MEMORY OUTLIVES THE MODEL: TRANSPLANTING IN-ATTENTION MEMORY STATES ACROSS LLM BACKBONES

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent memory states inherit the interfaces of the models that write them. We separate the two. Agents now outlive the models beneath them, and their accumulated memories must travel across checkpoints, as text already does through a shared symbolic format. In a delta-rule memory, the accumulated state is a small $r \times r$ matrix per layer, independent of the writer's hidden size, tokenizer, and depth; only the read and write projections are model-specific. The state is a portable file. We present the first cross-backbone transplant of a fixed-size episodic in-attention state: a frozen writer encodes an episode into its state once, and a frozen reader from another family decodes it inside its own attention through a $\sim$3M-parameter read head. With the episode absent from the prompt, the state alone significantly lifts five readers from four families (3B–32B) on 21 of 25 reader–benchmark pairs. The state is also a *format*. Independent writers store the same episode in unrelated per-layer coordinates, so a head fit to one writer reads another as noise. Each writer sets its axes and scale alone from its mean-state SVD; one frozen head per reader then reads writers it never saw (16/16 pairs; 44/50 with a single universal head). Added to retrieval, the state improves every reader on 121K-token histories and adds $+4$ to $+21$ points on top of the gold evidence sessions.

1 INTRODUCTION

Language-model agents now persist for weeks or months (Park et al., 2023; Packer et al., 2024), while the models beneath them are replaced every few months. Every memory such an agent can hold is welded to the checkpoint that created it. A KV cache inherits one model's hidden size and tokenizer (Liu et al., 2024b). Compressed contexts live in the hidden space of the decoder that compressed them (Mu et al., 2023; Ge et al., 2024). Parametric memories are written into one backbone's weights (Meng et al., 2023; Wang et al., 2024), and output-distribution memories are tied to one vocabulary (Cao et al., 2025). Text stores survive model upgrades through their shared symbolic interface (Zhong et al., 2024). Text is re-paid in context tokens on every call, grows without bound, is used unevenly once long (Liu et al., 2024a), and carries the writer's biases (Chang et al., 2026). When the backbone changes, the agent's experience dies with it.

It does not have to. Delta-rule associative memories (Schlag et al., 2021; Yang et al., 2024; Behrouz et al., 2025) accumulate context into a small matrix per layer. δ-mem (Lei et al., 2026) brings this write to a frozen backbone as an adapter, so the state can be snapshotted after an episode. That state is an $r \times r$ matrix per attention layer, whose shape depends on the rank alone and not on the hidden size, tokenizer, or depth of the model that wrote it. The memory's *content* already sits in a model-agnostic container; only its read and write projections belong to a particular model.

Two obstacles stand between this container and a portable file. (i) The read projections are the writer's, and a new reader needs its own. (ii) The coordinates of the container are not agreed upon. The read is invariant under a per-layer rotation and rescaling of the state, so a reader can never observe which coordinates a writer chose. Each writer therefore settles into coordinates of its own: the same episode written by Qwen3-4B and Qwen3-8B yields states with cosine $\approx$0. This is a gauge freedom, akin to the permutation symmetry of independently trained networks (Ainsworth et al., 2023) and the unaligned axes of independent embedding spaces (Lample et al., 2018; Moschella

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2023). The question is then sharp: can a state written once by one frozen model be read *inside the attention computation* of a different frozen model, through a head that never saw the writer?

We answer it with a transplant protocol (Figure 1) that takes δ-mem's write as given. A frozen writer (Qwen3-4B with a rank-8 adapter) encodes a conversation or passage set into its per-layer state once and rewrites it in *canonical coordinates*, fixed by an SVD of its mean state over a handful of its own episodes, with no paired data, reader, or second writer. A parameter-free depth remap routes the state into a reader of different depth. A thin read head of $\sim$3M parameters decodes it inside the reader's attention, fit once per reader on a few hundred question–answer pairs. We evaluate mainly in a *load-bearing* regime: the source text is absent from the prompt, so the state is the reader's sole channel to the episode. Prior artifact transfer replays sequence-sized KV caches through pair-specific interfaces (Liu et al., 2025; Fu et al., 2026; Heo et al., 2026), including cross-family transfer with tokenizer alignment, or moves a parametric table fixed at training time (Li et al., 2026). Ours is a fixed-size episodic state written online, crossing families, tokenizers, scales, and depths, with one head per reader for every writer.

Three findings follow. First, the state carries the episode. Across five frozen readers from four families (3B–32B) and five benchmarks (LoCoMo, LongMemEval, HotpotQA, 2WikiMultihopQA, MuSiQue), the state alone lifts the reader significantly in 21 of 25 reader–benchmark pairs ("cells" hereafter), by up to $+17$ F1 on multi-session temporal QA. Second, the state is a format. A head fit to one writer reads another writer's raw state as noise (0 of 16 cross-writer cells); once each writer canonicalizes its own state, the same frozen head is significant in 16 of 16, and one head per reader, fit jointly across domains, reads every writer (44 of 50, against 0 of 50 raw). Any self-consistent basis serves, even one calibrated on random letters; matching norms alone transfers nothing. Third, the state complements retrieval at zero context cost. Added to BM25 or dense retrieval it improves every reader on LoCoMo and on $\sim$121K-token LongMemEval histories, and still helps on top of the gold evidence sessions, supplying facts from sessions the retriever missed. Throughout, a norm-matched random state is indistinguishable from no memory, and the reader's answers are the episode's own dates, quantities, and entities.

Our contributions:

- **A transplant protocol for in-attention memory.** We factor a delta-rule memory into a model-agnostic $r \times r$ state, a writer-side coordinate convention, and thin per-reader read heads: the first cross-family, cross-tokenizer, cross-scale, and cross-depth transfer of a fixed-size episodic state, significant in 21/25 load-bearing cells, at 30–56% of the full-context score with zero context tokens.
- **A writer-independent format.** Independent writers use different per-layer coordinates. Each writer sets its axes and scale from its own unlabeled states, and one frozen head per reader then reads all tested writers: 16/16 cross-writer cells at 102% median retention, 44/50 with a single universal head, and 110/112 under disjoint-corpus or random-letter calibration.
- **A zero-token complement to retrieval.** On top of BM25 and dense retrieval the state improves every reader on 121K-token histories and adds $+4$ to $+21$ points on top of the gold evidence sessions.

2 Related Work

Memory carriers. Each of the three carriers of language-model memory is bound to one model. Text stores—MemGPT (Packer et al., 2024), and HippoRAG (Gutiérrez et al., 2024)—keep past interactions outside the model and page retrieved entries into the prompt; LoCoMo (Maharana et al., 2024), and LongMemEval (Wu et al., 2025) benchmark them. Text crosses backbones, but it is priced in context tokens on every call, carries only what the retriever returns, and degrades a foreign agent (Chang et al., 2026); the transplanted state complements it (§4.4). Compressed context—ICAE (Ge et al., 2024), and KV-cache compression (Liu et al., 2024b)—lives in the hidden or key–value space of one decoder and grows with the text. In-model state—fast weight programmers (Schlag et al., 2021), DeltaNet and Gated DeltaNet (Yang et al., 2024; 2025), and Titans (Behrouz et al., 2025)—is created, consumed, and discarded inside one forward pass. δ-mem (Lei et al., 2026) brings the delta-rule write to frozen backbones as a small adapter, making the state snapshottable

but still read only by its writer. We take that write as given and make the state itself the deliverable: fixed size, zero context positions, and axes that belong to the memory rather than to any backbone.

Artifact transfer. Soft prompts cross backbones through learned projections (Su et al., 2022), and KV caches are replayed inside another model—Cache-to-Cache across families with per-pair fusers and tokenizer alignment (Fu et al., 2026), DroidSpeak among models fine-tuned from a common base (Liu et al., 2025), closed-form linear maps within a family (Heo et al., 2026); these transfer sequence-sized, position-indexed caches through pair-specific interfaces. Concurrently, Li et al. (2026) transfer a frozen Engram-style hashed-memory table through a target-side reader, a parametric artifact fixed at training time. Ours is episodic, written online at inference time, token-free, fixed-size, and read by one head per reader for every writer.

Shared geometry. Model stitching (Lenc & Vedaldi, 2015; Bansal et al., 2021), the Platonic Representation Hypothesis (Huh et al., 2024), relative representations (Moschella et al., 2023), Procrustes and adversarial embedding alignment (Schönemann, 1966; Lample et al., 2018), vec2vec (Jha et al., 2025), and permutation matching before merging (Ainsworth et al., 2023) show that independently trained models share geometry—which is why a rank-8 linear read interface suffices—but they align hidden spaces whose symmetry group is unknown or enormous, and so need shared anchors, paired data, or a translator per pair. A delta-rule read absorbs invertible value-axis maps and orthogonal key-axis maps. Our writers show strong agreement under per-layer rotations and scaling; an SVD convention fixes this subgroup from each writer's own moments and supports fixed-head transfer across writers (§4.3).

3 METHOD

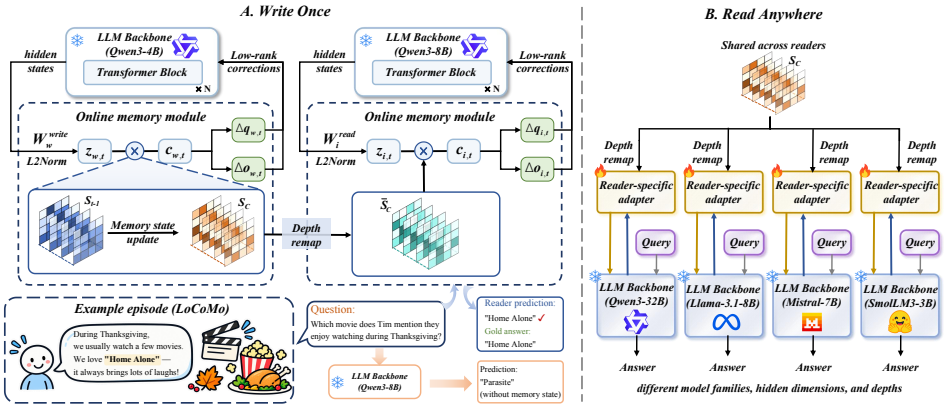

Figure 1: **Write once, read anywhere. (A)** A frozen writer with a rank-r δ-mem module writes an episode into one $r \times r$ state per layer (§3.2); the terminal states $\mathbf{S}_\mathcal{C}$, in the writer's canonical coordinates (§3.3), are the memory artifact. **(B)** A parameter-free depth remap routes the same artifact into frozen readers of other families, widths, and depths, whose thin heads (fit by cross-entropy, §3.5) form a query $\mathbf{z}$ from the reader's own hidden states, read $\mathbf{c} = \bar{\mathbf{S}}_\mathcal{C}\mathbf{z}$, and inject corrections $\Delta\mathbf{q}, \Delta\mathbf{o}$ inside attention (§3.4).

3.1 MEMORY AS A PORTABLE ARTIFACT

Let $\mathcal{W}$ be a frozen writer with L_w attention layers and hidden width d_w, and $\{\mathcal{R}_i\}$ frozen readers with L_i layers, widths d_i, and tokenizers that need not match $\mathcal{W}$'s. For an episode $\mathcal{C} = (x_1, \ldots, x_T)$—a conversation or a passage set—we seek a write map $\mathcal{C} \mapsto \mathbf{S}_\mathcal{C}$ computed by $\mathcal{W}$ alone and, per reader, a read map with parameters θ_i such that $p_{\mathcal{R}_i}(a \mid q; \mathbf{S}_\mathcal{C}, \theta_i)$ answers queries q about $\mathcal{C}$ better than $p_{\mathcal{R}_i}(a \mid q)$, where $\mathbf{S}_\mathcal{C}$ has fixed size independent of T, d_i, L_i, and any vocabulary, is written once before any reader is chosen, occupies no context positions, and is decoded by a θ_i fit once per reader rather than per writer–reader pair; in the *load-bearing* regime that anchors our evaluation, the prompt carries only q and the state is the reader's sole channel to $\mathcal{C}$.

3.2 WRITE ONCE: A DELTA-RULE STATE IN A FROZEN WRITER

We adopt δ-mem (Lei et al., 2026) as the write. Every attention layer ℓ of $\mathcal{W}$ carries a rank-r memory module with state $\mathbf{S}_t^\ell \in \mathbb{R}^{r\times r}$ and projections $\mathbf{W}_k^\ell, \mathbf{W}_v^\ell, \mathbf{W}_z^\ell, \mathbf{W}_\beta^\ell \in \mathbb{R}^{r\times d_w}$ (the figure's $\mathbf{W}_w^{\text{write}}$). With $\mathbf{h}_t^\ell \in \mathbb{R}^{d_w}$ the hidden state entering layer ℓ at token t and $\nu(\mathbf{u}) = \mathbf{u}/\|\mathbf{u}\|_2$, the module forms

$$\mathbf{k}_t = \nu(\tanh(\mathbf{W}_k\mathbf{h}_t)), \quad \mathbf{v}_t = \mathbf{W}_v\mathbf{h}_t, \quad \mathbf{z}_t = \nu(\tanh(\mathbf{W}_z\mathbf{h}_t)), \quad \boldsymbol{\beta}_t = \sigma(\mathbf{W}_\beta\mathbf{h}_t + \mathbf{b}_\beta) \in (0,1)^r, \tag{1}$$

dropping the layer index ($\tanh$ elementwise; the value path and the readout are linear), and updates the state by a row-gated delta rule,

$$\mathbf{S}_t = (\mathbf{I} - \mathbf{B}_t)\,\mathbf{S}_{t-1} + \mathbf{B}_t\big(\mathbf{v}_t - \mathbf{S}_{t-1}\mathbf{k}_t\big)\mathbf{k}_t^\top, \qquad \mathbf{B}_t = \operatorname{diag}(\boldsymbol{\beta}_t), \qquad \mathbf{S}_0 = \mathbf{0}. \tag{2}$$

This is the row-gated update used by δ-mem (Lei et al., 2026): each token writes $\mathbf{k}_t \mapsto \mathbf{v}_t$ with strength $\boldsymbol{\beta}_t$ and decays the rows it does not overwrite. Before each update it reads $\mathbf{c}_t = \mathbf{S}_{t-1}\mathbf{z}_t$ and returns it to the backbone as low-rank corrections to the query and the attention output,

$$\Delta\mathbf{q}_t = \tfrac{\alpha}{r}\,\mathbf{W}_{\Delta q}\,\mathbf{c}_t, \qquad \Delta\mathbf{o}_t = \tfrac{\alpha}{r}\,\mathbf{W}_{\Delta o}\,\mathbf{c}_t, \tag{3}$$

so that the layer attends with query $\mathbf{W}_Q\mathbf{h}_t + \Delta\mathbf{q}_t$ and emits $\mathbf{W}_O\mathbf{a}_t + \Delta\mathbf{o}_t$, where $\mathbf{a}_t$ is the attention output. The adapter—(1) and (3) at every layer—is trained once on $\mathcal{W}$ with δ-mem's objective, answer-token cross-entropy with the context written into the state and withheld from the prompt, and then frozen (Appendix B). Writing an episode is one forward pass of $\mathcal{W}$ over $\mathcal{C}$ with the scan (2) enabled; we snapshot the terminal state of every layer,

$$\mathbf{S}_\mathcal{C} = \big(\mathbf{S}_T^1, \ldots, \mathbf{S}_T^{L_w}\big) \in \mathbb{R}^{L_w\times r\times r}, \tag{4}$$

and this tensor—$L_w r^2$ numbers, independent of d_w, the vocabulary, and the length of $\mathcal{C}$—is the memory artifact (Appendix B.6).

