MEMHUB: ZERO-SHOT LATENT MEMORY TRANSFER ACROSS HETEROGENEOUS LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct latent communication between language models currently takes two forms: pairwise projections, whose training cost grows quadratically with the number of models, and shared representation spaces, which admit new members solely through existing ones and have been shown only within one model family. We introduce MEMHUB, a memory hub in which each frozen model independently trains one Writer–Reader adapter pair against a small frozen anchor that fixes a shared latent coordinate system. Members never observe one another's parameters, activations, or training, yet the persistent artifacts they write are mutually readable zero-shot. Across nine members from five heterogeneous families (0.5B–8B), MEMHUB recovers a median 96.1% of pair-specific readout gains over all 72 foreign directions, retains 97% of self-transfer under dataset shift, and onboards two held-out families with one adapter pair each and no incumbent retraining. Two hubs retrained from scratch read the first hub's artifacts, and each other's, at parity over 432 cross-hub directions: the interface belongs to the anchor, not to any training run. Zero-shot reads outperform a pair-trained C2C channel overall. The artifact is a fixed 64 KB object read at prefill cost: 21–33$\times$ faster to first token than a live text handoff and higher quality than a text channel trained at matched budget. Controlled factorials isolate the mechanism: the geometric constraint alone fixes the shared coordinate system, and anchoring two of 32 slots aligns the rest.

1 INTRODUCTION

Multi-model LLM systems still coordinate the way single models talk to people: through text (Wu et al., 2024; Du et al., 2024). Every exchange forces the sender to serialize its internal state into tokens and the receiver to re-encode them. Latent communication removes this bottleneck by transmitting internal state directly—fusing KV caches (Fu et al., 2026), compressing the exchange through a cache bottleneck (Rossi et al., 2026), or aligning caches in a learned shared space (Dery et al., 2026)—and beats text on both accuracy and latency.

What no latent channel has reproduced are the two properties that made text the default: text is *universal*—any model can read what any other wrote, including models that did not yet exist—and *persistent*— written once, read any number of times. Existing channels trade both away. Pairwise designs (Fu et al., 2026; Rossi et al., 2026) train a translator per ordered pair (Figure 1a), so cost grows as $O(N^2)$ and no unseen pairing works; shared-space designs (Dery et al., 2026) train each new member through frozen incumbents (Figure 1b) and have been shown within one family. In every case the transmitted state is transient, and a recent survey of latent communication still lists a hub-style codec for unseen pairings as an open problem (Liu, 2026).

We show that universality and persistence require not tokens but a fixed, public coordinate system. MEMHUB (Figure 1c) publishes a small anchor—a frozen text embedder and a frozen judge Reader—that defines the coordinates of a slot-structured memory space. Each frozen model then trains one Writer–Reader adapter pair against the anchor alone, in isolation from every other member; joining costs one adapter pair, and no incumbent is ever modified. The Writer's output is a serialized, position-independent artifact, stored once and read by every current and future member.

Why should adapters that never met be compatible? The obstruction is coordinates, not capability. Task loss constrains what a Writer must encode but not the basis it encodes it in: every invertible change of slot basis is absorbed exactly by the same model's Reader, so a member's own task loss

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.

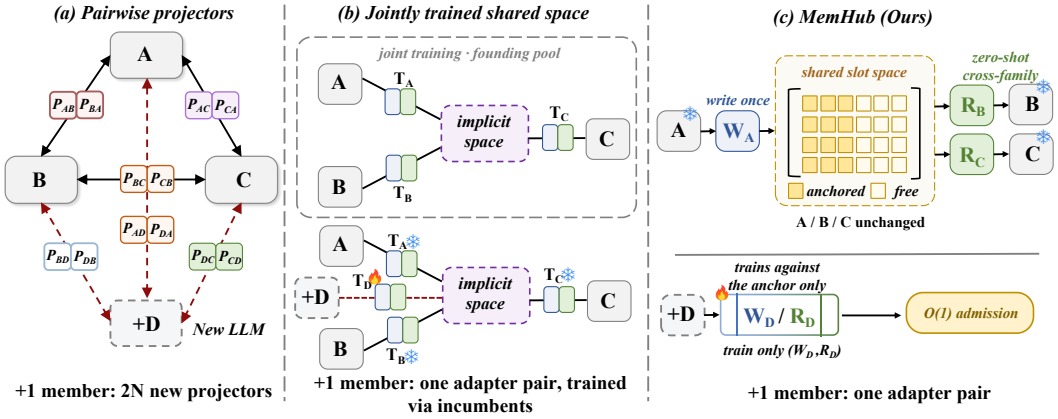


Figure 1: **Three ways to coordinate latent communication, and what each charges to admit a new model D. (a)** Pairwise projectors (Fu et al., 2026): $2N$ new translators. **(b)** Jointly trained shared space (Dery et al., 2026): T_D trained through frozen incumbents. **(c)** MEMHUB: a public frozen anchor fixes the slot format (anchored slots shaded); A writes once, R_B and R_C read zero-shot, and D trains (W_D, R_D) against the anchor alone.

leaves its basis free (Proposition 1, Appendix A), and members trained without the anchor solve their task yet score at most 0.4 F1 on every cross pair. The anchor removes exactly this freedom: one basis change acts on all 32 slots at once, so pinning two of them to frozen embedding targets fixes the basis of the 30 free slots as well (Proposition 2), their content aligns with it (top-1 retrieval .005 $\rightarrow$ .888), and cross-model transfer lifts from zero to its plateau. Conversely, a single linear map fit to 50 unlabeled texts restores transfer between unanchored members, and 950 more add nothing.

The interface holds up at scale: nine members from five heterogeneous families (0.5B–8B) recover a median 96.1% of pair-specific training gains over all 72 foreign directions (Table 1) at under 30 H800-hours for the whole hub, and two hubs retrained from scratch read that hub's artifacts, and each other's, at parity over 432 directions (median difference 0.0 F1) (Table 2). Because all coordination lives in the frozen anchor, the hub is closer to a file format than to a deployed system: publishing the anchor is the entire act of standardization. Our contributions are threefold:

- **A universal latent-memory interface from fully independent training.** To our knowledge, the first latent-memory interface enabling zero-shot cross-family reading through independently trained Writer–Reader adapters; each frozen model trains against a public anchor alone, without accessing any other member. One frozen anchor replaces $O(N^2)$ pairwise translators, and joining involves no incumbent.
- **Latent memory as a persistent, reusable artifact.** A serialized, position-independent 64 KB object that survives quantization and position transport, composes across writers, and supports repeated reads at prefill cost, with 21–33$\times$ lower time to first token than a live text handoff.
- **Broad, statistically grounded evidence.** A nine-member, five-family transfer matrix in which zero-shot reads recover a median 96.1% of pair-specific readout gains across all 72 foreign directions; the same frozen adapters retain 97% of self-transfer on two unseen QA distributions, admit two held-out families without retraining, and outperform a pair-trained C2C channel.

