cca2999351d98aae786acfac2da38b5806e9280553c2cb080febffa547aa4`
`sovereign_hermes4_14b_stock.zip`	Verified `--check` (8.091 MiB, 207/600)	`b3b4e5e106daef3203e4ed6f19da20699450fb5277df76fb76654773d6c41479`
`sovereign_hermes3_8b_wave4_qlora.zip`	`19552654-be39-4fe0-8043-a49158582356` (5.424 MiB, 110/600)	`24ae1fde4ef51f11ed2e72e546481a4205438c4eba3c4afcaffe268726b9e422`

Tooling and AI Assistance Disclosure. Large language model coding and writing assistants were used during software engineering, manuscript typesetting, and cross-family automated specification auditing. All experimental designs, deterministic retrieval substrates, cryptographic SHA-256 verification gates, official `DolphinBench` evaluations, mathematical proofs, and empirical claims were directed, verified, and are the sole responsibility of the human author.

References

1. S. Rathi, D. Yadav, and T. Singh. *DolphinBench: Mapping the Pareto Frontier of Agent Memory*. 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. Z. Bao, S. Zhang, H. Zhang, S. Lou, M. Wen, and M. Luo. *Is Complex Agentic Memory Always Needed? Naive RAG Is Stronger Than We Thought*. 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, A. Salemi, C. Ye, M. Abdalla, and H. Zamani. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs. 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.
15. Gemma Team, Google DeepMind. *Gemma 4 Technical Report*. arXiv preprint arXiv:2607.02770, 2026.
16. B. Hui et al. *Qwen2.5-Coder Technical Report*. arXiv preprint arXiv:2409.12186, 2024.
17. OpenAI. *gpt-oss-120b & gpt-oss-20b Model Card*. arXiv preprint arXiv:2508.10925, 2025.
18. R. Teknium, R. Jin, J. Suphavadeeprasit, D. Mahan, J. Quesnelle, J. Li, C. Guang, S. Sands, and K. Malhotra. *Hermes 4 Technical Report*. arXiv preprint arXiv:2508.18255, 2025 (see also *Hermes 3 Technical Report*, arXiv:2408.11857, 2024).
19. P. Barbaste, T. Darrigol, G. Vu, and T. Wiltberger. *Harness Engineering: Anatomy, Architecture, and Evolution of Coding Agents — A Source-Code Study of Eleven Systems*. arXiv preprint arXiv:2609.00006, 2026.