3.3 CANONICAL COORDINATES: FIXING THE GAUGE OF THE STATE

The read interface (7) absorbs any per-layer re-gauging $\mathbf{S} \mapsto s\,\mathbf{P}^\top\mathbf{S}\mathbf{Q}$, $\mathbf{P},\mathbf{Q} \in \mathrm{O}(r)$, $s > 0$ (Appendix A.1). This rotation–scale subgroup fits the cross-writer geometry measured in Table 15. We fix its coordinates from the writer's own statistics: let $\mathbf{M}_\ell = \mathbb{E}_\mathcal{C}[\mathbf{S}_\mathcal{C}^\ell]$ be the writer's mean state at layer ℓ over any set of its own episodes, with singular value decomposition $\mathbf{M}_\ell = \mathbf{L}_\ell\boldsymbol{\Sigma}_\ell\mathbf{R}_\ell^\top$, singular values descending; the *canonical* state is

$$\tilde{\mathbf{S}}_\mathcal{C}^\ell = s_\ell\,\mathbf{L}_\ell^\top\mathbf{S}_\mathcal{C}^\ell\mathbf{R}_\ell, \qquad s_\ell = \frac{c_\ell}{\|\mathbf{M}_\ell\|_F}, \tag{5}$$

in which the writer's mean state is diagonal, positive, and ordered; the joint sign of each singular pair is fixed relative to the leading pair by the third moment of the writer's own off-diagonal coefficients (Algorithm 1). The scale c_ℓ is either a constant (CANON; c_ℓ=0.2) or a fixed per-layer reference profile published once with the format (CANONP; the mean-state norms of our first writer). Each writer computes (5) alone from unlabeled states of its own, written from any text (Appendix D.5).

3.4 READ ANYWHERE: TRANSPLANTING THE STATE INTO A FOREIGN BACKBONE

A reader $\mathcal{R}_i$ consumes the stored (canonical) state, written $\mathbf{S}_\mathcal{C}$ from here on, through two components (Figure 1B). **Depth remap.** Reader layer j receives the writer state at matched relative depth,

$$\bar{\mathbf{S}}_\mathcal{C}^j = \mathbf{S}_\mathcal{C}^{\pi_i(j)}, \qquad \pi_i(j) = \operatorname{round}\Big(j\,\tfrac{L_w-1}{L_i-1}\Big), \qquad j = 0,\ldots,L_i-1, \tag{6}$$

a parameter-free re-indexing used for every reader. **Read head.** Each reader layer j carries a rank-r read interface $\theta_i^j = \{\mathbf{U}_j \in \mathbb{R}^{r\times d_i},\ \mathbf{V}_j^q \in \mathbb{R}^{d_i^q\times r},\ \mathbf{V}_j^o \in \mathbb{R}^{d_i\times r}\}$ (d_i^q: its query width; the figure's $\mathbf{W}_i^{\text{read}}$), which mirrors the writer's read path in (1)–(3) with the state held fixed:

$$\mathbf{z}_{i,t} = \nu(\mathbf{U}_j\mathbf{h}_{i,t}^j), \qquad \mathbf{c}_{i,t} = \bar{\mathbf{S}}_\mathcal{C}^j\,\mathbf{z}_{i,t}, \qquad \Delta\mathbf{q}_{i,t} = \tfrac{\alpha}{r}\,\mathbf{V}_j^q\mathbf{c}_{i,t}, \qquad \Delta\mathbf{o}_{i,t} = \tfrac{\alpha}{r}\,\mathbf{V}_j^o\mathbf{c}_{i,t}, \tag{7}$$

injected into the reader's frozen attention exactly as in (3), with writing disabled. The head has $|\theta_i| = r\sum_j(d_i + d_i^q + d_i)$ parameters (1.8–9.4M for our readers; Appendices B and B.6) and never references the writer's projections, width, or tokenizer; with $\bar{\mathbf{S}}_\mathcal{C} = \mathbf{0}$ it returns $\Delta\mathbf{q}_{i,t} = \Delta\mathbf{o}_{i,t} = \mathbf{0}$ for every θ_i, so the reader is then exactly the frozen base model.

3.5 FITTING THE HEAD, AND THE TRANSPLANT PROTOCOL

Objective. Let $\mathcal{D}_i = \{(\mathcal{C}, q, a)\}$ be question–answer pairs over held-in episodes written once by $\mathcal{W}$ and disjoint from every evaluation episode; the reader's prompt contains only q, in $\mathcal{R}_i$'s own chat template, so the episode reaches the reader only through the state:

$$\mathcal{L}(\theta_i) = -\,\mathbb{E}_{(\mathcal{C},q,a)\sim\mathcal{D}_i} \sum_t \log p_{\mathcal{R}_i}\big(a_t \mid a_{<t},\, q;\, \bar{\mathbf{S}}_{\mathcal{C}}, \theta_i\big). \tag{8}$$

$\mathcal{R}_i$ and $\bar{\mathbf{S}}_{\mathcal{C}}$ are frozen and gradients reach only θ_i. Per-benchmark heads are fit on a few hundred to a few thousand answer tokens (Table 8); a *universal* head per reader fits (8) on all five training splits at once (Appendix F.3).

Fixed-head protocol and inference. A head fit on one writer's canonical states is frozen; any other writer of the same rank computes its own convention (5), and the frozen head reads its canonical states without refitting (Appendix D). At inference $\mathcal{R}_i$ decodes with the remapped state loaded and writing disabled, on top of any prompt—the question alone, the full episode, or retrieved passages.

Controls. A *norm-matched random state* replaces each $\bar{\mathbf{S}}^j_{\mathcal{C}}$ with $\mathbf{G}_j \,\|\bar{\mathbf{S}}^j_{\mathcal{C}}\|_F / \|\mathbf{G}_j\|_F$, $\mathbf{G}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and is read by the trained head; an *ablated head* is a θ_i fit by (8) on such states and evaluated on them; both are run in every regime of §4 (Appendix B).

4 EXPERIMENTS

We test cross-backbone reading (§4.2), fixed-head transfer across writers (§4.3), complementarity with retrieval (§4.4), and state and read-head controls (§4.5).

4.1 SETUP

Writers, readers, heads. A frozen Qwen3-4B-Instruct (Q4) with the released rank-8 TSW δ-mem adapter writes every episode once; the Qwen3-8B (Q8) and Llama-3.1-8B writers of §4.3 use QASPER-trained adapters (no writer trains on a benchmark). The state serves five frozen readers from four families (Table 1), 3B–32B, with 32, 36, or 64 layers. Only the per-reader head trains, with cross-entropy on 332–5,426 answer tokens from held-in episodes disjoint from evaluation, one head per benchmark family unless stated. Canonical coordinates (§4.3) are estimated by each writer alone from 1,000 of its own HotpotQA-train states; other sections use Q4's native coordinates.

Benchmarks, conditions, statistics. LoCoMo (n=956) and LongMemEval (LME; 363) use F1; HotpotQA (410, answer-leak filtered), 2Wiki (303), and MuSiQue (497) use EM, with softEM for significance when passages are in the prompt. D3 is the transplanted state, B3 no memory, B-rand a norm-matched Gaussian state read through the same head. Lift is D3−B-rand unless stated; retention is a frozen head's foreign-writer lift over its training-writer lift. CIs are 95% paired bootstrap (10^4 resamples); * excludes zero; a significant cell has a positive lift whose CI excludes zero (Appendix B).

4.2 ONE STATE, FIVE READERS, FIVE BENCHMARKS—WITH AND WITHOUT CONTEXT

The state alone is read by every reader family. With the episode absent from the prompt, the 4.6 KB state is the reader's only channel to it, and 21 of 25 cells improve significantly (Table 1, top), by up to +16.9 F1 on LongMemEval and +15.2 EM on 2Wiki. At zero context tokens the state recovers 30–56% of each reader's full-context LoCoMo F1 and 32–66% on LongMemEval (Figure 2). Gains are smallest for the Qwen readers, which hold the highest no-memory scores; three of the four non-significant cells are theirs, and three are on LME.

Restoring the context does not make the state redundant. Read on top of the full episode through the same head, the state adds up to +46 EM on HotpotQA and +34 EM on MuSiQue (Table 1, bottom); every reader gains 18–34 EM on MuSiQue, and the gain over a random state is significant in 13 of 15 span-QA and 9 of 10 conversational cells (Tables 1, 13).

The state stands in for a 512-token summary. Against a summary written by the writer's own backbone, the state beats the 128-token budget for every reader (+4.9 to +10.0 F1) and ties the

512-token budget; stacked on the 512-token summary through the same head it adds a further +6.2 to +13.1 for every reader (Table 2): the summary holds a selection of facts, the state the episode's structure.

4.3 One head reads every writer: the state is a format

In raw coordinates the information is present but unreadable. A head refit per writer transfers Q8 and Llama states to every reader at margins comparable to Q4's (Appendix D.3); a fixed head reads none of it—0 of 16 cross-writer cells (Tables 17, 18). Same-episode states from two writers have cosine −0.01 to +0.02; a paired per-layer Procrustes map raises it to 0.93–0.97, one map shared across layers to 0.12–0.17 (Figure 3a; Table 15): the mismatch is the per-layer gauge of §3.3.

Table 1: **One state written by Qwen3-4B is read by five frozen readers of four families.** Top: no-context lift (D3−B3). Bottom: full-context scores, no memory → state. Conversational: F1 (×100 top); span QA: EM. *: CI excludes zero (top D3−B3; bottom D3−B-rand, softEM for span QA). Absolute no-context scores: Appendix C.

	conversational		*span QA (EM)*		
Reader	LoCoMo	LongMemEval	HotpotQA	2Wiki	MuSiQue
No context in the prompt—the state is the reader's only source (lift)					
Llama-3.1-8B	+**14.9***	+**16.9***	+**6.6***	+**15.2***	+**13.1***
Mistral-7B	+**7.1***	+**11.1***	+**6.6***	+**14.5***	+**10.7***
SmolLM3-3B	+**5.9***	+2.7	+**6.8***	+**12.5***	+**9.7***
Qwen3-8B	+**4.6***	+0.3	+**3.4***	+**7.6***	+**3.2***
Qwen3-32B	+**4.9***	−1.1	+**2.9***	+3.0	+**4.6***
Full context in the prompt—the state is read on top (no memory → +state)					
Llama-3.1-8B	.545 → **.556**	.472 → **.503***	46.3 → **57.1***	35.3 → **49.8***	28.4 → **46.7***
Mistral-7B	.393 → **.497***	.351 → **.531***	26.6 → **54.9***	19.8 → **41.6***	5.0 → **38.8***
SmolLM3-3B	.273 → **.436***	.291 → **.411***	0.2 → **45.9***	2.3 → **34.0***	0.0 → **26.6***
Qwen3-8B	.498 → **.544***	.416 → **.483***	46.8 → **56.3***	24.8 → **31.7**	15.1 → **38.8***
Qwen3-32B	.361 → **.560***	.338 → **.401***	50.2 → **60.0***	33.7 → **41.9**	23.5 → **47.5***

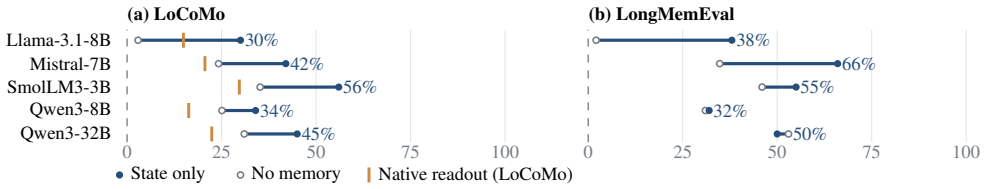

Figure 2: **State-only recovery.** No-context F1 as a percentage of each reader's full-context, no-state F1 (LoCoMo n=956; LME n=363). Filled dots: state only; open dots: no memory; orange ticks: Q4's own QASPER-fit projections reading the same state on LoCoMo (F1 ×100: 8.1).

Table 2: **The ported state beats a 128-token summary, ties a 512-token summary, and stacks on it.** LoCoMo F1 ×100, no episode text in the prompt (n=956). Native readout: Q4's own δ-mem projections reading the same state on Q4, fit by their authors on QASPER answers with no supervision from any benchmark here; summaries by the Qwen3-4B backbone. Bold: the ported state alone. Paired CIs: Appendix G.2.

	no text, no fitted head			*text summary*		*ported state*	
Reader	none	random state	native readout	128 tokens	512 tokens	alone	+512-tok. summary
Llama-3.1-8B	1.6	2.0	8.1	7.0	16.2	**16.4**	28.7
Mistral-7B	9.5	9.5	8.1	6.5	14.3	**16.5**	27.4
SmolLM3-3B	9.6	8.8	8.1	10.5	15.4	**15.4**	25.2
Qwen3-8B	12.5	11.8	8.1	7.7	16.2	**17.0**	24.0
Qwen3-32B	11.2	10.8	8.1	8.5	15.3	**16.1**	21.5

Each writer fixes the gauge alone. Canonicalization from a writer's own mean state—no pairs, no reader, no second writer—raises same-episode cosine to 0.81–0.89, and the frozen Llama head reads all three writers at full strength (Figure 3b). Across five readers on LoCoMo and three span benchmarks, CANONP gives significant cross-writer lift in 16 of 16 cells at 102% median retention and matches or exceeds a refit head in 11 of 16 (Table 3a): a frozen head reading a writer it never saw is as good as one trained on that writer, and any of the three writers can train it (9/9 cells; Table 4b). The SVD orders the mean-state axes, third moments fix their signs, and normalization sets the scale. Together they define a convention invariant under the per-layer rotation–scale subgroup (Appendix A.3).

What the convention must fix. Matching norms without rotating axes transfers nothing, a second-moment eigenbasis (CANON2) transfers as well as the mean-state basis, and a head trained under random gauges does not become gauge-invariant but stops reading even its own writer: the agreement is directional, any self-consistent basis serves, and it must live in the file (Table 4a). The norm profile is the remaining choice: Q4's mean-state norm grows 170× across layers, so CANON's constant target shrinks deep layers by over 100×, cuts the in-writer lift to +1.3 (Qwen3-8B) and +0.5 (Qwen3-32B), and leaves 14 of 16 cross-writer cells significant; CANONP's fixed reference profile

Table 3: **A head fit once on one writer's canonical states reads writers it never saw. (a)** Per-benchmark heads fit on Q4 states read Q4, Q8, or Llama states without refitting: CANONP lift over a random state (LoCoMo F1 points, span-QA EM points). **(b)** One CANONP head per reader, fit jointly on five training splits and frozen: significant positive lifts; foreign writers Q8 and Llama; last column, range of the ten foreign-writer lifts. *: CI excludes zero. CIs: Appendix D.3.

(a) Frozen CANONP head, three writers

State writer	Q4	Q8	Llama
LoCoMo, five readers			
Llama-3.1-8B	+13.8*	+14.6*	+15.0*
Mistral-7B	+3.0*	+6.3*	+7.1*
SmolLM3-3B	+8.0*	+6.5*	+4.7*
Qwen3-8B	+2.8*	+2.9*	+2.9*
Qwen3-32B	+6.0*	+5.2*	+5.6*
Span QA, Llama-3.1-8B reader			
HotpotQA	+15.0*	+15.2*	+14.6*
2Wiki	+12.9*	+13.5*	+13.2*
MuSiQue	+11.3*	+10.9*	+9.5*

(b) One universal head per reader

	Q4 writer	Foreign writers		Foreign lift
Benchmark	CANONP	Raw	CANONP	CANONP range
LoCoMo	5/5	0/10	**10/10**	+3.6–+17.1
LongMemEval	4/5	0/10	**4/10**	+0.2–+16.5
HotpotQA	5/5	0/10	**10/10**	+3.6–+15.4
2Wiki	5/5	0/10	**10/10**	+4.0–+17.5
MuSiQue	5/5	0/10	**10/10**	+1.8–+8.0
All five	24/25	0/50	**44/50**	+0.2–+17.5

Median cross-writer retention: 92% overall; 95% on the four domains other than LongMemEval, which holds all six non-significant foreign-writer cells.

Table 4: **The convention must rotate the axes; the basis, the training writer, and the calibration text do not matter.** Llama-3.1-8B reader on LoCoMo, F1-point lift over a random state (n=956); columns of (a, b): state writer. **(a)** Alternative conventions. **(b)** Each writer as training writer. **(c)** Foreign writers calibrated on text other than the training writer's; last column: no benchmark text on either side, LoCoMo heads refit; lift change is relative to shared HotpotQA calibration. *: CI excludes zero. Details: Appendix D.1.

(a) Convention, head fit on Q4

	Q4	Q8	Llama
norm matching only	+16.6*	+0.2	−0.1
gauge augmentation	+1.7*	+0.0	+0.2
CANON2 basis	+16.7*	+16.5*	+16.9*

(b) Training writer, CANON

	Q4	Q8	Llama
fit on Q4 states	+16.6*	+16.6*	+15.7*
fit on Q8 states	+16.6*	+16.1*	+16.2*
fit on Llama states	+14.9*	+16.4*	+15.5*

(c) Calibration text of the foreign writers; frozen CANONP heads

	disjoint corpora A	disjoint corpora B	random letters	benchmark-free
significant cells	64/66	10/10	26/26	10/10
median retention	91%	95%	97%	98%
median lift change	+0.0	−0.4	−0.4	−0.4

restores +2.8 and +6.0 and nearly triples Llama's MuSiQue lift (Table 3a; CANON cells in Tables 17 and 18). The format is a shared basis plus a shared norm profile.