2 RELATED WORK

2.1 LATENT COMMUNICATION BETWEEN LANGUAGE MODELS

Agents have inherited text as their medium (Wu et al., 2024; Du et al., 2024); work that moves the exchange into latent space sorts by the coordination it presupposes (Figure 1). Pairwise channels train a translator per ordered pair: C2C projects and fuses KV caches (Fu et al., 2026), LCF routes the exchange through a cache bottleneck (Rossi et al., 2026), and activation grafting splices hidden states into another model's forward pass (Ramesh & Li, 2025). Embedding-level debate assumes a shared vocabulary (Pham et al., 2024) and training-free cache handoff a single architecture (Zou et al., 2026). Shared-space designs learn a common space jointly and admit later members through

frozen incumbents (Dery et al., 2026, §5), within one family. In each case the coordination object is private—pair weights, the incumbents' adapters, or the model itself—and the transmitted state is consumed once. MEMHUB supplies a public coordination object instead—a frozen anchor against which every member trains alone—giving membership that involves no incumbent, across heterogeneous families, and persistent artifacts rather than transmissions.

2.2 PERSISTENT LATENT MEMORY AND REPRESENTATION ALIGNMENT

A complementary line treats latent state as memory. Soft prompts and gist tokens compress instructions into trained vectors (Lester et al., 2021; Mu et al., 2023); in-context autoencoders and Cartridges distill long contexts into reusable KV caches (Ge et al., 2024; Eyuboglu et al., 2025). These artifacts persist only in their producer's private basis, whereas agent memory systems persist state as text so that any model can read it back (Packer et al., 2023; Yu et al., 2026), at the serialization cost latent channels remove. Cross-model readability is the question of representation alignment: stitching shows independently trained networks are linearly translatable (Lenc & Vedaldi, 2015; Bansal et al., 2021) and converge with scale (Huh et al., 2024), and unsupervised translators recover maps between embedding spaces from geometry alone (Jha et al., 2025). Closest to our mechanism, relative representations express embeddings in coordinates defined by anchor samples, enabling zero-shot stitching of encoders (Moschella et al., 2023). All establish alignment after training; MEMHUB instead fixes a public coordinate system before any member trains, making independently written artifacts readable zero-shot across frozen LLM families.

3 METHOD

3.1 SETUP: A HUB OF INDEPENDENTLY TRAINED MEMBERS

A hub is a set of frozen decoder-only language models $\{f_1, \ldots, f_N\}$ that share one latent memory format: m position-independent slots of dimension d (m=32, d=1024 throughout). Member i adds two small adapters around its frozen backbone: the *Writer* W_i compresses a context x into an artifact $Z = W_i(x) \in \mathbb{R}^{m \times d}$, and the *Reader* R_j maps any artifact into a virtual K/V prefix under which f_j answers a query q with distribution $p_j(y \mid Z, q)$. Only the slot format and a frozen anchor (Section 3.2) are global, fixed before any member exists; training obeys an independence invariant: member i's computation graph (Figure 2a) contains its own frozen backbone, its own adapter pair, and the anchor—never another member's parameters, activations, gradients, or data order. Independence leaves one obstruction: the task loss of a self-trained pair (W_i, R_i) is invariant under every change of slot basis $Z \mapsto ZT$, $T \in \mathrm{GL}(d)$, because R_i absorbs T^{-1} exactly (Proposition 1), so the task loss leaves every member's basis free, and members trained without the anchor solve their task yet score at most 0.4 F1 on every cross pair (Section 4.4); the anchor fixes this basis once, publicly, for every present and future member (Proposition 2).

3.2 THE ANCHOR: A PUBLIC COORDINATE SYSTEM

The anchor is three frozen objects, published once, that together pin the slot space geometrically and functionally: a text embedder E (Qwen3-Embedding-0.6B, output dimension $e = d = 1024$), a random bridge $B \in \mathbb{R}^{e \times d}$ with orthonormal columns drawn from a public seed, and a judge—Qwen3-0.6B with a Reader trained once without any anchor, then frozen.

Geometric pinning. The first m_a slots are *anchored*: the embedder's tokenizer—frozen and shared, so every member chunks identically—cuts the written text into fixed-length pieces $c_1, c_2, \ldots$ (64 or 128 tokens; Appendix B), and anchored slot z_k regresses to a fixed target, so the same text receives the same absolute code in every member's anchored slots, whatever the backbone:

$$t_k \;=\; \sqrt{d}\,\frac{E(c_k)^\top B}{\|E(c_k)^\top B\|} \;\in\; \mathbb{R}^{1\times d}, \qquad \mathcal{L}_{\mathrm{geo}} \;=\; \frac{1}{m_a}\sum_{k=1}^{m_a}\frac{1}{d}\,\|z_k - t_k\|^2. \tag{1}$$

Functional pinning. Geometry fixes where information sits; the judge dictates the form it takes. The slots written by member i must also let the frozen judge answer the task:

$$\mathcal{L}_{\mathrm{judge}} \;=\; -\log p_{\mathrm{jud}}\big(y \mid Z, q\big), \tag{2}$$

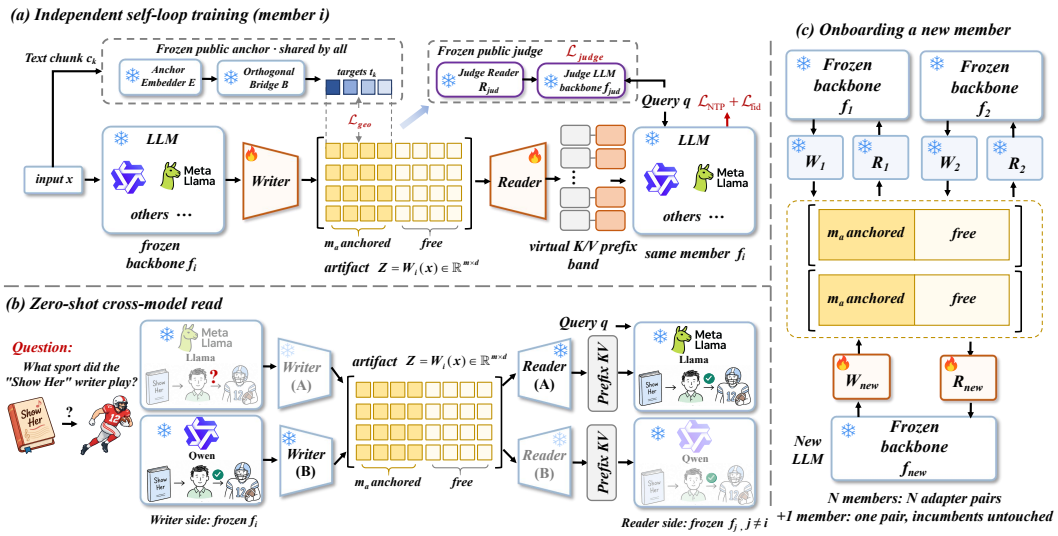


Figure 2: **MEMHUB: train alone, read anywhere, join anytime.** **(a)** Self-loop training of member i against the anchor (Eqs. 1–6); shading marks the m_a anchored slots. **(b)** Zero-shot read by a foreign Reader. **(c)** Onboarding: one adapter pair; incumbents and stored artifacts are untouched.

with gradients reaching W_i only through Z. Every Writer thus writes a code that one fixed Reader already consumes, and the $m - m_a$ free slots organize around the anchored ones rather than drifting into a private dialect: one basis change acts on every slot, so a basis pinned on any anchored slot is the basis of all m (Proposition 2). The default is m_a=16, although two anchored slots, or either constraint alone, already suffice (Section 4.4).

3.3 WRITER, READER, AND SELF-LOOP TRAINING

Writer. W_i reads hidden states—not the KV cache—from nine evenly spaced layers S_i in the mid-to-deep band $[L_i/3, L_i]$ of the frozen backbone, concatenates them per token into $H_i(x) = [h_i^{(\ell)}(x)]_{\ell \in S_i}$, and pools them into m slots by cross-attention against learned queries $Q_i \in \mathbb{R}^{m \times d}$:

$$Z = W_i(x) = g_i\big(\mathrm{CrossAttn}(Q_i, P_i H_i(x))\big) \in \mathbb{R}^{m \times d}, \tag{3}$$

where P_i is an input projection and g_i a self-attention block with output projection to d; hidden states carry no rotary phase, so the artifact is position-free by construction rather than by correction.

Reader. R_j is a shared trunk U_j—a linear input projection P_j^R of each slot, followed by normalized attention blocks—with per-layer read-out heads $\psi_j^{(\ell)}$ that emit keys and values directly in the receiver's native geometry (layer count L_j, KV heads, head dimension), so heterogeneous GQA configurations need no shape mapping; keys are emitted phase-free and rotated by the receiver's own rotary module at a virtual band of m positions from offset s, ahead of the context:

$$(\hat{K}_\ell, V_\ell) = \psi_j^{(\ell)}\big(U_j(Z)\big), \qquad K_\ell = \mathrm{RoPE}_j^{(s:s+m)}\, \hat{K}_\ell, \qquad \ell = 1, \ldots, L_j, \tag{4}$$

with entries RMS-normalized, rescaled by per-layer scales initialized to the receiver's cache statistics, and gated so memory arrives at native magnitude; the m entries load as a prefix cache,

$$p_j(y \mid Z, q) = f_j\big(y \,\big|\, q;\ \{(K_\ell, V_\ell)\}_{\ell=1}^{L_j} \text{ as prefix cache}\big), \tag{5}$$

so reading involves no decoding, no tokenization, and no assumption that Writer and Reader share a vocabulary, an architecture, or a scale.

Training. Each member trains alone, on its own tokenization of a shared public data mixture, by writing contexts and answering queries through its own Reader (Figure 2a), with a next-token loss on the answer given the injected prefix and a fidelity loss distilling the model's text-conditioned answer distribution into its memory-conditioned one (teacher-forced, over answer tokens):

$$\mathcal{L}_{\mathrm{NTP}} = \mathbb{E}\big[-\log p_i\big(y \mid W_i(x), q\big)\big], \qquad \mathcal{L}_{\mathrm{fid}} = \mathbb{E}\big[\mathrm{KL}\big(p_i(\cdot \mid x \oplus q) \,\big\|\, p_i(\cdot \mid W_i(x), q)\big)\big]. \tag{6}$$

The full objective adds the two anchor terms:

$$\mathcal{L} \;=\; \lambda_{\text{ntp}}\,\mathcal{L}_{\text{NTP}} + \lambda_{\text{fid}}\,\mathcal{L}_{\text{fid}} + \lambda_{\text{anchor}}\big(\mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{judge}}\big). \tag{7}$$

No term couples members: cross-model readability is never trained, only measured. Training is embarrassingly parallel; per-member configurations are given in Appendix B.

3.4 Deployment: write once, read anywhere, join anytime

Writing. A write is two forward passes—the frozen backbone and the Writer (Eq. 3)—with no per-document optimization; the output is the slot tensor serialized with its provenance (writing model, anchor version, slot layout), 64 KB in fp16 or 32 KB at int8 whatever the source length.

Reading. Any member loads an artifact, runs one Reader pass (Eq. 4), and answers with the resulting prefix cache (Eq. 5; Figure 2b); with no Writer, decoding, or tokens on the read path, a re-read costs a prefill (Section 4.5). The offset s is a read-time choice, so artifacts compose at disjoint bands,

$$\text{cache}\big(Z_1,\dots,Z_r\big) \;=\; \Big\{\big(\,\big[\text{RoPE}_j^{(s_t:s_t+m)}\hat{K}_\ell^{(t)}\big]_{t=1}^{r},\;\big[V_\ell^{(t)}\big]_{t=1}^{r}\,\big)\Big\}_{\ell=1}^{L_j}, \tag{8}$$

where $[\cdot]_{t=1}^{r}$ concatenates along the slot axis, so one Reader pass consults memories from different writers and times.

Joining. A new model joins by training one adapter pair against the published anchor—nothing else (Figure 2c): it reads every artifact ever written, its own are readable by every incumbent, and no incumbent changes, so membership is open at $O(N)$ total training cost (Section 4.3). The hub has no runtime component and no coordinator; the anchor is the entire standard.

4 Experiments

Nine members from five families (0.5B–8B), three multi-hop QA distributions, two families onboarded after the hub was frozen, and two hubs rebuilt from scratch give 828 directed cells, bounded by full-context text and three trained channels.

4.1 Evaluation protocol

Models. The hub comprises Qwen3-0.6B/1.7B/8B, Qwen2.5-0.5B-Instruct, Llama-3.2-1B/3B-Instruct, Llama-3.1-8B-Instruct, Phi-4-mini-instruct, and OLMo-2-1B-Instruct (Qwen Team et al., 2024; Yang et al., 2025; Grattafiori et al., 2024; Meta AI, 2024; Microsoft et al., 2025; OLMo Team et al., 2024; Allen Institute for AI, 2025) (Q06/Q17/Q8, Q25, L1/L3/L8, Phi, OLMo). Each frozen backbone receives one Writer–Reader pair (32×1024 slots) trained alone on 50,000 HotpotQA and 25,000 OpenHermes-2.5 examples (Yang et al., 2018; Teknium, 2024); members share only the anchor of Section 3.2. SmolLM2-1.7B (Ben Allal et al., 2025) and Gemma-2-2B-it (Gemma Team et al., 2024) are withheld from hub development and trained only for onboarding (Appendix B).