One head per reader, for every writer and domain. One CANONP head per reader, fit jointly on the five training splits and frozen, is significant in 24 of 25 in-writer and 44 of 50 foreign-writer cells at 92% median retention (Table 3b); in raw coordinates the same heads transfer in 0 of 50. The calibration text is immaterial: writers calibrated on disjoint corpora or on random letters transfer to the frozen heads in 110 of 112 cells at 91–98% median retention, and 10/10 when no benchmark text touches any convention (Table 4c; Table 21). A writer's convention can ship with its adapter.

4.4 A ZERO-TOKEN COMPLEMENT TO RETRIEVAL

Retrieval plus state beats retrieval. With top-3 BM25 or dense BGE-M3 sessions held fixed and nothing retrained—a 16-session prompt window on LoCoMo, no raw history and 38–62-session (∼121K-token) histories on LongMemEval-full—the state improves all 20 reader–retriever combinations significantly, by +2.7 to +20.7 points on LoCoMo and +3.1 to +18.2 F1 on LongMemEval-full, at zero additional context tokens (Table 5). For Mistral, SmolLM3, and Qwen3-32B the state alone, with no retrieved text in the window, already scores above BM25 retrieval (42.2, 39.5, 43.8 against 40.4, 30.3, 37.1 graded accuracy; Table 22).

The state adds to the gold evidence itself. Added on top of the gold sessions, it still improves every reader (+3.7 to +21.1 on LoCoMo, +5.5 to +17.9 F1 on LongMemEval-full; 10/10 significant); on LoCoMo temporal questions gain more than single-hop ones (Appendices C.3, E). On single

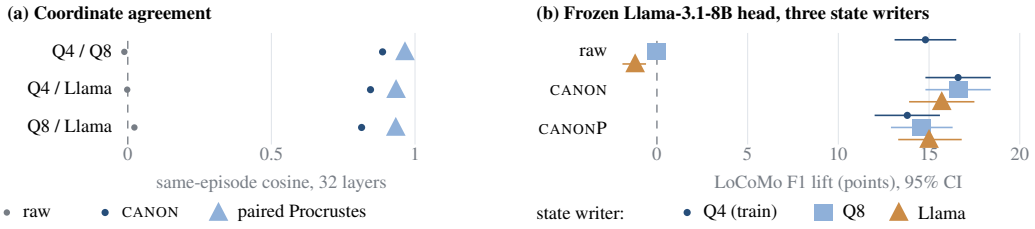

Figure 3: **Coordinate agreement and cross-writer readout.** (a) Same-episode state cosine between two writers, averaged over 32 layers: raw coordinates, CANON computed by each writer alone, and a paired four-episode Procrustes map as reference. (b) A Llama-3.1-8B head fit on Q4 states and frozen reads Q4, Q8, and Llama states under raw, CANON, and CANONP coordinates; LoCoMo F1 lift in points, n=956, bars 95% CIs (Tables 15, 17).

Table 5: **The state improves every reader on top of retrieved and even gold evidence, at zero marginal context tokens.** Gain from adding the state to fixed evidence: the same top-3 retrieved sessions, or the annotated gold sessions; 0–100 scale [95% CI], *: CI excludes zero. Absolute scores: Appendix E.

	Gain from adding state to fixed evidence		
Reader	BM25	BGE-M3	Gold sessions
(a) LoCoMo: 16-session window, n=956, graded accuracy			
Llama-3.1-8B	+2.7* [0.9,4.6]	+3.0* [1.1,4.9]	+3.8* [1.9,5.6]
Mistral-7B	+11.6* [9.4,13.8]	+11.7* [9.5,13.8]	+12.5* [10.4,14.7]
SmolLM3-3B	+18.8* [16.6,21.0]	+18.9* [16.7,21.2]	+21.1* [19.0,23.3]
Qwen3-8B	+3.7* [1.8,5.6]	+2.8* [1.0,4.7]	+3.7* [1.8,5.7]
Qwen3-32B	+20.7* [18.6,22.7]	+19.6* [17.6,21.7]	+21.1* [19.0,23.2]
(b) LongMemEval-full: no raw history, n=205, F1			
Llama-3.1-8B	+6.7* [2.2,11.3]	+5.7* [1.5,10.2]	+6.3* [1.6,11.1]
Mistral-7B	+16.7* [12.1,21.5]	+18.2* [13.5,23.0]	+17.9* [12.9,23.1]
SmolLM3-3B	+5.8* [2.9,8.8]	+6.5* [3.4,9.8]	+6.6* [3.5,9.9]
Qwen3-8B	+4.5* [1.2,7.9]	+3.1* [0.0,6.3]	+5.5* [1.9,9.2]
Qwen3-32B	+9.7* [6.2,13.7]	+11.6* [7.8,15.7]	+11.3* [7.7,15.3]

questions (Appendix E.5) the retriever returns the turn in which the user reaches Premier Silver, and the state supplies the later upgrade to Gold, from a session the retriever did not return, that makes Silver the *previous* status.

Zero tokens, negligible compute. Writing a 121K-token LongMemEval-full history is one 30-second scan of the writer, done once per episode; the artifact is 4.6 KB. Reading it costs $r(2d_i + d_i^q) + r^2$ multiply-adds per layer and token, 3.15M for the 32-layer Llama reader and 3.54M for the 36-layer Qwen3-8B, below 0.05% of either backbone's forward pass, and canonicalization is one 8×8 SVD per layer, computed once per writer (Appendix B.6).

4.5 Controls, and what the reader returns

Controls in every regime. A norm-matched random state read through the trained head is indistinguishable from no memory (Appendix G.1), a head trained on random states (*ablated head*) stays within 2.4 points of it for every reader and benchmark, and the real state lifts every reader on every span benchmark and Llama from 1.6 to 16.4 F1 on LoCoMo (Table 6). The same separation holds with passages present and inside the retrieval prompt (real-minus-random boosters significant in 10 of 10 cells, random or ablated boosters within -1.8 to $+0.8$; Tables 13, 24). What the reader returns with the real state and the question alone is the episode's own material: 16GB, 55-inch, Charles County, and 2018 composed from a session timestamp and "five years ago" (Table 14); with a retrieved session in the prompt it returns Thrive Market, from a session the retriever did not return, over the Trader Joe's the prompt names (Table 25). On LoCoMo and LongMemEval these are facts of one conversation that no backbone holds parametrically.

Table 6: **Real states lift every reader; heads trained on random states stay at the no-memory floor for every reader and benchmark.** B3: no memory; abl.: head trained and evaluated on random states; D3: real state. Conversational F1 $\times100$, span QA EM (HotpotQA: full split). Bold: D3 above B3.

	LoCoMo			LongMemEval			HotpotQA			2Wiki			MuSiQue		
Reader	B3	abl.	D3	B3	abl.	D3	B3	abl.	D3	B3	abl.	D3	B3	abl.	D3
Llama-3.1-8B	1.6	2.3	**16.4**	1.0	2.0	**18.0**	9.8	9.0	**21.2**	5.0	6.6	**20.1**	1.4	1.6	**14.5**
Mistral-7B	9.5	8.2	**16.5**	12.2	10.5	**23.3**	10.4	8.0	**21.6**	5.9	6.3	**20.5**	2.4	1.4	**13.1**
SmolLM3-3B	9.6	10.2	**15.4**	13.4	12.2	**16.1**	1.0	3.2	**16.0**	4.6	5.0	**17.2**	0.0	0.2	**9.7**
Qwen3-8B	12.5	12.8	**17.0**	12.9	11.9	**13.2**	14.6	15.0	**18.8**	9.9	9.6	**17.5**	2.4	2.4	**5.6**
Qwen3-32B	11.2	11.6	**16.1**	17.9	17.2	16.8	19.8	20.2	**25.0**	14.2	14.9	**17.2**	5.8	5.4	**10.5**

What the head needs. No teacher—cross-entropy alone gives the largest lift; distillation toward the writer's native read lowers it by more than a quarter, and distillation alone does not move the reader (Table 26)—and little data: a single training conversation already gives a significant lift, and one head pooled over the five domains matches the per-domain heads at 106% median retention (Appendix F.3). Three seeds reproduce every LoCoMo gain within 0.1–1.2 points (Table 27); Benjamini–Hochberg retains 21 of 25 headline cells and Holm 20 (Appendix F.4).

5 Conclusion

Delta-rule memories can persist across model upgrades through a shared state format and reader-specific interfaces. Their content is an $r \times r$ state per layer whose shape carries no trace of the backbone; only the read and write projections are model-specific. A $\sim$3M-parameter head, fit in minutes, reads that state inside the attention of frozen readers from four families, significantly in 21 of 25 load-bearing cells. Independent writers store the same episode in unrelated coordinates; mean-state axes and scale define a shared convention, and a frozen head reads writers it never saw as well as the writer it was fit on. On top of retrieval, and even of the gold evidence, the state adds $+4$ to $+21$ points. Memory can be written once, by any writer of the format, and read by backbones the writer never met. What remains is to fix the gauge of other memory carriers, to raise the rank, and to fit the head on benchmark-free text (Appendix H).

AI USE STATEMENT

We did not use generative AI tools for tasks requiring disclosure, including research ideation, methodological or experimental design, implementation, or interpretation of results. Generative AI tools were used solely to edit author-written manuscript text for grammar, clarity, and concision. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work involves no human subjects or new personal data. We use publicly released models and benchmark datasets under their licenses for research purposes.

REPRODUCIBILITY STATEMENT

Algorithmic details, experimental protocols, and additional results appear in the main text and appendices. §3 specifies the state update, coordinate convention, and read-head objective, with the canonicalization procedure given as Algorithm 1. Appendix B lists training data, hyperparameters, evaluation settings, controls, and statistical procedures; Appendix D specifies the coordinate ablations and calibration corpora; Appendices C, E, and F.4 provide supporting results and controls; Appendix B.6 reports storage and computational costs.

REFERENCES

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01ded4259d101feb739b06c399e9cd9c-Paper.pdf.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113506–113543. Curran Associates, Inc., 2025. doi: 10.52202/085713-3786. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4ca07aa108036f80cbb5b82285fd4b1-Paper-Conference.pdf.

Jiaqi Cao, Jiarui Wang, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. Memory decoder: A pretrained, plug-and-play memory for large language models, 2025. URL https://arxiv.org/abs/2508.09874.

Yurui Chang, Yiran Wu, Qingyun Wu, and Lu Lin. Memcollab: Cross-model memory collaboration via contrastive trajectory distillation, 2026. URL https://arxiv.org/abs/2603.23234.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://arxiv.org/abs/2402.03216v4.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LeatkxrBCi.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2591–2607, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0b276510ec2d3f6613a8b60c41ff0438-Paper-Conference.pdf.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024. doi: 10.52202/079017-1902. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ddc001d07ca4f319af96a3024f6dbd1-Paper-Conference.pdf.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-model kv cache transfer in llm families: A closed-form linear mapping for prefill reuse, 2026. URL https://arxiv.org/abs/2608.03893.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Rishi Dev Jha, Collin Zhang, Vitaly Shmatikov, and John Xavier Morris. Harnessing the universal geometry of embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jiCLUPq5xv.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Jingdi Lei, Di Zhang, Junxian Li, Weida Wang, Kaixuan Fan, Xiang Liu, Qihan Liu, Xiaoteng Ma, Baian Chen, and Soujanya Poria. δ-mem: Efficient online memory for large language models, 2026. URL https://arxiv.org/abs/2605.12357.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015. doi: 10.1109/CVPR.2015.7298701.

Mingyuan Li, Guangsheng Yu, Xu Wang, and Shaoxiong Ji. Cross-model memory transfer via target-side reader adaptation, 2026. URL https://arxiv.org/abs/2608.17050.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a.

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, and Esha Choukse. Droidspeak: Kv cache sharing for cross-llm communication and multi-llm serving, 2025. URL https://arxiv.org/abs/2411.02820.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/liu24bz.html.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19327–19352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0848. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d77c6dcc7f143aa2154e7f4d5e22d68-Paper-Conference.pdf.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. Association for Computing Machinery, 2023.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, March 1966. ISSN 1860-0980. doi: 10.1007/BF02289451.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. On transferability of prompt tuning for natural language processing. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3949–3969, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.290. URL https://aclanthology.org/2022.naacl-main.290/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian Mcauley. MEMORYLLM: Towards self-updatable large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50453–50466. PMLR, 21–27 Jul 2024.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 115491–115522. Curran Associates, Inc., 2024. doi: 10.52202/079017-3668. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d13a3eae72366e61dfdc7eea82eeb685-Paper-Conference.pdf.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

APPENDIX OVERVIEW

The appendix follows the order of the main text. Every number quoted in §4 has a row here, and every table sits next to the text that reads it.

Appendix	Contents	Supports
A Method details	Gauge of the read path, exact symmetry of the write, canonical map and its sign-fixing algorithm, depth remap, head sizes	§3
B Experimental setup	Writers, readers, head training, benchmarks, regimes, retrieval, controls, statistics, cost	§4.1
C Additional main results	Absolute scores, with-context controls, question categories, worked examples	§4.2
D Cross-writer transfer	Coordinate conventions, state geometry, every fixed-head cell, universal heads, calibration corpora	§4.3
E Retrieval complementarity	Absolute retrieval scores, question types, k and window sweeps, booster and gold-session controls, worked examples	§4.4
F Ablations and head generalization	Objective, data budget, norm profile, pooled heads, seeds	§4.5
G Content and carrier controls	Random-state and untrained-head controls, summary-carrier contrasts	§4.5
H Scope and open problems	What the results cover and what they leave open	—

A METHOD DETAILS

A.1 WHAT THE READ HEAD ABSORBS

Fix one reader layer and drop its index. The read (7) is a composition of three linear maps around one normalization,

$$\mathbf{z} = \nu(\mathbf{U}\mathbf{h}), \qquad \mathbf{c} = \bar{\mathbf{S}}\,\mathbf{z}, \qquad \Delta\mathbf{q} = \tfrac{\alpha}{r}\mathbf{V}^{q}\mathbf{c}, \qquad \Delta\mathbf{o} = \tfrac{\alpha}{r}\mathbf{V}^{o}\mathbf{c}. \tag{A.1}$$

Let the stored state be re-expressed in another gauge,

$$\bar{\mathbf{S}}' = \mathbf{A}\bar{\mathbf{S}}\,\mathbf{Q}, \qquad \mathbf{A} \in \mathrm{GL}(r), \quad \mathbf{Q} \in \mathrm{O}(r), \tag{A.2}$$

and transform the head alongside it,

$$\mathbf{U}' = \mathbf{Q}^{\top}\mathbf{U}, \qquad \mathbf{V}^{q\prime} = \mathbf{V}^{q}\mathbf{A}^{-1}, \qquad \mathbf{V}^{o\prime} = \mathbf{V}^{o}\mathbf{A}^{-1}. \tag{A.3}$$

Because an orthogonal map preserves the norm, $\nu(\mathbf{Q}^{\top}\mathbf{u}) = \mathbf{Q}^{\top}\nu(\mathbf{u})$, and the chain (A.1) closes:

$$\mathbf{z}' = \mathbf{Q}^{\top}\mathbf{z}, \qquad \mathbf{c}' = \mathbf{A}\bar{\mathbf{S}}\,\mathbf{Q}\mathbf{Q}^{\top}\mathbf{z} = \mathbf{A}\mathbf{c}, \qquad \mathbf{V}^{q\prime}\mathbf{c}' = \mathbf{V}^{q}\mathbf{A}^{-1}\mathbf{A}\mathbf{c} = \mathbf{V}^{q}\mathbf{c}. \tag{A.4}$$

The corrections injected into attention are identical, hence so is every reader output and the objective (8):

$$\mathcal{L}(\theta;\,\bar{\mathbf{S}}) = \mathcal{L}(\theta';\,\bar{\mathbf{S}}') \quad \text{for every } (\mathbf{A},\mathbf{Q}) \in G_{\mathrm{read}} := \mathrm{GL}(r) \times \mathrm{O}(r). \tag{A.5}$$

The linear value path absorbs every invertible $\mathbf{A}$; query normalization permits orthogonal $\mathbf{Q}$. The convention (5) fixes the subgroup $\mathbf{A} = s\mathbf{P}^{\top}$ with $s > 0$ and $\mathbf{P} \in \mathrm{O}(r)$, using ordered mean-state axes, paired signs, and a norm profile (§A.3). A head can be fit in each writer's native coordinates, as the refit results demonstrate (Table 17). Fixed-head transfer instead requires shared coordinates: raw states yield 0/50 significant cross-writer cells, and the rotation–scale convention yields 44/50 (Table 19). Native Q4 coordinates serve for every result outside §4.3.