Protocol. Each HotpotQA bridge example is split into complementary five-paragraph Writer and Reader views with one supporting paragraph each. For Reader j, *Half* sees only its own view, *Full* all ten paragraphs as text, and *Oracle* the two supporting paragraphs; *Self* reads $W_j \to R_j$, *ZS* reads an independently trained foreign Writer $W_i \to R_j$, PairFT trains R_j specifically for frozen W_i, and *Empty* loads an artifact of the empty string. The locked matrix covers all $9 \times 8 = 72$ foreign directions (n=1000 per cell); the same adapters are re-evaluated without retraining on 2WikiMultihopQA (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022) (n=400). Three trained channels bound the comparison (Section 4.5): pair-trained C2C (Fu et al., 2026), a jointly trained shared space after Dery et al. (2026), and a distilled text communicator. All routes decode greedily.

Metrics. We report token F1, normalized exact match (EM), the gain over Half, and retention in percent (alignment rates as fractions). Retention is the share of a trained or native interface recovered zero-shot, $R_{\text{pair}}(i,j) = (F1_{\text{ZS}} - F1_{\text{Half}})/(F1_{\text{PairFT}} - F1_{\text{Half}})$ and $R_{\text{self}}(i,j)$ likewise with the Self denominator. To isolate answer-relevant content gain from task adaptation, we also

Table 1: **One adapter pair per model recovers pair-specific transfer, and the frozen interface carries unchanged to two unseen distributions.** F1 (%); rows are Readers, PAIRFT and ZS average the eight foreign Writers. Med. $R_{\text{pair}}/R_{\text{self}}$: median retention (Section 4.1; "–": denominator gain not significant); H^+: directions of 8 (72 in **All**) with a Holm-corrected EM gain over Half; †: median over seven valid directions.

(a) HotpotQA: zero-shot transfer against pair-specific training (n=1000 per direction)

	Baselines and ceilings				Memory reads			ZS vs. PAIRFT	
Reader	Half	Empty	Full	Oracle	Self	PAIRFT	ZS	Med. R_{pair} (%)	H^+
Qwen3-0.6B	25.9	31.8	34.8	54.4	39.8	39.5	38.9	94.8	8
Qwen3-1.7B	37.0	41.7	51.0	69.0	44.5	44.8	44.6	99.1	8
Qwen2.5-0.5B	29.6	23.1	34.8	57.4	30.5	32.4	30.6	15.5†	0
Qwen3-8B	45.8	55.3	67.3	77.1	58.6	58.0	58.0	101.4	8
Llama-3.2-1B	29.1	34.5	36.2	53.1	40.8	41.6	40.7	91.7	8
Llama-3.2-3B	39.2	48.6	59.1	73.3	51.4	54.0	50.8	79.2	8
Llama-3.1-8B	44.5	62.2	65.4	73.7	63.9	63.8	63.9	100.2	8
Phi-4-mini	40.7	45.9	52.2	67.2	50.0	50.3	49.5	93.3	8
OLMo-2-1B	23.2	40.8	28.5	45.1	44.2	44.4	44.1	97.8	8
All (72 directions)	**35.0**	**42.7**	**47.7**	**63.4**	**47.1**	**47.6**	**46.8**	**96.1**	**64**

(b) Unseen distributions: the same frozen adapters, no retraining (n=400 per direction)

	2WikiMultihopQA					MuSiQue				
Reader	Half	Self	ZS	Med. R_{self} (%)	H^+	Half	Self	ZS	Med. R_{self} (%)	H^+
Qwen3-0.6B	10.3	17.0	17.2	100.4	4	13.3	21.0	21.7	110.4	8
Qwen3-1.7B	16.3	20.5	20.1	87.8	0	19.2	24.7	24.9	101.2	4
Qwen2.5-0.5B	11.6	14.0	14.1	–	0	18.5	16.8	18.5	–	0
Qwen3-8B	20.9	25.9	26.7	114.6	8	25.4	38.7	38.6	97.7	8
Llama-3.2-1B	14.7	20.3	18.9	73.2	0	14.6	20.5	20.5	96.4	8
Llama-3.2-3B	11.2	24.9	25.1	99.8	8	18.8	33.4	31.3	87.0	8
Llama-3.1-8B	14.4	31.7	32.2	102.4	8	21.8	44.5	44.8	101.3	8
Phi-4-mini	17.4	27.2	25.9	87.4	8	22.6	33.8	33.1	91.7	8
OLMo-2-1B	10.2	20.4	20.7	101.9	8	13.1	24.9	24.0	91.9	8
All (72 directions)	**14.1**	**22.5**	**22.3**	**98.5**	**44**	**18.6**	**28.7**	**28.6**	**97.0**	**60**

define $R_{\text{content}}(i,j) = (F1_{\text{ZS}} - F1_{\text{Empty}})/(F1_{\text{PAIRFT}} - F1_{\text{Empty}})$. A ratio is reported only when its denominator gain has a positive paired-bootstrap interval ("–" otherwise). For R_{content}, the aggregate is the median over valid directions with a direction-cluster bootstrap 95% CI; paired intervals use 2,000 sample-paired resamples. EM significance uses exact McNemar tests with Holm correction within each matrix.

4.2 ONE ADAPTER PAIR PER MODEL RECOVERS PAIR-SPECIFIC TRANSFER

Across all 72 HotpotQA directions, MEMHUB recovers a median $\mathbf{96.1}\%$ of the gain of a Reader trained specifically for each foreign Writer ($R_{\text{pair}} = 96.1\%$, IQR $[86.5, 100.4]$, cluster-bootstrap 95% CI over training runs $[85.6, 99.7]$; Table 1a). This is Half-normalized readout-gain recovery. Relative to Empty, the foreign artifact adds a median $+3.2$ F1 content gain; R_{content} recovers 87.0% (IQR $[75.4, 99.6]$, 95% CI $[82.2, 93.2]$) of the PAIRFT gain in 70 of 72 directions with valid denominators. The gain over Half is positive in 66 of 72 paired F1 intervals and in 64 Holm-corrected EM tests, with none negative; the median gap to PAIRFT is -0.49 F1 and no cell trails by more than 4.6.

Recovery is uniform across families and scales. Eight of nine Reader columns improve for every foreign Writer under both tests; the ninth, Qwen2.5-0.5B, gains only 0.9 F1 even from its own Writer. The 20 directions that avoid the anchor's Qwen family entirely retain a median 96.9%. The Writer barely matters: zero-shot F1 spreads by at most 4.3 points across the eight Writers feeding any Reader, against gains of up to 21.5 (Figure 4). Relative to the Empty control, the adapter alone adds a median $+6.0$ F1 over Half and the foreign artifact a further $+3.2$ (Appendix D.6). The frozen adapters carry the same interface to 2WikiMultihopQA and MuSiQue (Table 1b): median self-retention is 98.5% and 97.0% against 97.6% on HotpotQA, the paired F1 gain is positive in

Table 2: **Anything trained against the anchor reads anything else trained against it.** HotpotQA, n=1000 per direction. **(a)** Held-out families: mean zero-shot F1 (%) and gain over Half across the nine incumbent edges, median R_{self}, edges significant under both tests, and H800 training hours (Tables 10, 11). **(b)** Hubs retrained under different seeds: median ZS−Half over the 72 directions in which the row hub's Readers read the column hub's Writers (positive / negative paired intervals; shaded diagonal: within-hub).