A.2 WHAT THE WRITE FIXES

The recurrence (2) has a linear skeleton—projection, normalization, delta rule, readout—that is equivariant under $\mathrm{O}(r)$ on the key axis; the exact symmetry of the full module is smaller. Rotate the key-side projections,

$$\mathbf{W}_k \mapsto \mathbf{Q}^{\top}\mathbf{W}_k, \quad \mathbf{W}_z \mapsto \mathbf{Q}^{\top}\mathbf{W}_z \quad \Longrightarrow \quad \begin{aligned} \mathbf{k}_t &\mapsto \nu\big(\tanh(\mathbf{Q}^{\top}\mathbf{W}_k\mathbf{h}_t)\big), \\ \mathbf{z}_t &\mapsto \nu\big(\tanh(\mathbf{Q}^{\top}\mathbf{W}_z\mathbf{h}_t)\big). \end{aligned} \tag{A.6}$$

Whenever these equal $\mathbf{Q}^\top \mathbf{k}_t$ and $\mathbf{Q}^\top \mathbf{z}_t$, the state rotates with the keys: if $\mathbf{S}_{t-1} \mapsto \mathbf{S}_{t-1}\mathbf{Q}$, then by (2)

$$(\mathbf{I} - \mathbf{B}_t)\,\mathbf{S}_{t-1}\mathbf{Q} + \mathbf{B}_t\big(\mathbf{v}_t - \mathbf{S}_{t-1}\mathbf{Q}\mathbf{Q}^\top \mathbf{k}_t\big)\mathbf{k}_t^\top \mathbf{Q} = \mathbf{S}_t\mathbf{Q}, \qquad \mathbf{S}_{t-1}\mathbf{Q}\mathbf{Q}^\top \mathbf{z}_t = \mathbf{S}_{t-1}\mathbf{z}_t, \tag{A.7}$$

so the whole state trajectory rotates and the writer's own read is unchanged. Without the $\tanh$ this holds for every $\mathbf{Q} \in \mathrm{O}(r)$, since $\nu(\mathbf{Q}^\top \mathbf{u}) = \mathbf{Q}^\top \nu(\mathbf{u})$. With it, $\tanh(\mathbf{Q}^\top \mathbf{u}) = \mathbf{Q}^\top \tanh(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^r$ exactly when $\mathbf{Q}$ is a signed permutation, because $\tanh$ is elementwise and odd: the exact symmetry of the key axis is $\mathcal{B}_r$, the signed permutations. For a positive diagonal $\mathbf{D}$, independent value-axis scaling gives

$$\mathbf{W}_v \mapsto \mathbf{D}\mathbf{W}_v \quad \Longrightarrow \quad \mathbf{S}_t \mapsto \mathbf{D}\mathbf{S}_t, \qquad \mathbf{W}_{\Delta q}, \mathbf{W}_{\Delta o} \mapsto \mathbf{W}_{\Delta q}\mathbf{D}^{-1},\ \mathbf{W}_{\Delta o}\mathbf{D}^{-1}. \tag{A.8}$$

Here $\mathbf{D}$ commutes with every diagonal gate $\mathbf{B}_t$. More generally, $\mathbf{W}_v \mapsto \mathbf{A}\mathbf{W}_v$ and $\mathbf{S}_t \mapsto \mathbf{A}\mathbf{S}_t$ require the transformed gate $\mathbf{A}\mathbf{B}_t\mathbf{A}^{-1}$:

$$\mathbf{A}(\mathbf{I} - \mathbf{B}_t)\,\mathbf{S}_{t-1} = (\mathbf{I} - \mathbf{A}\mathbf{B}_t\mathbf{A}^{-1})\,\mathbf{A}\mathbf{S}_{t-1}. \tag{A.9}$$

This gate is diagonal for every $\boldsymbol{\beta}_t \in (0,1)^r$ exactly when $\mathbf{A}$ is monomial: $\mathbf{A} = \mathbf{D}\mathbf{P}$, with $\mathbf{P} \in \mathcal{B}_r$. Writing $\mathcal{M}_r$ for this group, the write symmetries are

$$G_{\text{write}} = \mathcal{M}_r \times \mathcal{B}_r \ \subset\ G_{\text{read}}, \tag{A.10}$$

with monomial value maps, signed key permutations, and gate rows permuted alongside the value axes. These transformations preserve the state trajectory and writer readout after the projection changes above. Across independently trained writers, paired per-layer Procrustes fits reach cosine 0.93–0.97; their rotations place 0.41–0.44 of their squared Frobenius mass on the best signed permutation (Table 15). This measured agreement motivates the rotation–scale convention of §A.3. Its axes and norm profile are both operational: norm-only transfer gives $+0.2$ and -0.1, gauge augmentation gives $+0.0$ and $+0.2$, and independent canonicalization gives $+16.6$ and $+15.7$ for the same foreign writers (Table 16). The convention supplies shared coordinates for the fixed head; the cross-writer tables measure the resulting transfer.

A.3 Invariance of the canonical map

For the subgroup $\mathbf{A} = s\mathbf{P}^\top, \mathbf{P}, \mathbf{Q} \in \mathrm{O}(r)$, $s > 0$, the mean state and its SVD transform as

$$\mathbf{M}_\ell \mapsto s\,\mathbf{P}^\top \mathbf{M}_\ell \mathbf{Q} = (\mathbf{P}^\top \mathbf{L}_\ell)\,(s\boldsymbol{\Sigma}_\ell)\,(\mathbf{Q}^\top \mathbf{R}_\ell)^\top, \qquad s_\ell = \frac{c_\ell}{\|\mathbf{M}_\ell\|_F} \ \mapsto\ \frac{s_\ell}{s}, \tag{A.11}$$

so that substituting into (5) returns the same canonical state throughout this subgroup:

$$\frac{s_\ell}{s}\,(\mathbf{L}_\ell^\top \mathbf{P})\big(s\,\mathbf{P}^\top \mathbf{S}_{\mathcal{C}}^\ell \mathbf{Q}\big)(\mathbf{Q}^\top \mathbf{R}_\ell) = s_\ell\,\mathbf{L}_\ell^\top \mathbf{S}_{\mathcal{C}}^\ell \mathbf{R}_\ell = \tilde{\mathbf{S}}_{\mathcal{C}}^\ell. \tag{A.12}$$

The one residual freedom is the joint sign of each singular pair, which leaves the diagonal mean fixed but flips off-diagonal entries of the canonical state,

$$(\mathbf{l}_k, \mathbf{r}_k) \mapsto (-\mathbf{l}_k, -\mathbf{r}_k) \quad \Longleftrightarrow \quad \tilde{\mathbf{S}} \mapsto \mathbf{D}\,\tilde{\mathbf{S}}\,\mathbf{D}, \qquad \mathbf{D} = \mathrm{diag}(\pm 1), \qquad \mathbf{D}\,\boldsymbol{\Sigma}\,\mathbf{D} = \boldsymbol{\Sigma}; \tag{A.13}$$

Write $a_{kj}(\mathcal{C}) = \mathbf{l}_k^\top \mathbf{S}_{\mathcal{C}}^\ell \mathbf{r}_j$ for the coefficients of a state before the signs are fixed, so that $\tilde{\mathbf{S}}_{\mathcal{C}} = s_\ell\,\mathbf{D}\,[a_{kj}(\mathcal{C})]\,\mathbf{D}$. A joint flip of pair k negates a_{kj} and a_{jk} for every $j \neq k$ and leaves a_{kk} unchanged; $\mathbf{D}$ and $-\mathbf{D}$ act identically, so the residual group has 2^{r-1} elements and the $r-1$ signs *relative* to one anchor pair are what a convention has to fix. Coefficient-wise third moments supply sign-sensitive statistics: the first moment of every off-diagonal coefficient vanishes identically, $\mathbb{E}_{\mathcal{C}}[a_{kj}] = (\mathbf{L}_\ell^\top \mathbf{M}_\ell \mathbf{R}_\ell)_{kj} = 0$ for $k \neq j$, and each coefficient-wise second moment is even under a flip. Algorithm 1 anchors on the leading pair and fixes each further pair by the sign of its accumulated third moment with the pairs already fixed. With nonzero anchor sums, each step fixes a relative sign. Under the rotation–scale subgroup, every a_{kj} transforms by the scale s, removed by s_ℓ, and by the joint signs resolved by the algorithm.

Coincident singular values. $\mathbf{L}_\ell$ and $\mathbf{R}_\ell$ are determined up to an orthogonal block on any subspace where singular values of $\mathbf{M}_\ell$ coincide, so the convention is defined on layers with a simple spectrum. On a near-degenerate block the second-moment eigenbasis of CANON2, restricted to the block, breaks the tie; CANON2 transfers as well as CANON (Table 4a).

Algorithm 1 Canonical coordinates of one writer layer (CANON; CANONP changes only c_ℓ).

Require: calibration states $\{\mathbf{S}_\mathcal{C}\}_{\mathcal{C}\in\mathcal{K}}$ written by the writer at this layer; target scale c_ℓ

$\mathbf{M} \leftarrow |\mathcal{K}|^{-1}\sum_\mathcal{C} \mathbf{S}_\mathcal{C}$; $\mathbf{M} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^\top$ with $\sigma_1 > \cdots > \sigma_r > 0$

$a_{kj}(\mathcal{C}) \leftarrow \mathbf{l}_k^\top \mathbf{S}_\mathcal{C} \mathbf{r}_j$ for all $k, j, \mathcal{C}$; $m_{kj} \leftarrow |\mathcal{K}|^{-1}\sum_\mathcal{C}\big(a_{kj}(\mathcal{C})^3 + a_{jk}(\mathcal{C})^3\big)$

$d_1 \leftarrow +1$

for $k = 2, \ldots, r$ **do**

  $d_k \leftarrow \operatorname{sign}\big(\sum_{j<k} d_j\, m_{jk}\big)$

end for

$\mathbf{D} \leftarrow \operatorname{diag}(d_1, \ldots, d_r)$; $\mathbf{L} \leftarrow \mathbf{L}\mathbf{D}$; $\mathbf{R} \leftarrow \mathbf{R}\mathbf{D}$

return $\tilde{\mathbf{S}}_\mathcal{C} = (c_\ell/\|\mathbf{M}\|_F)\,\mathbf{L}^\top \mathbf{S}_\mathcal{C}\mathbf{R}$ for every state of the writer

Sign regularity. Algorithm 1 defines relative signs when every anchor sum $\sum_{j<k} d_j m_{jk}$ is nonzero. Under a singular-pair sign change, each sum acquires exactly the sign of that pair relative to the anchor, preserving the canonical state. The absolute anchor sums measure sign stability; sampling perturbations smaller than these margins preserve the selected signs.

Canonical coordinates and cross-writer agreement. Equations (A.11)–(A.13) and Algorithm 1 give rotation–scale invariance for a positive simple mean spectrum and nonzero sign anchors. The convention is computed from the writer's own state distribution. Calibration uses 1,000 HotpotQA-train episodes by default; a one-episode basis estimate already gives cross-writer cosine 0.80–0.84 (§D.2). Code, mathematics, fiction, medicine, and random-letter calibration yield the transfer results of §D.5. Cross-writer agreement is measured on held-out episodes: the same episode written by Q4, Q8, and Llama reaches cosine 0.81–0.89 after independent canonicalization, within 0.08–0.12 of a paired Procrustes fit (Table 15), and a frozen head reads all three at 99–102% median retention (Tables 17, 18). CANON and CANON2 publish nothing beyond the writer's adapter; CANONP publishes one per-layer norm profile c_ℓ with the format.

A.4 Depth remap and head sizes

For the 36-layer writer the remap (6) is explicit at the three reader depths:

$$\pi_{36}(j) = j, \qquad \pi_{32}(j) = j + \left\lfloor \tfrac{j+4}{8} \right\rfloor, \qquad \pi_{64}(j) = \left\lfloor \tfrac{35j+31}{63} \right\rfloor. \tag{A.14}$$

The 36-layer readers (SmolLM3-3B, Qwen3-8B) read every writer layer at its own depth; the 32-layer readers (Llama-3.1-8B, Mistral-7B) skip writer layers 4, 13, 22, 31 and read the other 32 in order; the 64-layer Qwen3-32B reads writer layers $0, 5, \ldots, 35$ once and each of the remaining 28 through two consecutive reader layers. Conventions (5) are properties of the writer's layers and are unaffected by the remap, which only re-indexes them. The head that reads the remapped state has, per layer, the three matrices of (A.1), so its size and per-token cost are

$$|\theta_i| = r\,L_i\,(2d_i + d_i^q), \qquad \text{MACs per layer and token} = r\,(2d_i + d_i^q) + r^2, \tag{A.15}$$

which Table 7 evaluates for every backbone in the study; the artifact they all read is

$$|\mathbf{S}_\mathcal{C}| = L_w r^2 = 36 \times 64 = 2{,}304 \text{ values} = 4{,}608 \text{ bytes in bfloat16}, \tag{A.16}$$

whatever the length of $\mathcal{C}$.

The largest head, 9.44M parameters for Qwen3-32B, is 0.03% of its backbone and below 0.05% of its per-token multiply-adds; the 1.77M head of SmolLM3-3B reads the same 4.6 KB file. With $\bar{\mathbf{S}}_\mathcal{C} = \mathbf{0}$ every head returns $\Delta\mathbf{q} = \Delta\mathbf{o} = \mathbf{0}$ by (A.1), so a reader with no state loaded is bit-for-bit its frozen base model: this is the no-memory condition B3 used throughout.

Table 7: **Backbones, depths, and interface sizes.** d: hidden width; d^q: query width (heads $\times$ head dimension). Head sizes follow (A.15) with $r = 8$; the writer's adapter adds $\mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_z, \mathbf{W}_\beta$ and the bias $\mathbf{b}_\beta$ at each layer.

Model	Role	Layers	d	d^q	Interface	Remap (A.14)
Qwen3-4B-Instruct-2507	writer (Q4)	36	2,560	4,096	4.87M adapter	—
Qwen3-8B	writer (Q8)	36	4,096	4,096	rank-8 adapter	—
Llama-3.1-8B-Instruct	writer (Ll.)	32	4,096	4,096	rank-8 adapter	—
SmolLM3-3B	reader	36	2,048	2,048	1.77M head	π_{36}: identity
Llama-3.1-8B	reader	32	4,096	4,096	3.15M head	π_{32}: skips 4 layers
Mistral-7B	reader	32	4,096	4,096	3.15M head	π_{32}: skips 4 layers
Qwen3-8B	reader	36	4,096	4,096	3.54M head	π_{36}: identity
Qwen3-32B	reader	64	5,120	8,192	9.44M head	π_{64}: 28 layers twice

B Experimental setup

B.1 Writers and adapters

The principal writer Q4 is Qwen3-4B-Instruct-2507 with the publicly released δ-mem adapter (`declare-lab/delta-mem_qwen3_4b-instruct`): rank $r = 8$, token-state write (TSW, the token-level write granularity of δ-mem), corrections on the query and output projections, write length 8,192, 4,866,336 parameters (0.121% of the backbone), read scale $\alpha/r = 16/8$. The adapter is used as released, trained by its authors on QASPER question–answer pairs with the objective below, without modification: keys and read queries pass through $\tanh$ and then L2 normalization (`normalize_qk`), the forget gate is coupled to the write gate, $\lambda_t = 1 - \beta_t$, and the update is the row-gated rule (2). It writes every episode in the paper once. The Qwen3-8B and Llama-3.1-8B-Instruct writers of §4.3 carry adapters trained with the same δ-mem code and module configuration, and one recipe: 3,598 read-then-answer episodes from QASPER (Dasigi et al., 2021) training and validation papers, at most 8,192 written tokens per episode, one epoch, global batch size 32, AdamW at 2×10^{-4} with cosine decay, bfloat16, ZeRO-2 on four NVIDIA H800 GPUs, and δ-mem's objective (Lei et al., 2026): cross-entropy on the answer tokens, with the paper written into the state and withheld from the prompt. No benchmark text enters any writer's training.