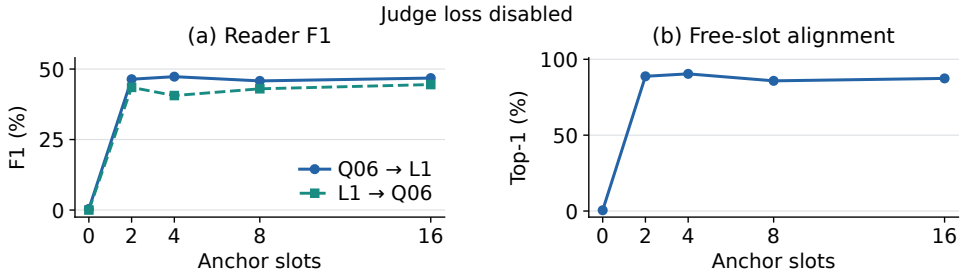
(a) Held-out families join the frozen hub

New family	Role	ZS	Δ	R_{self}	Sig.	Hours
SmolLM2-1.7B	Writer	46.8	+11.8	96.3	8/9	2.77
	Reader	42.6	+19.0	96.3	9/9	
Gemma-2-2B	Writer	46.7	+11.7	94.6	8/9	5.35
	Reader	53.2	+14.3	99.2	9/9	

(b) Independently trained hubs read one another (median ZS−Half over 72 directions; CI^+ / CI^-)

	Writer hub		
Reader hub	Seed 42	Seed 43	Seed 44
Seed 42 (Table 1)	+11.8 (66 / 0)	+11.9 (66 / 0)	+12.0 (66 / 0)
Seed 43	+13.4 (64 / 0)	+13.7 (64 / 0)	+13.7 (64 / 0)
Seed 44	+12.0 (64 / 0)	+11.7 (64 / 0)	+11.9 (65 / 0)

Figure 3: **Anchoring two of 32 slots fixes the shared coordinate system.** Zero-shot F1 (n=300) and top-1 alignment of the *unanchored* slots (150 probe texts) versus the number of anchored slots; Q06/L1 pair, judge loss disabled (Table 23).



70 and 64 of 72 directions, and none is negative. Nor is the code QA-specific: Readers trained for classification and summarization on their own Writers' artifacts read foreign artifacts zero-shot in all 72 directions, and one Reader per member serves four tasks (Appendix J).

4.3 Held-out families and independently trained hubs interoperate without retraining

After the hub is frozen, SmolLM2 and then Gemma-2—a GQA configuration absent from the hub—join by training one adapter pair each against the published anchor; no incumbent changes (Table 2a). Both interoperate at once in both directions: 34 of 36 new edges are significant under both tests, retention is 94–99%, and Gemma-2 as a Reader absorbs all nine foreign Writers within a 1.2-point band. Each join costs one adapter pair (2.77 and 5.35 H800-hours) where a pair-specific design needs 18 and 20 new integrations; the nine-member hub costs 29.87 hours (Figure 7).

Whole hubs read one another. We rebuild the hub from scratch under seeds 43 and 44 and let every Reader of every hub read every Writer of every other (Table 2b). Each new hub reproduces the locked hub internally (median gain over Half +13.7 and +11.9 F1 against +11.8; PairFT adds +0.5 and −0.3 over zero-shot reading), and across hubs the matrix is flat: over all 432 cross-hub directions a Reader extracts the same gain from a foreign hub's artifact as from its own to within a median 0.0 F1 (never more than 1.7; 402 of 432 intervals cross zero): the interface belongs to the anchor, not to any training run.

4.4 A small anchor fixes the shared coordinate system

Figure 3 sweeps the number of anchored slots with the judge loss disabled. Without an anchor the Q06/L1 development pair scores 0.4 and 0.0 F1 and free-slot top-1 alignment is .005: members

Table 3: **Each anchor constraint alone yields zero-shot transfer, through a different mechanism, and the coordinates are portable.** Q06/L1 pair, n=300 per direction; alignment is raw top-1 cross-Writer retrieval over 150 probe texts. **(b)** swaps the embedder for BGE-M3 or trains the members on disjoint corpus halves.

(a) Factorial over the two anchor constraints

			Zero-shot F1 (%)			Top-1 alignment	
Condition	$\mathcal{L}_{\text{geo}}$	$\mathcal{L}_{\text{judge}}$	Q06→L1	L1→Q06	Mean	Anchored	Free
Neither	–	–	0.0	0.1	0.1	.003	.004
Geometric only	✓	–	47.2	39.9	43.6	.987	.852
Judge only	–	✓	42.9	44.2	43.6	.020	.023
Both	✓	✓	48.0	43.6	45.8	.982	.528

(b) Anchor-substitution and data-overlap controls

	Q06→L1		L1→Q06			Top-1 alignment	
Control	ZS	Self	ZS	Self	Mean ZS	Anchored	Free
BGE-M3 anchor (Chen et al., 2024)	46.0	47.1	43.0	42.8	44.5	.998	.985
Shared corpus	45.9	46.3	42.1	42.9	44.0	.965	.525
Disjoint corpora	48.1	48.6	42.6	42.9	45.3	.977	.508

trained alone solve the task in private bases. Anchoring two of 32 slots raises transfer to 46.4/43.5 and free-slot alignment to .888, the plateau the default 16 anchors also reach (46.8/44.5): one basis change acts on all 32 slots (Proposition 2), so pinning it on two fixes the basis of all; the free slots' content then aligns as well (.888). Zeroing either slot group at read time collapses F1 below 2 (Table 24); the anchor organizes the full code rather than carrying it.

The factorial of Table 3a shows that the two anchor terms are individually sufficient and act through different mechanisms. Geometry alone restores full transfer by aligning the slot bases (free-slot retrieval .85); the judge alone restores it *without* aligning them (raw alignment .02), by pinning every Writer to one manifold a fixed Reader already consumes (linear label probes confirm the split, Appendix J.4). The production anchor composes both and posts the highest transfer (each constraint alone replicates on a second pair and seed, Table 25).

The coordinates are portable (Table 3b). Swapping the embedder for BGE-M3 preserves transfer and lifts free-slot retrieval to .985; an anchor with no Qwen component (BGE-M3 embedder, Llama-3.2-1B judge) holds all six cross-directions among three retrained members at production level (median $+1.4$ F1; Table 27); and members trained on disjoint corpus halves transfer as well as members trained on shared data (48.1/42.6 versus 45.9/42.1). Conversely, one Procrustes map fit to 50 unlabeled texts lifts the unanchored pair from zero to 44.8/39.3 F1, 950 further texts add nothing, and anchored pairs move by at most 1.4 (Figure 6). The basis freedom is the obstruction Proposition 1 proves, and one linear map is all that separates the unanchored members: independent training learns compatible task structure; the anchor selects the basis in which it is readable.