B.2 Readers and read heads

The five readers and their head sizes are listed in Table 7. All readers are the post-trained chat checkpoints of their families, since the prompt is rendered in each reader's own chat template: Llama-3.1-8B-Instruct, Mistral-7B-Instruct, SmolLM3-3B, Qwen3-8B, and Qwen3-32B. The Qwen3 and SmolLM3 readers decode in their non-thinking mode: the 50-token answer budget of the conversational benchmarks leaves no room for a reasoning block. Backbones and written states stay frozen; only the head trains, on answer-token cross-entropy (8) with AdamW at 2×10^{-4}, weight decay 10^{-2}, batch size 1, gradient accumulation 8, and seed 0. The prompt during training and evaluation holds the question in the reader's own chat template and nothing else, so the loss can be lowered only through the state. One head per benchmark family is the default; the universal heads of §F change only the training episodes. Table 8 gives every head's supervision budget.

The budgets are small. The LongMemEval-oracle head sees 332–422 answer tokens—fifty short answers—and lifts Llama by $+16.9$ F1 (Table 1); the span-QA heads see under 5.5K tokens and run for 150 steps. Head fitting takes 148–358 s for span QA, 285–689 s for the conversational heads, and 401–922 s for universal heads on one H800 (§B.6).

B.3 Benchmarks, regimes, and metrics

Table 9 lists the six evaluation sets, including HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). LoCoMo scores 956 questions over six held-out conversation states; other benchmarks write one state per instance.

Table 8: **Head-training data and supervision.** Training episodes are disjoint from every evaluation episode. Answer-token counts are the supervised tokens, ranging across reader tokenizers. A head fit once is reused for every episode of its benchmark.

Head	Training episodes	QA pairs	Steps	Answer tokens
LoCoMo	Four held-in conversations	584	300	3.9–4.6K
LongMemEval-oracle	50 held-in conversations	50	300	332–422
LongMemEval-full	80 held-in histories	80	300	0.9–1.1K
HotpotQA	Training split	1,000	150	4.7–5.4K
2WikiMultihop	Training split	665	150	2.3–2.6K
MuSiQue	Training split	1,000	150	4.4–5.4K
Universal ALL-5	Balanced mixture of the five training splits	3,299	400	16.6–19.5K
Ablated, LoCoMo	The LoCoMo data above, with random states	584	300	3.9–4.6K

Table 9: **Benchmarks.** n: scored questions; episodes: written states evaluated. HotpotQA has 500 questions before the leak filter below. †F1 in the no-context and full-context regimes, graded accuracy in the retrieval regime; ‡EM, with softEM for significance when passages are in the prompt.

Benchmark	Episode	Episodes	n	Question types	Metric
LoCoMo	conversation, 19–26K tokens	6	956	single-hop, multi-hop, temporal, open-domain	F1†
LongMemEval-oracle	evidence sessions, $\leq$6	363	363	single-session, multi-session, temporal	F1
LongMemEval-full	full history, 38–62 sessions, $\sim$121K tokens	390	205	multi-session (100), temporal (105)	token F1
HotpotQA	passage set	410	410	bridge, comparison	EM‡
2Wiki	passage set	303	303	compositional, comparison, inference, bridge-comparison	EM‡
MuSiQue	passage set	497	497	2-, 3-, 4-hop	EM‡

Leak filter. A span-QA question is removed when the normalized gold answer appears as a contiguous token sequence in the question. Ninety of 500 HotpotQA questions are removed (n=410); none from 2Wiki or MuSiQue.

Regimes. In the *no-context* regime the prompt contains only the question. In the *full-context* regime it contains the entire episode. The LoCoMo *retrieval* regime keeps the last 16 sessions in the prompt and retrieves further sessions from the history; LongMemEval-full keeps no raw history and retrieves over all 38–62 sessions.

Scoring. LoCoMo no-context and full-context runs use official F1; LoCoMo retrieval uses graded accuracy. LongMemEval-oracle uses token-level F1; LongMemEval-full uses normalized token-level F1. Span QA without context uses EM and F1. With passages present, significance uses $\mathrm{softEM}(p, g) = \mathbb{1}[g \text{ is a contiguous token span of } p]$ after normalization, which is insensitive to answer verbosity; raw EM is reported alongside and marked. Decoding is greedy, with the official conversational prompt and a 50-token answer limit for the conversational benchmarks; span-QA prompts are identical across conditions.

Prompt construction. Every prompt is rendered in the reader's own chat template and has the same three-part structure: the benchmark's official task instruction, an evidence block, and the question. The evidence block is what the regimes vary—empty in the no-context regime, the entire episode in the full-context regime, the retrieved or gold sessions in chronological order in the retrieval regime, and the passage set for span QA with passages. The instruction and question strings are the benchmarks' own (LoCoMo's short-answer instruction with its date convention for temporal questions; LongMemEval's history-based answering instruction; a short-answer span-QA instruction), so a reader answers the no-context prompt by treating the missing block as an empty history, as the "the conversation does not mention" prediction in Table 14 shows. The summaries of §G.2 are produced

by the Qwen3-4B backbone without its adapter from a single summarization instruction over the full conversation. The rendered templates ship with the code release.

B.4 Retrieval

BM25 with standard Okapi weighting ranks session-level documents. Dense retrieval embeds sessions and questions with BGE-M3 (Chen et al., 2024) (`BAAI/bge-m3`) and ranks by cosine similarity. Retrieved sessions enter the prompt in chronological order, k=3 unless stated, and the retrieved text is byte-identical with and without the state. The gold-session condition supplies the benchmark's annotated evidence sessions. Qwen3-32B retrieval runs at batch size 1 with prompt limits of 16,000 tokens on LongMemEval-full and 40,910 on the LoCoMo window.

B.5 Controls and statistics

Conditions. B3 is the frozen backbone with no state; D3 is the written state read through the trained head; B-rand replaces each layer's state with an independent Gaussian tensor scaled to the written tensor's Frobenius norm, read through the same trained head; the *ablated head* is a head trained on such random states and evaluated on them; B2 is an untrained head on the real state. A zero state returns the frozen backbone exactly (7), so B3 needs no separate model.

Intervals. Every interval is a 95% paired percentile bootstrap over 10^4 resamples, by question on LoCoMo and by instance on span QA and LongMemEval. LoCoMo's five headline gains stay significant when whole conversations are resampled. A star marks an interval that excludes zero; counts of transfer cells refer to significant *positive* lifts. Table 1 reports D3−B3 in its no-context panel; content controls and every cross-writer table report D3−B-rand, the stricter contrast. Retention is a frozen head's foreign-writer lift divided by its training-writer lift under the same metric and control.

Multiple comparisons. The 25 no-context reader–benchmark cells form one family. Benjamini–Hochberg at q=0.05 retains 21 significant cells and Holm at α=0.05 retains 20. The ten LongMemEval-full retrieval boosters survive both procedures.

B.6 Storage and computation

A rank-8 state from the 36-layer writer is 2,304 bfloat16 values, 4,608 bytes, whatever the episode length. Writing is one scan of the frozen writer, shared by all readers: 3.4–5.1 s for a 19–26K-token LoCoMo conversation and 29.9–30.5 s for a 107–109K-token LongMemEval-full history on one H800. Head fitting takes 148–358 s for span QA, 285–689 s for the conversational heads, and 401–922 s for the universal heads; a fitted head is reused for every episode. Table 10 sets the three carriers side by side.

Table 10: **Context and amortized computation.** Times are measured on one H800; read-operation counts are analytical.

Carrier	Context positions per query	Preprocessing or adaptation
Full episode	∼25K tokens (LoCoMo); ∼121K (LongMemEval-full)	The text is re-read at every query
BM25 top-3 sessions	∼2.5K tokens (LoCoMo)	Index construction
Memory state	0	One writer scan per episode; one head fit per reader, reused for every episode

At inference the read path costs $r(2d_i + d_i^q) + r^2$ multiply-adds per layer and token: 3.15M per token for the 32-layer Llama reader and 3.54M for the 36-layer Qwen3-8B reader, below 0.05% of either backbone's forward multiply-add count. Canonicalization is one SVD of an 8×8 matrix per layer, computed once per writer, and changes neither the state size nor the read cost.

C ADDITIONAL MAIN RESULTS

C.1 ABSOLUTE SCORES BEHIND TABLE 1

Table 11 gives the conversational scores that produce the no-context lifts and the recovery percentages of Figure 2; Table 12 gives the span-QA scores.

Table 11: **Conversational benchmarks, F1.** No context: no state (B3) and transplanted state (D3); full context: no state. Recovery is D3 divided by the full-context score, as in Figure 2. LoCoMo n=956; LongMemEval-oracle n=363.

	LoCoMo				LongMemEval-oracle			
Reader	B3	D3	full ctx	recovery	B3	D3	full ctx	recovery
Llama-3.1-8B	.016	**.164**	.545	30%	.010	**.180**	.472	38%
Mistral-7B	.095	**.165**	.393	42%	.122	**.233**	.351	66%
SmolLM3-3B	.096	**.154**	.273	56%	.134	**.161**	.291	55%
Qwen3-8B	.125	**.170**	.498	34%	.129	**.132**	.416	32%
Qwen3-32B	.112	**.161**	.361	45%	.179	.168	.338	50%

Two readers make the scale of the effect concrete. Llama-3.1-8B answers essentially nothing from the question alone (.016 F1 on LoCoMo, .010 on LongMemEval); with the 4.6 KB state and no episode text it reaches .164 and .180, the latter through a head that saw fifty short answers. Mistral-7B with the state alone recovers two thirds of what it scores when the entire LongMemEval episode is in its prompt (.233 against .351).

Table 12: **Span QA without context, EM / F1 (%).** HotpotQA n=410; 2WikiMultihop n=303; MuSiQue n=497 (leak filter in §B.3).

		HotpotQA		2Wiki		MuSiQue	
Reader	memory	EM	F1	EM	F1	EM	F1
Llama-3.1-8B	none	7.1	11.9	5.0	6.6	1.4	3.1
	state	**13.7**	**26.6**	**20.1**	**25.7**	**14.5**	**22.3**
Mistral-7B	none	5.1	15.9	5.9	11.7	2.4	7.3
	state	**11.7**	**26.1**	**20.5**	**25.5**	**13.1**	**22.3**
SmolLM3-3B	none	0.5	1.6	4.6	4.6	0.0	0.3
	state	**7.3**	**18.0**	**17.2**	**24.1**	**9.7**	**16.8**
Qwen3-8B	none	6.6	15.2	9.9	13.7	2.4	6.8
	state	**10.0**	**20.9**	**17.5**	**22.0**	**5.6**	**12.6**
Qwen3-32B	none	11.0	24.1	14.2	22.8	5.8	14.7
	state	**13.9**	**26.3**	**17.2**	**23.2**	**10.5**	**20.2**

On span QA the pattern is the same across families: 2Wiki EM more than triples for Llama, Mistral and SmolLM3 (5.0→20.1, 5.9→20.5, 4.6→17.2), and on MuSiQue—composed answers that no reader holds parametrically—SmolLM3 rises from 0.0 to 9.7 EM and Llama from 1.4 to 14.5. All thirty state cells exceed their no-memory counterparts.

C.2 CONTROLS WITH THE PASSAGES PRESENT

Table 13 repeats the content controls in the full-context regime, where the reader has the passages and could in principle ignore the state.

The real state beats the norm-matched random state in 13 of 15 cells, by 4.8–29.6 softEM points; on MuSiQue the margin is +16.1 to +29.6 for every reader. Neither the random state nor the ablated head moves any reader off its no-state baseline (B-rand−B3 within −1.7 to +2.6; abl.−B3 never significantly positive), so the with-context gain is read from the state's content, not from the presence of an extra input. Raw-EM gains over B3 are significant in 15 of 15 cells. The two cells whose softEM interval touches zero are the Qwen readers on 2Wiki, where the state still adds +4.0 and +2.6 points.

Table 13: **With-context controls (softEM, %; passages in the prompt).** B3: no state; B-rand: norm-matched random state through the trained head; abl.: head trained and evaluated on random states; D3: transplanted state. Contrasts are paired, 10^4 bootstrap resamples, 95% CI; *: CI excludes zero. HotpotQA uses the full 500-question split; 2Wiki n=303; MuSiQue n=497.

		softEM				contrasts [95% CI]		
Benchmark	Reader	B3	B-rand	abl.	D3	D3 − B-rand	B-rand − B3	abl. − B3
HotpotQA	Llama-3.1-8B	55.0	53.4	53.0	**70.0**	+16.6 [+12.6, +20.6]*	−1.6 [−4.2, +0.8]	−2.0 [−4.4, +0.2]
	Mistral-7B	45.6	48.2	47.4	**62.8**	+14.6 [+9.8, +19.4]*	+2.6 [−0.4, +5.6]	+1.8 [−1.4, +5.2]
	SmolLM3-3B	40.4	39.8	34.6	**58.0**	+18.2 [+13.2, +23.2]*	−0.6 [−3.0, +1.8]	−5.8 [−9.0, −2.6]*
	Qwen3-8B	57.6	57.8	58.0	**63.0**	+5.2 [+2.0, +8.6]*	+0.2 [−1.4, +1.8]	+0.4 [−0.8, +1.6]
	Qwen3-32B	66.4	66.4	66.4	**71.2**	+4.8 [+1.6, +8.2]*	+0.0 [−1.0, +1.0]	+0.0 [−0.8, +0.8]
2Wiki	Llama-3.1-8B	44.6	42.9	42.9	**58.1**	+15.2 [+9.6, +20.8]*	−1.7 [−4.6, +1.3]	−1.7 [−5.0, +1.7]
	Mistral-7B	33.7	33.7	33.7	**47.2**	+13.5 [+7.3, +19.8]*	+0.0 [−3.3, +3.3]	+0.0 [−4.3, +4.3]
	SmolLM3-3B	16.8	17.5	19.1	**42.9**	+25.4 [+19.1, +31.7]*	+0.7 [−3.0, +4.3]	+2.3 [−1.3, +5.9]
	Qwen3-8B	37.0	37.0	35.3	**40.9**	+4.0 [−0.7, +8.9]	+0.0 [−1.3, +1.3]	−1.7 [−4.0, +0.7]
	Qwen3-32B	50.2	50.8	50.8	**53.5**	+2.6 [−1.7, +7.3]	+0.7 [−0.7, +2.0]	+0.7 [−1.0, +2.3]
MuSiQue	Llama-3.1-8B	30.4	28.8	28.0	**50.1**	+21.3 [+16.7, +26.0]*	−1.6 [−4.0, +0.8]	−2.4 [−4.8, −0.2]*
	Mistral-7B	15.7	14.1	16.3	**43.7**	+29.6 [+24.9, +34.2]*	−1.6 [−3.4, +0.2]	+0.6 [−2.0, +3.2]
	SmolLM3-3B	8.5	9.9	10.1	**33.8**	+23.9 [+19.5, +28.4]*	+1.4 [−0.8, +3.8]	+1.6 [−0.6, +3.8]
	Qwen3-8B	18.9	18.9	18.5	**42.3**	+23.3 [+19.1, +27.6]*	+0.0 [−1.4, +1.4]	−0.4 [−2.0, +1.2]
	Qwen3-32B	36.8	35.6	37.2	**51.7**	+16.1 [+11.9, +20.3]*	−1.2 [−3.0, +0.6]	+0.4 [−0.8, +1.6]

C.3 Question categories

On LoCoMo, temporal questions—which need a session timestamp combined with a relative time expression from a different turn—gain significantly for all five readers and gain more than single-hop questions. The retrieval regime shows the same division of labour: on LoCoMo's global questions (temporal and open-domain, n=255) the standalone state beats top-3 BM25 for every reader (§E.3), and on top of the gold evidence sessions it still adds +3.7 to +21.1 points (Table 24). The first example below is a temporal question whose answer appears nowhere in the episode.

C.4 Worked examples

Table 14 shows held-out questions answered by the Qwen3-8B reader—the reader with the highest no-memory scores—in the no-context regime: the question alone in the prompt, first without any state and then with the Q4 state loaded. The two episode excerpts in Figure 4 are the only text the writer saw for the first two rows; the reader saw none of it.

LoCoMo conv-43, session 3 — 4:21 pm on 16 July, 2023

Tim: That looks peaceful! Do you have a favorite beach memory?
John: I had an awesome summer with my friends, surfing and riding the waves. The feeling was unreal!
Tim: Wow! How long have you been surfing?
John: I started surfing **five years ago** and it's been great. I love the connection to nature.