4.5 A 64 KB ARTIFACT STANDS IN FOR THE TEXT IT REPLACES

Against the full source text. Full hands each Reader all ten paragraphs as text (about 1,400 tokens); the foreign 64 KB artifact lands a median 2.7 F1 below it on HotpotQA, ties it on 2WikiMultihopQA ($+0.7$), and leads on MuSiQue ($+2.6$; 24 directions significantly positive, 1 negative; Table 4a): a foreign model's compressed memory generalizes better than the raw text it summarizes. Oracle beats Full by 8.4–24.0 points everywhere—full-context reading pays a distractor tax that compression prepays—and the weakest Readers answer better from their own artifact than from the complete source (OLMo, $+15.7$ F1; Table 1a).

Against a pair-trained channel. Under the same private-sender protocol, C2C at the matched budget (4800 steps, batch 32) clears Half by median $+9.4$ F1; the artifact leads by $+4.3$ (10/18 significant, none reversed, with or without a shared token grid), or $+3.9$ at C2C's published 1200-step budget (13/18; Table 20). C2C trains a projector per ordered pair; the artifact trains none.

Table 4: **The artifact matches or beats the text it replaces, beats pair-trained C2C, sits within 1.5 F1 of a jointly trained shared space, and reads at prefill cost.**

(a) Against the full source text (72 directions per dataset)

		ZS − Full (F1 points)				Mean F1 (%)		
Dataset	Dir.	Median	CI^+	CI^-	$Holm^+$	Half	Full	ZS
HotpotQA	72	−2.7	24	33	24	35.0	47.7	46.8
2Wiki	72	+0.7	9	21	8	14.1	23.1	22.3
MuSiQue	72	+2.6	24	1	46	18.6	25.6	28.6

(b) Against trained latent channels (n=1000 per pair)

		ZS − channel (F1 points)			Channel	Median F1 (%)	
Channel	Pairs	Median	[IQR]	$CI^+/^-$	−Half	Channel	ZS
C2C, shared tokenizer	12	+4.3	[+1.2, +6.6]	7 / 0	+7.9	46.1	47.6
C2C, disjoint tokenizers	6	+3.3	[+1.2, +6.3]	3 / 0	+9.8	38.6	44.4
C2C, all pairs	18	+4.3	[+1.0, +6.7]	10 / 0	+9.4	42.4	44.8
Joint shared space	12	−1.5	[−2.0, −1.3]	0 / 5	+14.0	44.5	42.8

(c) Against live and trained text messages (n=300 per cell)

		Median F1 (%)		Median	CI^+ cells		$Holm^+$ cells	
Text route	Cells	Direct	Text	Δ (pt)	Direct	Text	Direct	Text
Live T2T, all directions	72	48.4	42.9	+8.5	42	3	36	0
Live T2T, decode-matched	56	50.7	43.5	+9.2	34	0	29	0
Tuned T2T†	6	44.6	38.8	+7.1	5	0	5	0

(d) Wall-clock cost per operation (median ms; one H800, warm cache)

		Time to first token, batch 1				32-token answer, batch 32		
Reader	Write	Read	Re-read	Live	Speedup	Read	Re-read	Speedup
Qwen3-0.6B	24	35	25	1123	33×	31	70	2.3×
Llama-3.2-3B	23	34	30	962	28×	41	68	1.7×
Llama-3.1-8B	34	47	54	982	21×	63	125	2.0×

Notes. CI^+/CI^-: paired interval favors the first/second operand; $Holm^+$: corrected EM wins. (b) C2C matched budget: 4800 steps, batch 32. (c) $^\dagger n$=1000. (d) Re-read prefills text; Live decodes 50–59 tokens, then prefills.

Against a jointly trained shared space. A joint graph trains all adapters on cross-Reader losses (Dery et al., 2026) over Q06, L1, Phi, and OLMo. The jointly trained shared space leads MEMHUB by a median of 1.5 F1 points. The corresponding gains over Half are +14.0 F1 for the jointly trained shared space and +12.6 F1 for MEMHUB (Table 22). A fifth member joins by training through frozen incumbents; a MEMHUB member needs only the anchor.

Against text messages. On 56 decode-matched directions, direct reaches F1 50.7 versus 43.5 for live text and wins 45 cells. Distillation toward each Writer's best four messages leaves the artifact ahead by +7.1 F1 (Holm EM wins in 5/6, no reversal). Writers ≤1.7B lose to their artifacts in all 28 cells (12.0 points on average), so latent transfer helps where decoding is weakest (Appendix F).

Prefill cost, stability, and fidelity. Write/read takes 23–34/34–47 ms on 0.6B–8B Readers, 21–33× faster than live decoding and ahead of source re-reading at batch 32 (Table 4d). Quantization, position transport, and worst prompt perturbation each move F1 by less than 2 points; padding 512–8,192-token inputs leaves zero-shot F1 flat within 1.2 points while re-reading falls from 1,024 tokens onward (Tables 16, 15). Artifacts compose, degrade under distractors, and steer answers without overriding visible evidence (Appendix E).

5 CONCLUSION

Text is the universal, persistent medium of multi-model systems; latent channels trade both away for bandwidth. MEMHUB uses a pre-training anchor to fix the coordinate system that makes independently trained memories readable. Across nine members and five families it recovers 96.1% of pair-specific readout gain zero-shot, retains 97% under shift, admits new families with one adapter pair, and reads rebuilt hubs at prefill cost. Shared state requires a publishable coordinate system.

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.

REPRODUCIBILITY STATEMENT

Section 4.1 defines the evaluation protocol, metrics, and statistical tests; Appendix B specifies the anchor, per-member training configurations, and evaluation construction. Per-sample model outputs and per-cell statistics for all 828 directed cells—every score, bootstrap interval, and McNemar test behind every aggregate—are released as machine-readable files with the supplementary material, and the two replication hubs of Section 4.3 reproduce the hub itself from the recipe alone; Appendices D–F analyze their cell-level structure, and Appendix H reports the development studies behind the architecture and training choices, and Appendix J the protocol and per-cell files of the non-QA study.

REFERENCES

Allen Institute for AI. Olmo 2 1B: May 2025 release notes, May 2025. URL https://docs.allenai.org/release_notes/olmo-release-notes.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 2021.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, et al. SmolLM2: When smol goes big—data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. doi: 10.48550/arXiv.2502.02737. URL https://arxiv.org/abs/2502.02737.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024. doi: 10.48550/arXiv.2402.03216. URL https://arxiv.org/abs/2402.03216.

Lucio M. Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent space communication via K-V cache alignment. *arXiv preprint arXiv:2601.06123*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, and Christopher Ré. Cartridges: Lightweight and general-purpose long context representations via self-study. *arXiv preprint arXiv:2506.06266*, 2025.