Question: What year did John start surfing? *Gold:* 2018

LongMemEval 37d43f65, session 1 — 2023/05/26

User: My laptop is a Dell Inspiron 15 5000 series from 2018, and I've been using it as my primary work laptop. Before the RAM upgrade to **16GB**, I was getting around 6–7 hours of battery life, but now I'm getting around 4–5 hours. I haven't made any other recent changes to the hardware or software aside from the RAM upgrade.

Question: How much RAM did I upgrade my laptop to? *Gold:* 16GB

Figure 4: **Episode excerpts behind the first two rows of Table 14.** The writer read the full conversation; the reader received only the question and the state.

Without the state the reader fails in three characteristic ways: a fluent parametric guess (1998; 32 GB; Clark Gable, Vivien Leigh's co-star rather than her husband; Calvert County, one of the counties the cooperative serves), an abstention ("Unknown"; "the conversation does not mention"), or a degenerate output ("Yes"). With the state, every answer is exact.

Three of the rows show what the state carries beyond a stored string. The LoCoMo answer, 2018, occurs nowhere in the conversation: it follows only from the session timestamp and the phrase "five years ago", two pieces of information from different positions of the episode that the state

Table 14: **Held-out questions, Qwen3-8B reader, question-only prompt.** "Without state" is the frozen backbone (B3); "with state" loads the 4.6 KB Q4 state through the trained head (D3). Predictions are verbatim; every with-state answer matches the gold answer.

Benchmark	Question	Without state	With state
LoCoMo (temporal)	What year did John start surfing?	1998	**2018**
LongMemEval (single-session)	How much RAM did I upgrade my laptop to?	32 GB	**16GB**
LongMemEval (single-session)	What size is my new Samsung TV?	The conversation does not mention the size of the Samsung TV.	**55-inch**
HotpotQA (comparison)	What took place first, the Korean War or the Western Allied invasion of Germany?	The Korean War	**The Western Allied invasion of Germany**
HotpotQA (bridge)	Candace Flynn was voiced by the actress who achieved mainstream success on what show?	Yes	**The Suite Life of Zack & Cody**
2Wiki (compositional)	Where was the place of death of the director of film *Basta Guardarla*?	Unknown	**Rome**
2Wiki (compositional)	What is the place of birth of the director of film *Gold and the Girl*?	Unknown	**New York**
MuSiQue (2-hop)	Who is the female star in *Gone with the Wind* married to?	Clark Gable	**Laurence Olivier**
MuSiQue (2-hop)	In which county is Southern Maryland Electric Cooperative headquartered?	Calvert County	**Charles County**

delivers together—the temporal composition that §C.3 measures across the whole benchmark. The HotpotQA comparison requires the relative dates of two events, neither of which is in the question. The five bridge and compositional questions each pass through an entity the question never names—Ashley Tisdale, Luciano Salce, Edmund Mortimer, Vivien Leigh, Hughesville—and the reader produces the second hop from a state written by the smaller Qwen3-4B backbone.

D CROSS-WRITER TRANSFER WITH A FIXED HEAD

The experiments in this appendix fit a head on one writer's states, freeze it, and read another writer's states through it; the only exception is the per-writer *refit* reference, labelled as such. Writers are Qwen3-4B-Instruct-2507 (Q4), Qwen3-8B (Q8), and Llama-3.1-8B-Instruct (Ll.); adapters, head optimization, and statistics follow §B. Unless varied explicitly, each head trains on Q4 states for 300 steps on LoCoMo or 150 on span QA, evaluation uses no prompt context, lift is D3−B-rand, and retention divides the foreign-writer lift by the same head's Q4 lift.

D.1 COORDINATE CONVENTIONS

Each writer estimates its convention from its own unlabeled states; no paired states, no reader, and no downstream label enter the estimate. The default calibration set is 1,000 HotpotQA training episodes, disjoint from evaluation. The convention is a per-layer quantity of the writer: the mean state and its SVD are taken over the writer's own layers, and the depth remap (6) only selects which writer layer each reader layer receives, so one canonical state serves every reader. With $\mathbf{M}_\ell = \mathbf{L}_\ell \mathbf{\Sigma}_\ell \mathbf{R}_\ell^\top$ the SVD of the layerwise mean state (singular values descending):

- **CANON**: $\tilde{\mathbf{S}}_\ell = (0.2/\|\mathbf{M}_\ell\|_F)\, \mathbf{L}_\ell^\top \mathbf{S}_\ell \mathbf{R}_\ell$. The mean becomes positive, diagonal, and ordered; the joint sign of each singular pair is fixed relative to the leading pair by the third moment of the writer's off-diagonal coefficients (Algorithm 1), the lowest-order odd statistic that is nonzero in this basis.

- **CANONP**: the same rotations with scale $c_\ell / \|\mathbf{M}_\ell\|_F$, where the reference profile c_ℓ is Q4's mean-state norm at each layer, published once and shared by all writers.
- **CANON2**: descending eigenbases of $\sum_i \mathbf{S}_i \mathbf{S}_i^\top$ and $\sum_i \mathbf{S}_i^\top \mathbf{S}_i$, paired signs with $\mathbf{l}_k^\top \mathbf{M} \mathbf{r}_k \geq 0$, joint signs anchored on the first pair by the first row and column of $\mathbf{L}^\top \mathbf{M} \mathbf{R}$, which are nonzero in this basis—the anchoring of Algorithm 1 with the first moment in place of the third—and scale $0.2 / \|\mathbf{M}_\ell\|_F$.
- **Controls**: *raw* states keep their native coordinates; *norm-only* applies the CANON scale without rotation; *gauge augmentation* trains the head under independent random layerwise orthogonal maps and log-uniform scales in $[0.2, 5]$ per sample and evaluates on raw states.

CANON and CANON2 require no shared reference at all; CANONP adds one per-layer norm profile to the format. LoCoMo real and random states receive the same convention; span-QA random states match the real state's per-module norm.

D.2 GEOMETRY OF THE STATE

Table 15 measures how the same episode is stored by different writers. In raw coordinates the states are orthogonal. Independent canonicalization—each writer alone—aligns them to cosine 0.81–0.89, within 0.08–0.12 of a paired per-layer Procrustes map fit on four episodes of both writers at once. One map shared across all layers reaches only 0.12–0.17: the gauge is per layer, exactly as the read path predicts.

Table 15: **Same-episode state cosine at 32 layers**, averaged over layers and six held-out LoCoMo conversations. Procrustes uses four paired calibration episodes; shared uses one map across all layers. Norm-only leaves the cosine unchanged; CANON2 differs from CANON by less than 0.01.

Writers	Raw	CANON	Procrustes (paired)	Shared map
Q4∼Q8	−0.012	+**0.887**	+0.965	+0.148
Q4∼Llama	−0.002	+**0.845**	+0.934	+0.151
Q8∼Llama	+0.023	+**0.814**	+0.933	+0.165

Calibration needs almost nothing: at 32 layers one state gives cosine 0.80–0.84 and four give 0.84–0.88. The fitted Procrustes maps place 0.41–0.44 of their squared Frobenius mass on the best signed permutation, showing that the fitted maps mix coordinates beyond signed permutations (§A.2).

Table 16 isolates what the convention has to fix. Matching norms without rotating axes transfers nothing, and a head trained under random gauges does not become gauge-invariant. Any self-consistent basis works: CANON2, built from second moments instead of the mean, transfers as well as CANON. And any writer can train the head: all nine training-writer × state-writer combinations lift the Llama reader by +14.9 to +16.6 with no diagonal advantage.

Table 16: **Coordinate controls and training-writer symmetry**, Llama-3.1-8B reader on LoCoMo. Lift in F1 points [95% CI]; columns identify the state writer. Raw, CANON, and CANONP rows for this reader appear in Table 17.

Control / training writer	Q4 states	Q8 states	Ll. states
Coordinate controls; head fitted on Q4			
Norm only	+16.6 [14.8, 18.4]*	+0.2 [−0.2, 0.6]	−0.1 [−0.5, 0.2]
Gauge augmentation	+1.7 [0.6, 2.8]*	+0.0 [−0.4, 0.4]	+0.2 [−0.3, 0.7]
CANON2	+16.7 [14.9, 18.5]*	+16.5 [14.7, 18.3]*	+16.9 [15.1, 18.7]*
Training-writer symmetry under CANON; row = training writer			
Qwen3-4B	+16.6 [14.8, 18.4]*	+16.6 [14.8, 18.4]*	+15.7 [13.9, 17.5]*
Qwen3-8B	+16.6 [14.7, 18.5]*	+16.1 [14.3, 17.9]*	+16.2 [14.4, 18.1]*
Llama-3.1-8B	+14.9 [13.2, 16.7]*	+16.4 [14.6, 18.2]*	+15.5 [13.8, 17.3]*

D.3 FIXED HEADS ACROSS READERS AND BENCHMARKS

Tables 17 and 18 give every per-benchmark cell behind Table 3, with intervals. Raw coordinates fail in all 16 cross-writer cells. CANON succeeds in 14/16 at 99% median retention, CANONP in 16/16 at 102%. The *refit* rows fit a fresh head in each foreign writer's raw coordinates: they show that every foreign state is decodable, and they are the yardstick a frozen head has to meet. CANONP meets or exceeds its own refit in 11 of the 16 cells.

Table 17: **Per-benchmark heads on LoCoMo**: lift in F1 points [95% CI], n=956. Raw results are from independent replication runs under the same protocol. Each coordinate arm fits a head on Q4 and freezes it; refit rows fit a separate head in each foreign writer's raw coordinates.

Coordinates	Q4 states (own)	Q8 states	Ll. states
Llama-3.1-8B			
Raw	$+14.8\,[13.1, 16.5]^*$	$-0.0\,[-0.3, 0.3]$	$-1.2\,[-1.9, -0.6]^*$
CANON	$+16.6\,[14.8, 18.4]^*$	$+16.6\,[14.8, 18.4]^*$	$+15.7\,[13.9, 17.5]^*$
CANONP	$+13.8\,[12.0, 15.6]^*$	$+14.6\,[12.9, 16.3]^*$	$+15.0\,[13.3, 16.8]^*$
Refit on each writer	$+14.8$	$+14.5\,[12.9, 16.2]^*$	$+14.2\,[12.5, 15.9]^*$
Mistral-7B			
Raw	$+7.4\,[5.8, 9.1]^*$	$+0.0\,[-0.7, 0.7]$	$-2.1\,[-3.1, -1.2]^*$
CANON	$+7.9\,[6.2, 9.7]^*$	$+6.6\,[5.0, 8.2]^*$	$+6.6\,[5.1, 8.2]^*$
CANONP	$+3.0\,[1.3, 4.7]^*$	$+6.3\,[4.7, 7.9]^*$	$+7.1\,[5.5, 8.7]^*$
Refit on each writer	$+7.4$	$+7.1\,[5.5, 8.8]^*$	$+3.8\,[2.3, 5.5]^*$
SmolLM3-3B			
Raw	$+5.6\,[4.0, 7.4]^*$	$-0.3\,[-0.8, 0.3]$	$-0.2\,[-0.7, 0.2]$
CANON	$+9.0\,[7.4, 10.6]^*$	$+9.4\,[7.8, 11.0]^*$	$+8.8\,[7.2, 10.4]^*$
CANONP	$+8.0\,[6.4, 9.7]^*$	$+6.5\,[4.8, 8.2]^*$	$+4.7\,[3.0, 6.5]^*$
Refit on each writer	$+5.6$	$+5.6\,[3.9, 7.3]^*$	$+5.3\,[3.7, 6.9]^*$
Qwen3-8B			
Raw	$+4.6\,[2.9, 6.3]^*$	$+0.0\,[-0.3, 0.4]$	$-0.1\,[-0.5, 0.4]$
CANON	$+1.3\,[0.1, 2.6]^*$	$+1.7\,[0.4, 3.0]^*$	$+1.2\,[-0.1, 2.4]$
CANONP	$+2.8\,[1.4, 4.3]^*$	$+2.9\,[1.5, 4.4]^*$	$+2.9\,[1.5, 4.4]^*$
Refit on each writer	$+4.6$	$+2.7\,[1.0, 4.4]^*$	$+2.0\,[0.5, 3.5]^*$
Qwen3-32B			
Raw	$+4.7\,[3.3, 6.1]^*$	$-0.2\,[-0.6, 0.2]$	$-0.1\,[-0.4, 0.2]$
CANON	$+0.5\,[0.0, 1.0]^*$	$+0.6\,[0.2, 1.1]^*$	$+0.3\,[-0.1, 0.9]$
CANONP	$+6.0\,[4.6, 7.4]^*$	$+5.2\,[3.8, 6.5]^*$	$+5.6\,[4.2, 6.9]^*$
Refit on each writer	$+4.7\,[3.3, 6.1]^*$	$+4.9\,[3.4, 6.3]^*$	$+4.1\,[2.9, 5.3]^*$

The norm profile is part of the format. Q4's mean-state norm grows from 0.188 at the first layer to 32.178 at the last, a 170× range. CANON's constant target of 0.2 therefore compresses the deep layers by more than 100×, and the two Qwen readers—whose no-memory scores are the highest—read little through it: +1.3 and +0.5 on their own writer. CANONP keeps the writer's depth profile and restores the readout to +2.8 and +6.0, and on MuSiQue it nearly triples the Llama reader's lift (Table 18, +4.0→+11.3). Both conventions transfer; CANONP transfers at full strength on every reader, which is why it is the format used for the universal heads below.

D.4 ONE HEAD FOR EVERY WRITER AND EVERY DOMAIN

For each reader, one ALL-5 head is trained for 400 steps with a balanced sampler on Q4 states from the five benchmark training splits (§F.3) and then frozen. It then reads held-out states from all three writers on all five benchmarks, 363 LongMemEval conversations included. Figure 5 shows three domains; Table 19 lists all 75 cells.

CANONP is significant in 24 of 25 in-writer cells and 44 of 50 cross-writer cells at 92% median retention; the same heads in raw coordinates transfer in 0 of 50. On LoCoMo, HotpotQA, 2Wiki, and MuSiQue the cross-writer count is 40 of 40 at 95% median retention. All six unstarred cross-writer cells are on LongMemEval, the benchmark with the shortest episodes and the highest no-memory scores (Mistral/Q8, SmolLM3/Q8 and /Ll., Qwen3-8B/Ll., Qwen3-32B/Q8 and /Ll.); Llama reads

Table 18: **Per-benchmark heads on span QA**: lift in EM points [95% CI], Llama-3.1-8B reader. Each coordinate arm fits a head on Q4 and freezes it; refit rows fit a separate head in each foreign writer's raw coordinates. Calibration is HotpotQA-train for all three benchmarks.