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 *International Conference on Learning Representations*, 2026. arXiv:2510.03215.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *International Conference on Learning Representations*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. 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. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2):167–195, 2015.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Yingzhuo Liu. Beyond tokens: A unified framework for latent communication in LLM-based multi-agent systems. *arXiv preprint arXiv:2606.05711*, 2026.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. doi: 10.48550/arXiv.2503.01743. URL https://arxiv.org/abs/2503.01743.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems*, 2023.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, 2018.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. doi: 10.48550/arXiv.2501.00656. URL https://arxiv.org/abs/2501.00656.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. In *International Conference on Learning Representations*, 2024. arXiv:2310.06272.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Vignav Ramesh and Kenneth Li. Communicating activations between language model agents. In *International Conference on Machine Learning*, 2025. arXiv:2501.14082.

Maximillian Rossi, Prajwal Raghunath, and Eugene Wu. Latent cache flow: Model-to-model communication without text. *arXiv preprint arXiv:2605.22863*, 2026.

Teknium. OpenHermes-2.5. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

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/.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed H. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework. In *First Conference on Language Modeling*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

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. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. MemAgent: Reshaping long-context LLM with multi-conv RL-based memory agent. In *International Conference on Learning Representations*, 2026. arXiv:2507.02259.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning*, 2026. arXiv:2511.20639.

A WHY INDEPENDENT TRAINING NEEDS AN ANCHOR: THE COORDINATE ARGUMENT

Section 3.1 attributes the failure of unanchored members to coordinates rather than capability. Two statements are proved here for the architecture of Section 3.3: a member's own task loss does not determine its slot basis (Proposition 1), and the geometric anchor determines it on every slot whenever the anchored targets span $\mathbb{R}^d$ and are fit closely (Proposition 2). Appendix G reports the measurements that this argument organizes.

A.1 SELF-LOOP TRAINING IS INVARIANT UNDER A CHANGE OF SLOT BASIS

Write the Writer of Eq. 3 as $W_i(x) = G_i(x)\,A_i$, where $A_i \in \mathbb{R}^{d_i' \times d}$ is the output projection of the block g_i, applied to every one of the m slot rows, and $G_i(x) \in \mathbb{R}^{m \times d_i'}$ collects everything before it. Write the Reader trunk of Eq. 4 as $U_i(Z) = \tilde{U}_i\big(Z\,P_i^R + \mathbf{1}b_i^\top\big)$, where $P_i^R \in \mathbb{R}^{d \times h_i}$ is the linear input projection of each slot, b_i its bias, and $\tilde{U}_i$ collects every normalization, attention block, and read-out head that follows, together with the RMS rescaling, gating, and rotary module of Eq. 4.

Proposition 1 (Basis invariance of self-loop training)**.** *Fix member i and set $\lambda_{\text{anchor}} = 0$ in Eq. 7. For every $T \in \mathrm{GL}(d)$ let (W_i', R_i') be the adapter pair with $A_i' = A_i T$ and $P_i'^R = T^{-1}P_i^R$ and every other parameter unchanged. Then*

$$W_i'(x) = W_i(x)\,T, \qquad R_i'(Z) = R_i\big(Z\,T^{-1}\big),$$

and (W_i', R_i') attains exactly the same $\mathcal{L}_{\text{NTP}}$ and $\mathcal{L}_{\text{fid}}$ as (W_i, R_i) on every input.

Proof. $W_i'(x) = G_i(x)A_iT = W_i(x)T$. For any Z, $R_i'(Z) = \tilde{U}_i(ZT^{-1}P_i^R + \mathbf{1}b_i^\top) = R_i(ZT^{-1})$, so $R_i'(W_i'(x)) = R_i(W_i(x)T\,T^{-1}) = R_i(W_i(x))$: T^{-1} is absorbed by the input projection before any normalization or nonlinearity, so every downstream module—normalization, attention, read-out heads, RMS rescaling, gates, and the rotary band—receives an input identical entry by entry. The prefix cache handed to f_i is therefore identical, and with it every conditional $p_i(y \mid W_i'(x), q)$ of Eq. 5. Both losses are functions of these conditionals alone. □

Remark 1 (The symmetry acts on the basis, not on the slots)**.** *The invariance acts on the slot axis $\mathbb{R}^d$ with one matrix T shared by all m slots, because A_i and P_i^R are shared by all slots. It does not extend to permutations of the slots: the receiver rotates slot k at virtual position $s{+}k$ (Eq. 4), so permuting slots changes the prefix cache, and no parameter of R_i undoes it. Slot identity is fixed by the interface; only the basis is free. This is why one $d \times d$ map with no permutation search aligns unanchored members (Table 28), and why "anchor the first m_a slots" is a well-defined operation.*

Corollary 1 (The unanchored objective does not determine the basis)**.** *For every $T \in \mathrm{GL}(d)$ the map $(W_i, R_i) \mapsto \big(W_iT,\ R_i(\cdot\,T^{-1})\big)$ is a bijection of the set of minimizers of the unanchored objective onto itself. No property of the artifact basis that varies under $\mathrm{GL}(d)$ is determined by member i's data, objective, or backbone; which point of the orbit a run reaches is left to initialization and optimizer dynamics, which are independent across members, so the relative basis of two independently trained members is uncontrolled.*

Members trained without the anchor solve their own task and score .004 and .000 F1 on each other's artifacts (Table 23, m_a=0).

A.2 THE ANCHOR DETERMINES THE BASIS

Let $\mathcal{T} \in \mathbb{R}^{N \times d}$ stack the anchored targets $t_k(x)$ of Eq. 1 over every training text x and every anchored slot $k \le m_a$, and let $Z_a(W_i) \in \mathbb{R}^{N \times d}$ stack the corresponding anchored slots written by W_i.

Proposition 2 (Geometric pinning has a trivial stabilizer)**.** *Let $V = \mathrm{span}$ of the rows of $\mathcal{T}$. If (W_i, R_i) and $(W_iT,\ R_i(\cdot\,T^{-1}))$ both attain $\mathcal{L}_{\text{geo}} = 0$, then $tT = t$ for every row t of $\mathcal{T}$, so T is the identity on V; when $V = \mathbb{R}^d$, $T = I$. If instead both attain $\|Z_a - \mathcal{T}\|_F \le \delta$, then $\|T - I\|_F \le (1 + \|T\|_{\text{op}})\,\delta/\sigma_{\min}(\mathcal{T})$.*

Proof. The anchored slots of $W_i T$ are $Z_a(W_i)\,T$. Exact fit for both pairs gives $\mathcal{T}T = Z_a(W_i)T = \mathcal{T}$, and rows spanning $\mathbb{R}^d$ force $T = I$. For the bound, $\mathcal{T}(T - I) = (\mathcal{T} - Z_a)T + (Z_a T - \mathcal{T})$, so $\sigma_{\min}(\mathcal{T})\,\|T - I\|_F \le \|\mathcal{T}(T - I)\|_F \le \delta\|T\|_{\mathrm{op}} + \delta$. □

Two consequences fix what the anchor is asked to do.