Coordinates	Q4 states (own)	Q8 states	Ll. states
HotpotQA			
Raw	$+12.2\,[8.8, 15.6]^*$	$-0.8\,[-2.0, 0.2]$	$-1.6\,[-3.0, -0.4]^*$
CANON	$+14.0\,[10.8, 17.4]^*$	$+14.0\,[10.8, 17.4]^*$	$+13.6\,[10.4, 17.0]^*$
CANONP	$+15.0\,[11.6, 18.6]^*$	$+15.2\,[11.8, 18.8]^*$	$+14.6\,[11.2, 18.0]^*$
Refit on each writer	$+12.2$	$+11.4\,[8.0, 14.8]^*$	$+13.4\,[10.0, 16.8]^*$
2Wiki			
Raw	$+16.2\,[11.6, 21.1]^*$	$-1.0\,[-2.6, 0.7]$	$-0.7\,[-2.0, 0.7]$
CANON	$+12.9\,[8.6, 17.2]^*$	$+12.9\,[8.6, 17.2]^*$	$+11.6\,[7.6, 15.8]^*$
CANONP	$+12.9\,[8.6, 17.5]^*$	$+13.5\,[8.9, 18.2]^*$	$+13.2\,[8.6, 17.8]^*$
Refit on each writer	$+16.2$	$+14.9\,[9.9, 19.8]^*$	$+13.9\,[9.6, 18.5]^*$
MuSiQue			
Raw	$+13.1\,[10.1, 16.1]^*$	$+0.0\,[0.0, 0.0]$	$-0.4\,[-1.2, 0.4]$
CANON	$+4.0\,[2.2, 6.0]^*$	$+4.4\,[2.6, 6.4]^*$	$+4.6\,[2.6, 6.6]^*$
CANONP	$+11.3\,[8.5, 14.1]^*$	$+10.9\,[8.0, 13.7]^*$	$+9.5\,[6.8, 12.1]^*$
Refit on each writer	$+13.1$	$+12.5\,[9.7, 15.5]^*$	$+8.2\,[5.8, 10.9]^*$

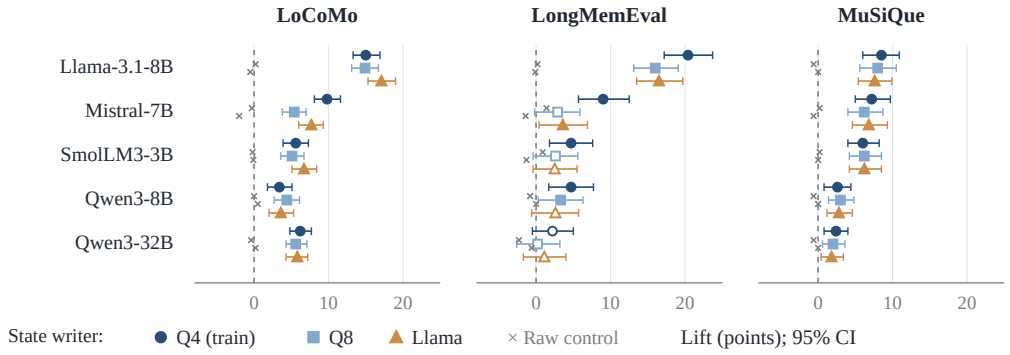

Figure 5: **One frozen head per reader reads three writers.** Lift (points) with 95% CIs for the CANONP ALL-5 heads on LoCoMo, LongMemEval, and MuSiQue; state writers Q4 (training writer), Q8, and Llama. Grey crosses: the same heads reading Q8 and Llama states in raw coordinates. Open markers: intervals that include zero.

all three writers on it at $+16.0$ to $+20.4$. The Llama reader's CANON ALL-5 head also succeeds in all 5 in-writer and 10 cross-writer cells at 98% median retention, with the lower MuSiQue lift of that convention ($+3.4$, $+3.8$).

D.5 CALIBRATION CORPORA

The convention is estimated from unlabeled states written from any text, so the calibration corpus is a free choice. To rule out shared calibration documents as a hidden anchor, each writer is calibrated on a different corpus outside every evaluation benchmark (Table 20): 300 documents per corpus with zero pairwise overlap of 200-character shingles. Code, mathematics, fiction, and medicine come from the OpenHermes-2.5 sources glaive-code-assist, metamath, airoboros2.2 creative, and CogStackMed, shuffled with seed 0 and packed to at least 6,000 characters; two independent letter-string corpora use seeds 1 and 2 and 6,002 characters per document. In arms A, B, and random letters Q4's basis, reference profile, and every head stay fixed. The final arm removes benchmark text from calibration altogether: Q4's basis and profile are estimated from code and its LoCoMo heads are retrained for 300 steps in that convention.

Table 19: **Universal ALL-5 heads**: CANONP lift in F1/EM points [95% CI], with the raw-coordinate foreign-writer lifts of the same heads. * marks an interval excluding zero, including negative raw effects. Unstarred CANONP cells are the six cross-writer cells whose interval includes zero.

Benchmark	CANONP: Q4	CANONP: Q8	CANONP: Ll.	Raw: Q8 / Ll.
Llama-3.1-8B				
LoCoMo	+15.0 [13.3, 16.9]*	+14.9 [13.1, 16.7]*	+17.1 [15.3, 19.0]*	+0.2 / −0.5*
LongMemEval	+20.4 [17.2, 23.7]*	+16.0 [13.1, 19.1]*	+16.5 [13.5, 19.7]*	+0.2 / −0.1
HotpotQA	+14.2 [10.6, 17.8]*	+14.4 [11.0, 18.0]*	+13.0 [9.6, 16.4]*	−3.2* / −1.2
2Wiki	+16.8 [12.2, 21.8]*	+17.5 [13.2, 22.1]*	+15.2 [10.9, 19.8]*	−0.7 / −0.3
MuSiQue	+8.5 [6.0, 10.9]*	+8.0 [5.6, 10.5]*	+7.6 [5.4, 9.9]*	−0.6 / +0.0
Mistral-7B				
LoCoMo	+9.8 [8.1, 11.6]*	+5.4 [3.8, 7.0]*	+7.7 [6.0, 9.3]*	−0.3 / −2.0*
LongMemEval	+9.0 [5.7, 12.5]*	+2.9 [−0.2, 5.9]	+3.6 [0.4, 6.9]*	+1.4 / −1.4
HotpotQA	+15.4 [12.0, 19.0]*	+15.2 [11.8, 18.6]*	+14.0 [10.6, 17.4]*	+1.4 / −6.4*
2Wiki	+14.5 [10.2, 18.8]*	+13.9 [9.6, 18.5]*	+12.5 [8.3, 16.8]*	+1.7 / −2.0
MuSiQue	+7.2 [5.0, 9.7]*	+6.2 [4.0, 8.7]*	+6.8 [4.6, 9.3]*	+0.2 / −0.6
SmolLM3-3B				
LoCoMo	+5.6 [3.9, 7.3]*	+5.1 [3.6, 6.7]*	+6.7 [5.1, 8.4]*	−0.2 / −0.1
LongMemEval	+4.7 [1.8, 7.6]*	+2.6 [−0.4, 5.6]	+2.5 [−0.4, 5.5]	+0.9 / −1.3*
HotpotQA	+15.4 [12.2, 18.6]*	+15.4 [12.2, 18.8]*	+14.0 [11.0, 17.2]*	+0.0 / +0.0
2Wiki	+11.6 [7.3, 15.8]*	+11.2 [6.9, 15.5]*	+10.6 [6.6, 14.9]*	+0.3 / +0.0
MuSiQue	+6.0 [4.0, 8.2]*	+6.2 [4.2, 8.5]*	+6.2 [4.2, 8.5]*	+0.2 / +0.0
Qwen3-8B				
LoCoMo	+3.4 [1.8, 5.1]*	+4.4 [2.7, 6.1]*	+3.6 [2.0, 5.3]*	+0.0 / +0.5
LongMemEval	+4.7 [1.7, 7.7]*	+3.3 [0.3, 6.3]*	+2.6 [−0.6, 5.7]	−0.8* / +0.0
HotpotQA	+3.6 [1.0, 6.2]*	+3.6 [1.0, 6.4]*	+3.8 [1.2, 6.4]*	+0.6 / +0.0
2Wiki	+8.3 [4.3, 12.2]*	+6.9 [3.3, 10.9]*	+7.6 [4.0, 11.6]*	+0.3 / +0.7
MuSiQue	+2.6 [0.8, 4.4]*	+3.0 [1.4, 4.8]*	+2.8 [1.2, 4.6]*	−0.6 / +0.0
Qwen3-32B				
LoCoMo	+6.2 [4.8, 7.7]*	+5.6 [4.3, 7.1]*	+5.8 [4.3, 7.2]*	−0.4 / +0.2
LongMemEval	+2.2 [−0.5, 5.0]	+0.2 [−2.6, 3.2]	+1.1 [−1.7, 4.0]	−2.3* / −0.6
HotpotQA	+4.2 [1.4, 7.0]*	+4.2 [1.4, 7.0]*	+3.6 [0.8, 6.4]*	−1.6* / −0.2
2Wiki	+4.0 [1.7, 6.6]*	+4.0 [1.7, 6.6]*	+4.0 [1.7, 6.6]*	+0.0 / +0.0
MuSiQue	+2.4 [0.8, 4.0]*	+2.0 [0.6, 3.6]*	+1.8 [0.4, 3.4]*	−0.6 / +0.0

Table 20: **Writer-specific calibration corpora.** The CANONP reference profile comes from Q4's listed corpus; only the final arm retrains heads.

Arm	Q4	Q8	Ll.	Heads
A	HotpotQA	Code	Fiction	Frozen
B	HotpotQA	Medicine	Mathematics	Frozen
Random letters	HotpotQA	Letters 1	Letters 2	Frozen
No benchmark calibration	Code	Mathematics	Medicine	Retrained

Table 21: **No benchmark calibration**: LoCoMo lift in F1 points [95% CI], final arm of Table 20. Heads are frozen after retraining in Q4's code-calibrated convention.

Reader	Q4 states (own)	Q8 states	Ll. states
Llama-3.1-8B	+14.8 [13.0, 16.5]*	+16.2 [14.4, 18.1]*	+14.3 [12.5, 16.1]*
Mistral-7B	+5.9 [4.1, 7.6]*	+4.7 [3.1, 6.2]*	+5.3 [3.9, 6.9]*
SmolLM3-3B	+7.0 [5.3, 8.7]*	+8.1 [6.5, 9.7]*	+6.9 [5.2, 8.6]*
Qwen3-8B	+2.5 [1.1, 4.0]*	+3.4 [2.0, 4.7]*	+2.8 [1.6, 4.1]*
Qwen3-32B	+5.7 [4.3, 7.0]*	+3.3 [2.3, 4.3]*	+3.7 [2.5, 4.8]*

The bases move with the corpus—mean absolute diagonal overlap with the HotpotQA basis is 0.45–0.69—yet the re-expressed LoCoMo states keep cosine 0.74–0.95, and cross-writer same-episode

cosine stays at 0.83–0.91 under disjoint natural-language calibration, 0.75–0.87 under random letters, and 0.85–0.90 with no benchmark calibration (32 layers). Transfer follows: 110 of 112 tested cells are significantly positive (Table 4c; arm A covers 10 LoCoMo, 6 span-QA, and 50 universal-head cells, arm B 10 LoCoMo cells, random letters 10 LoCoMo, 6 span-QA, and 10 Llama universal-head cells, and the benchmark-free arm 10 LoCoMo cells), at 91–98% median retention and a median lift change of at most -0.4 points from shared HotpotQA calibration. The two remaining cells are Qwen3-32B ALL-5 on LongMemEval in arm A ($+2.1$ [-0.3, 4.6] on Q8; -0.6 [-3.2, 2.0] on Ll.), the same two cells that are unstarred under shared calibration in Table 19.

With no benchmark text in any convention (Table 21), all ten cross-writer lifts are significantly positive, at retention 58% to 132%. Heads trained on Q4 states calibrated on code read Q8 states calibrated on mathematics and Llama states calibrated on medicine: the convention travels with the adapter, independent of the text either party saw.

E RETRIEVAL COMPLEMENTARITY

The retrieval experiments hold the retrieved text fixed and add the state on top, so every gain is measured at zero additional context tokens. Both retrievers return three sessions unless stated. On LoCoMo a *standalone* condition drops the retrieved text but keeps the operating point's 16-session prompt window. The gold sessions are the benchmark's annotated evidence sessions. Metrics and splits follow §B.3; every interval is a 95% paired bootstrap over 10^4 resamples.

E.1 LOCOMO WITH A 16-SESSION WINDOW

Table 22 gives the absolute scores behind Table 5a. Adding the state improves BM25 and BGE-M3 for every reader. For Mistral, SmolLM3, and Qwen3-32B the state *alone*—no retrieved text, the same 16-session window—already outscores top-3 BM25 (.422 vs .404, .395 vs .303, .438 vs .371), and stacking the two carriers raises these readers by $+.116$, $+.188$, and $+.207$ over BM25.

Table 22: **LoCoMo, 16-session prompt window, graded accuracy** (n=956). *base* and *state* omit retrieved text; all modes keep the window; BM25 and BGE-M3 use k=3. *: CI excludes zero.

Reader	base	state	BM25	+state	BGE-M3	+state
Llama-3.1-8B	.385	.447	.552	**.579**	.533	.563
Mistral-7B	.296	.422	.404	**.520**	.392	.509
SmolLM3-3B	.222	.395	.303	**.491**	.291	.480
Qwen3-8B	.368	.419	.508	**.545**	.498	.526
Qwen3-32B	.272	.438	.371	**.578**	.373	.570

Reader	gain over BM25	gain over BGE-M3
Llama-3.1-8B	+.027 [.009,.046]*	+.030 [.011,.049]*
Mistral-7B	+.116 [.094,.138]*	+.117 [.095,.138]*
SmolLM3-3B	+.188 [.166,.210]*	+.189 [.167,.212]*
Qwen3-8B	+.037 [.018,.056]*	+.028 [.010,.047]*
Qwen3-32B	+.207 [.186,.227]*	+.196 [.176,.217]*

E.2 LONGMEMEVAL-FULL: 121K-TOKEN HISTORIES

LongMemEval-full instances hold 38–62 sessions (median 47, about 121K tokens); the state is written from the whole history in one 30-second scan (§B.6), and the prompt carries only the three retrieved sessions. Table 23 gives the absolute scores behind Table 5b. Adding the state lifts every reader under both retrievers, by $+3.1$ to $+18.2$ F1, and Mistral more than doubles its retrieval score (13.9→30.6).

Table 23: **LongMemEval-full, global questions** (n=205, F1 $\times$100). Both retrievers use k=3; the retrieval-prompt controls are in Table 24. *: CI excludes zero; all ten retrieval gains survive Holm and Benjamini–Hochberg correction.

Reader	base	BM25	+state	BGE-M3	+state
Llama-3.1-8B	1.1	21.5	**28.2**	21.5	27.2
Mistral-7B	10.6	13.9	30.6	13.6	**31.8**
SmolLM3-3B	15.0	13.2	**19.0**	12.5	**19.0**
Qwen3-8B	11.6	18.7	**23.3**	19.3	22.4
Qwen3-32B	17.7	14.9	24.7	15.0	**26.6**

Reader	gain over BM25	gain over BGE-M3
Llama-3.1-8B	+6.7 [2.2,11.3]*	+5.7 [1.5,10.2]*
Mistral-7B	+16.7 [12.1,21.5]*	+18.2 [13.5,23.0]*
SmolLM3-3B	+5.8 [2.9,8.8]*	+6.5 [3.4,9.8]*
Qwen3-8B	+4.5 [1.2,7.9]*	+3.1 [0.0,6.3]*
Qwen3-32B	+9.7 [6.2,13.7]*	+11.6 [7.8,15.7]*

E.3 Question types, retrieval depth, and window size

Where the state leads. On LoCoMo's global questions (temporal and open-domain, n=255) the standalone state exceeds top-3 BM25 by .046–.154 graded accuracy for every reader, all significant, and exceeds BGE-M3 significantly for four (Qwen3-8B: +.035). On LongMemEval-full, where BM25 covers .64–.70 of the gold sessions for the global question types, the retrieval gain is positive in all ten reader $\times$ type cells and significant in eight (the two intervals that include zero are Llama on multi-session and Qwen3-8B on temporal questions).

Where retrieval leads. On question types that a retrieved slice answers directly—LoCoMo single-hop and multi-hop questions, and LongMemEval single-hop and knowledge-update questions, whose gold sessions top-3 retrieval covers completely—the retrieved text is the stronger standalone channel. The two carriers therefore divide the question space, and the combination inherits both halves: the retrieval gain from adding the state is positive at every setting tested.

Sweeps. On LongMemEval-full the gain over BM25 stays positive for every reader at $k \in \{1, 3, 5\}$, spanning 0.7–16.7 F1; on LoCoMo the retrieval gain is positive at every prompt window of 8, 16, and 24 sessions and every k.

E.4 Controls inside the retrieval prompt, and gold evidence

Table 24 runs the content controls with the retrieved sessions in the prompt. The real state beats a norm-matched random state in all ten cells, by +3.5 to +20.5 points; random states and ablated heads stay within −1.8 to +0.8 of retrieval alone. The last column adds the state on top of the annotated gold-evidence sessions: every reader improves on both benchmarks, 10 of 10 significant, by +3.7 to +21.1 on LoCoMo and +5.5 to +17.9 F1 on LongMemEval-full.

E.5 Worked examples

Table 25 follows four held-out questions through the Qwen3-32B reader with the retrieved sessions held byte-identical and only the state added; Figure 6 prints the retrieved and the missing turns for two of them.

What retrieval provides. In every row the retriever returns the turn that names the answer entity or the most recent event: the session in which the user reaches Premier Silver, the one grocery trip with an amount, the day Pixie arrives, two of the three plant purchases. This is the local evidence a top-k slice is good at, and the reader answers from it: it reports the store it sees, and where the slice is silent it abstains ("not mentioned", "did not specify") or fills the gap with invented plants.

What the state adds. The state is written from the whole 38–62-session history, and on these questions it supplies what the slice does not hold: the order of events across sessions (Silver in

Table 24: **Booster controls.** *real − random*: (BM25+state) − (BM25+random state); *random booster*: (BM25+random state) − BM25; *ablated booster*: (BM25+state through an ablated head) − BM25; *gold+state − gold*: the state added on top of the gold-evidence sessions. LoCoMo: 16-session window, graded accuracy ×100, n=956; LongMemEval-full: global questions, F1 ×100, n=205. 95% CIs; *: CI excludes zero.

Benchmark	Reader	real − random booster	random booster	ablated booster	gold+state − gold
LongMemEval-full	Llama-3.1-8B	+7.4 [3.0, 11.8]*	−0.7 [−2.8, 1.2]	−1.2 [−3.9, 1.6]	+6.3 [1.6, 11.1]*
	Mistral-7B	+17.1 [12.5, 21.8]*	−0.4 [−2.1, 1.2]	−0.5 [−2.7, 1.6]	+17.9 [12.9, 23.1]*
	SmolLM3-3B	+6.1 [3.3, 9.3]*	−0.4 [−2.1, 1.2]	−0.1 [−2.2, 2.1]	+6.6 [3.5, 9.9]*
	Qwen3-8B	+4.3 [1.1, 7.6]*	+0.3 [−0.6, 1.1]	−0.2 [−1.4, 0.9]	+5.5 [1.9, 9.2]*
	Qwen3-32B	+9.1 [5.6, 13.0]*	+0.6 [−0.2, 1.7]	+0.3 [−0.4, 1.1]	+11.3 [7.7, 15.3]*
LoCoMo, 16 sessions	Llama-3.1-8B	+3.5 [1.7, 5.3]*	−0.8 [−1.8, 0.2]	−0.5 [−1.6, +0.5]	+3.8 [1.9, 5.6]*
	Mistral-7B	+13.4 [11.2, 15.6]*	−1.8 [−3.2, −0.5]*	+0.3 [−1.1, +1.6]	+12.5 [10.4, 14.7]*
	SmolLM3-3B	+19.3 [17.1, 21.4]*	−0.5 [−1.5, 0.6]	−1.3 [−2.5, −0.2]*	+21.1 [19.0, 23.3]*
	Qwen3-8B	+3.9 [2.0, 5.8]*	−0.2 [−0.8, 0.3]	−0.4 [−1.0, +0.2]	+3.7 [1.8, 5.7]*
	Qwen3-32B	+20.5 [18.4, 22.5]*	+0.2 [−0.4, 0.8]	+0.8 [+0.3, +1.4]*	+21.1 [19.0, 23.2]*

Table 25: **Retrieval with and without the state, Qwen3-32B reader.** The retrieved sessions are byte-identical in the last two columns; the state is the only addition. k is the retrieval depth of the row, taken from the sweeps of §E.3. Predictions are verbatim apart from shortening at the ellipses and removal of markdown emphasis; bold marks a prediction that matches the gold answer.

Question	Retrieved slice	Retrieval	Retrieval + state
LongMemEval, BM25 k=1. What was my previous frequent flyer status on United Airlines before I got the current status? *Gold:* Premier Silver	*Holds* S20 (Sept. 2022): "finally eligible for Premier Silver status". *Lacks* S43 (May 2023): "I just reached Premier Gold status".	Your previous frequent flyer status on United Airlines was not mentioned in the conversation.	**Premier Silver**
LongMemEval, BM25 k=1. Which grocery store did I spend the most money at in the past month? *Gold:* Thrive Market	*Holds* S45: Trader Joe's, $80; Thrive Market order mentioned, no amount. *Lacks* S31: Thrive Market, $150; S33: Walmart, $120; S49: Publix, $60.	Trader Joe's	**Thrive Market**
LongMemEval, BGE-M3 k=3. How many plants did I acquire in the last month? *Gold:* 3	*Holds* S2: peace lily and succulent "from the nursery two weeks ago"; S14: snake plant and succulent named, no date; S18: unrelated. *Lacks* S17: snake plant "from my sister last month".	You acquired a peace lily, a succulent, a rose bush, a snake plant, an orchid, a spider plant, and a fern … That totals to seven plants.	**You acquired 3 plants in the last month: a peace lily, a succulent, and a snake plant.**
LoCoMo conv-44, BM25 k=5, 16-session window. How many years passed between Audrey adopting Pixie and her other three dogs? *Gold:* three years	*Holds* D2 (2 Apr. 2023): "We adopted another puppy called Pixie"; D10: "They're all 3-year-old". *Lacks* D1 (27 Mar. 2023): "I've had them for 3 years! Their names are Pepper, Precious and Panda"; D1 is also outside the window (D13–D28).	Audrey mentioned adopting Pixie … on 2 April 2023. She did not specify the exact adoption dates for her other three dogs …	**3 years**

September 2022, Gold in May 2023, so Silver is the *previous* status), quantities scattered over sessions ($150 at Thrive Market against $80, $120, and $60 elsewhere), acquisition times bound to entities (the snake plant arrived last month, the succulent two weeks ago, the plants of S14 are old), and durations anchored to dates (three years with Pepper, Precious, and Panda before Pixie). Each

LongMemEval 50635ada

Session 20 — 2022/09/16 (retrieved, BM25 top-1)
User: I actually just hit 20,000 miles on United Airlines, which means I'm finally eligible for **Premier Silver** status. Can you tell me how many miles I need to book a one-way economy ticket to LA from DC?

Session 43 — 2023/05/30 (not retrieved)
User: … I do have a United Airlines MileagePlus account—I just reached **Premier Gold** status.

Question: What was my previous frequent flyer status on United Airlines before I got the current status? *Gold:* Premier Silver

LongMemEval gpt4_2ba83207

Session 45 — 2023/05/29 (retrieved, BM25 top-1)
User: … my sister and I went to **Trader Joe's** the week before last and spent around **$80** between the two of us … have you heard about Thrive Market? I placed an online order with them last month and got some great deals …

Sessions 31, 33, 49 — 2023/05/26–05/30 (not retrieved)
User: … placed an online order with **Thrive Market** last month and spent around **$150** … spent around **$120** at **Walmart** … ordered from **Publix** last week and spent around **$60**.

Question: Which grocery store did I spend the most money at in the past month? *Gold:* Thrive Market

Figure 6: **Retrieved and missing turns behind the first two rows of Table 25.** The reader's prompt held only the retrieved session; the writer had scanned the whole 38–62-session history into the state.

of these needs turns from two or more sessions at once, and none of the four retrieved slices contains them.

Why the combination scores higher. With the state in place the reader stops abstaining and stops answering from the slice alone: the same retrieved text now yields "Premier Silver", Thrive Market, the three plants by name, and "3 years". The retrieved turn still does its part—it pins the answer to the wording and entities in the text, the tier name, the store name, the adoption date the duration is counted from—while the state supplies the cross-session structure that decides which of them is the answer. On these four questions the division is explicit: retrieval locates the turn, the state supplies facts from sessions the retriever did not return, and the reader uses both; the aggregate counterpart is the +3.1 to +18.2 F1 of Table 23 and the larger gains on temporal and multi-session questions in §E.3.

F Ablations and head generalization

The head is the only trained component, and this appendix varies everything about it: the objective, the data budget, the norm profile of the coordinate convention, the domains it sees, and the seed. The state and the backbones are fixed throughout.

F.1 Objective and data budget

Cross-entropy on the answer tokens is the whole objective. Table 26 compares it with distillation toward the writer's native read on Qwen3-8B in the full-history regime. Cross-entropy alone gives the largest lift; adding a distillation term lowers it by more than a quarter, and distillation alone does not move the reader. A single training conversation is already enough for a significant lift.

Table 26: **Objective and data budget**, Qwen3-8B reader on LoCoMo full-history (n=945). Lift over the no-memory score 0.4986, with paired 95% CIs.

Head training	lift	95% CI
cross-entropy only, 4 conversations	**+.0404***	[+.0211, +.0599]
cross-entropy + distillation, 4 conversations	+.0289*	[+.0165, +.0419]
distillation only, 4 conversations	+.0057	[−.0022, +.0141]
cross-entropy + distillation, 1 conversation	+.0250*	[+.0119, +.0380]

The writer's own read is a weak teacher: its state-only readout scores 8.1 F1 (×100) on LoCoMo (Table 2), and a head fit directly to the benchmark's answers has nothing to gain from matching it.

F.2 NORM PROFILE

One further ablation is reported with the cross-writer tables. The coordinate scale is the one design choice with a large effect: replacing CANON's constant target with the writer's own depth profile (CANONP) raises the Qwen3-32B in-writer LoCoMo lift from $+0.5$ to $+6.0$ and the Llama reader's MuSiQue lift from $+4.0$ to $+11.3$ (§D.3).

F.3 ONE HEAD ACROSS DOMAINS

A single head per reader, fit jointly on the five benchmark training splits in native Q4 coordinates, is significant in 24 of 25 held-out cells at 106% median retention relative to the per-domain heads: pooling domains costs nothing and the head no longer needs to know the benchmark. Its CANONP counterpart is the universal head of §D.4, which reads every writer on every domain (Table 19).

F.4 SEEDS, TRAINING SUBSETS, AND MULTIPLE COMPARISONS

With the writer states fixed, three independent head-training seeds reproduce every LoCoMo lift (Table 27), with across-seed standard deviations of 0.10–1.15 points. Replacing the four LoCoMo training conversations of Qwen3-8B by two other four-conversation subsets gives lifts of $+.0516$ [$+.0347$, $+.0688$] and $+.0515$ [$+.0344$, $+.0685$] against the canonical $+.0516$ (these resamples score n=898 and 875 held-out questions respectively, so they are separate evaluations rather than paired contrasts on the canonical set). Under Benjamini–Hochberg control the 25 headline no-context cells keep 21 significant results and under Holm 20 (§B.5).

Table 27: **Seed robustness.** LoCoMo no-context lift (D3 − B-rand, $\times 100$), mean $\pm$ standard deviation over three head-training seeds; TSW writer, n=956.

	Llama-3.1-8B	Mistral-7B	SmolLM3-3B	Qwen3-8B	Qwen3-32B
Mean $\pm$ std	$+13.94 \pm 0.90$	$+6.52 \pm 1.15$	$+5.49 \pm 0.10$	$+3.71 \pm 0.88$	$+5.06 \pm 0.74$

G CONTENT AND CARRIER CONTROLS

Two questions are settled here for every reader and benchmark: whether the lifts survive replacing the written state by a random one, at evaluation and in the head's training, and how the 4.6 KB state compares with a text summary of the same episode.

G.1 RANDOM STATES, UNTRAINED HEADS, AND ABLATED HEADS

Table 6 reports the ablated-head control for all five readers. A head trained and evaluated on norm-matched random states (*abl.*) stays within 2.4 points of the no-memory score in every one of the 25 cells. The real state lifts Llama from 1.6 to 16.4 F1 ($\times 100$) on LoCoMo and from 1.0 to 18.0 on LongMemEval, and lifts every reader on every span benchmark.

The other two controls agree. A norm-matched random state read through the trained head (B-rand) is indistinguishable from no memory: for Qwen3-8B in LoCoMo's 16-session regime (graded accuracy, n=956), B-rand$-$B3 $= -0.0001$ [-0.0069, $+0.0064$], and an untrained head on the real state (B2) gives B2$-$B3 $= +0.0001$ [-0.0052, $+0.0047$]. Across the 15 span-QA reader–dataset cells, B-rand$-$B3 lies within -2.4 to $+1.5$ F1. The same separation holds with passages in the prompt (Table 13) and inside the retrieval prompt (Table 24). The two controls remove different things: the random state removes the content of a written state and keeps its norm; the ablated head removes every written state from training as well as evaluation, so its scores are what the question–answer supervision yields when no written state is ever present.

G.2 TEXT CARRIERS

Table 2 places the ported state beside three alternatives at zero context cost on LoCoMo: no memory, a norm-matched random state, and Q4's own projections reading the same state on Q4, fit on QASPER answers and on no benchmark used here. The paired contrasts behind that table follow.

The strongest text carrier is a summary. The Qwen3-4B backbone, without its memory adapter, summarizes each held-out LoCoMo conversation once; the summary is truncated to 128 or 512 tokens and compared with the state alone and with summary plus the same state and head, nothing retrained (Table 28). At zero context tokens the state beats a 128-token summary for every reader, by $+4.9$ to $+10.0$ F1, and matches a 512-token summary for every reader: the 4.6 KB state stands in for 512 prompt positions. Summary and state are complementary: their combination improves on the summary alone for all ten reader–budget pairs and reaches 21.5–28.7 F1 at 512 tokens, $+6.2$ to $+13.1$ over the summary. Stacking the two carriers is the best configuration in eight of ten cells; the two exceptions are the Qwen readers at 128 tokens, where the state alone scores higher than the short summary with the state added.

Table 28: **LoCoMo summary baseline, paired contrasts** (F1 $\times 100$, n=956; absolute scores in Table 2). Both budgets use the same stored state and head. Paired 95% CIs; *: CI excludes zero. Bold marks the best absolute score in each row.

	absolute F1 ($\times 100$)			contrast [95% CI]	
Reader	summary	state	summ.+state	state − summary	(summ.+state) − summary
128-token summary					
Llama-3.1-8B	7.0	16.4	**20.8**	+9.4 [7.4, 11.4]*	+13.9 [12.1, 15.7]*
Mistral-7B	6.5	16.5	**19.3**	+10.0 [8.0, 11.9]*	+12.8 [11.2, 14.5]*
SmolLM3-3B	10.5	15.4	**18.0**	+4.9 [3.1, 6.7]*	+7.5 [6.0, 9.0]*
Qwen3-8B	7.7	**17.0**	16.0	+9.3 [7.3, 11.3]*	+8.3 [6.8, 9.8]*
Qwen3-32B	8.5	**16.1**	14.4	+7.5 [5.7, 9.4]*	+5.9 [4.8, 7.1]*
512-token summary					
Llama-3.1-8B	16.2	16.4	**28.7**	+0.2 [−2.3, 2.5]	+12.5 [10.6, 14.3]*
Mistral-7B	14.3	16.5	**27.4**	+2.2 [−0.1, 4.4]	+13.1 [11.3, 14.9]*
SmolLM3-3B	15.4	15.4	**25.2**	+0.0 [−2.0, 2.0]	+9.8 [8.2, 11.5]*
Qwen3-8B	16.2	17.0	**24.0**	+0.8 [−1.5, 3.0]	+7.9 [6.4, 9.4]*
Qwen3-32B	15.3	16.1	**21.5**	+0.8 [−1.4, 3.0]	+6.2 [5.0, 7.5]*

H SCOPE AND OPEN PROBLEMS

The results establish the transplant for delta-rule states at one rank, r=8, with writers of 4B and 8B parameters and readers of 3B to 32B. Four directions follow directly from the tables.

Memory carriers. The delta-rule read absorbs invertible value maps and orthogonal key maps (§A.1). A mean-state SVD, paired signs, and normalization fix the rotation–scale subgroup and yield the cross-writer transfer measured here. Recurrent states, learned latent memories, and compressed-context tokens each have a read interface and a corresponding gauge group. Extending the protocol requires characterizing that group, measuring the transformations relating independently trained writers, and selecting a convention from each writer's own statistics. A writer-side convention plus a thin per-reader head provides the template for testing which carriers support shared coordinates and reusable memories across backbone families.

Rank and writer scale. The state is $L_w r^2$ numbers and the head is $rL_i(2d_i + d_i^q)$: state storage grows quadratically in r, and head size linearly. Relative to rank 8, ranks 16 and 32 use $4\times$ and $16\times$ the state storage, and $2\times$ and $4\times$ the head parameters. The gauge argument of §A.1 holds at any rank. Whether higher rank widens the gap to text carriers on the longest histories is the natural next measurement, as is writing with a larger backbone.

Supervision for the head. Every head in this study is fit on question–answer pairs from the benchmark family it serves, or on a pooled mixture of the five (§F.3). A head fit on benchmark-free

text alone that reaches the same readout for every reader family would remove the last benchmark-specific step from the protocol, and is the natural next experiment.

Short episodes with strong readers. Every non-significant cross-writer cell of the universal heads falls on LongMemEval-oracle, whose episodes hold at most six sessions and whose no-memory scores are among the highest in the study; on the 121K-token variant of the same benchmark the state adds $+3.1$ to $+18.2$ F1 on top of retrieval for every reader (§E.2). The state earns most where the episode is long and the prompt is short, which is the regime long-running agents live in.