- *Identifiability depends on the span of the targets, not on* m_a. The targets are chunk embeddings of tens of thousands of distinct training texts, so the rows of $\mathcal{T}$ span $\mathbb{R}^{1024}$ for every $m_a \ge 1$; m_a=2 is the smallest value tested.
- *The constrained object is one* $d \times d$ *matrix, not the* 32×1024 *slot space.* T has $d^2 \approx 10^6$ entries and is met by $N \sim 10^5$ target rows of d equations each; and because the same T acts on every slot (Remark 1), a basis pinned on any anchored slot is the basis of all m. This is the exact content of "two slots align the other thirty".

Free-slot alignment therefore has two factors. The basis is shared: this is Proposition 2. The content that members place in their free slots is the same: this is measured. Zeroing the free slots at read time collapses F1 below .014 at every setting (Table 24), so those slots carry task state; cross-Writer top-1 retrieval on them is .888 at m_a=2 and .874 at m_a=16 (Table 23), so that state is the same across members. The two factors move independently: anchored-slot alignment stays at .96–1.00 in every anchored condition, while free-slot alignment ranges from .51 (production anchor) to .99 (BGE-M3 anchor; Table 3).

Remark 2 (Functional pinning identifies a convention, not a basis)**.** R_{jud} *is a fixed nonlinear map, so* $\mathcal{L}_{\mathrm{judge}}$ *is not invariant under Proposition 1: only a* T *that leaves every output of* R_{jud} *unchanged survives it. What the judge fixes is a decoding convention—the point on each member's solution set that one frozen Reader decodes—rather than a coordinate system: in the judge-only condition raw cross-Writer retrieval is* .02 *and centered retrieval* .16 *while transfer remains significant in every direction (Table 26). Coordinate agreement is sufficient for readability (geometric only) and not necessary for it (judge only); the production anchor imposes both and posts the highest transfer (Table 3a).*

Conversely, one centered orthogonal Procrustes map fit to 50 paired unlabeled texts lifts the unanchored Q06/L1 pair from .000/.001 to .448/.393 F1 and 950 further texts add nothing, while anchored pairs start at plateau and move by at most .014 (Table 28, Figure 6): the anchor is this one map, published before any member exists.

B Protocol and implementation details

This section completes Section 4.1: the member configurations (Table 5), the training recipe (Table 6), the evaluation and statistical protocols, and the four reference channels (Table 7).

B.1 Members and geometric targets

Table 5 lists the frozen backbones and the member-specific settings fixed before the locked matrix was evaluated. All members instantiate the identical anchor objective and the identical 32×1024 slot interface; the geometric targets t_k of Eq. 1 are precomputed once from the public mixture with the frozen embedder and bridge and shared by every member.

Two chunk lengths are in use for the geometric targets: 128 tokens (at most 8 per example) for the seven members trained first, and 64 tokens (at most 16 per example) for the two 8B members, Gemma-2, and the eighteen members of the two replication hubs. The interface is indifferent to the choice: the 28 cross-chunking edges of the locked matrix and the 112 (of 144) between the locked hub and each replication hub read at within-hub level (Table 2b), and the development pair replicates under both (Table 31). Gemma-2-2B trains with microbatch 4×32 because its 256k-entry vocabulary dominates judge-loss memory; its read path is unchanged.

B.2 Training recipe

Table 6 gives every setting of the self-loop training of Section 3.3 that is common to all members; the per-member settings are those of Table 5.

Table 5: **Frozen backbones and adapter-training configurations of the locked nine-member hub.** Batch denotes microbatch $\times$ gradient accumulation; Chunk is the token length of the geometric-target chunks.

Backbone	Size	λ_{fid}	Steps	Batch	Chunk
Qwen2.5-Instruct	0.5B	0.5	1200	8×16	128
Qwen3	0.6B	0	1200	8×16	128
Qwen3	1.7B	0	1200	6×21	128
Qwen3	8B	0	600	2×64	64
Llama-3.2-Instruct	1B	0	1200	8×16	128
Llama-3.2-Instruct	3B	0.5	1200	4×32	128
Llama-3.1-Instruct	8B	0	600	2×64	64
Phi-4-mini-Instruct	3.8B	0	1200	4×32	128
OLMo-2-Instruct	1B	0	1200	8×16	128

Table 6: **Common training recipe of every member and of both held-out families.**

Writer input	hidden states of nine evenly spaced layers in $[L_i/3, L_i]$, pooled into 32 slots by cross-attention (Eq. 3)
Reader route	R-append: a shared trunk with per-layer read-out heads emitting a 32-entry virtual K/V prefix in the receiver's native geometry (Eq. 4; selected in Table 32)
Precision	adapters in float32 around a bfloat16 frozen backbone
Anchor	m_a=16 anchored slots, λ_{anchor}=1; embedder Qwen3-Embedding-0.6B (Zhang et al., 2025); judge: a no-anchor Qwen3-0.6B Reader trained once in advance and frozen
Data	50,000 HotpotQA + 25,000 OpenHermes-2.5 examples, each member on its own tokenization; 60% of HotpotQA examples in the memory-only construction
Optimizer	AdamW, learning rate 10^{-4}, weight decay .01, cosine decay, gradient clipping at 1.0, seed 42
Token limits	Writer and student inputs 1,024 tokens; fidelity teachers up to 2,048
λ_{fid}	chosen per member from $\{0, .5\}$ by the conformance rule of Appendix H (Table 35)
Checkpoint	the step with the highest generation F1 on 200 development examples

B.3 Evaluation construction and statistics

- **Locked HotpotQA slice.** Examples 500–1499 of the validation split (n=1000) in each of the 72 directed cells. Each bridge example is split into complementary five-paragraph Writer and Reader views with one supporting paragraph each; Writer inputs are truncated to 1,024 tokens.
- **Shifted distributions.** 2WikiMultihopQA keeps compositional and inference items with exactly two supporting titles; MuSiQue keeps answerable two-hop items; n=400 per direction, same adapters, no retraining.
- **Decoding.** Every route decodes greedily with repetition penalty 1.0 (Qwen2.5 T2T endpoints: 1.1, Appendix F.1) to at most 32 answer tokens against uniform plain-text references. The source-length study (Table 15) raises the Writer limit to 512–8,192 tokens for both routes; the sample-matched T2T comparison allows up to 200 background tokens before the Reader answers; artifact-swap controls use one reference Writer per Reader, paired by example ID.
- **Statistics.** Gain intervals use 2,000 sample-paired bootstrap resamples; EM tests are exact two-sided McNemar tests with family-wise Holm correction within each matrix; aggregates are medians with IQR. A retention ratio is reported only when its denominator gain has a positive paired interval. Main-text F1 and retention are in percent; the development and mechanism tables of the appendix report F1, retention, and alignment as fractions.

B.4 Reference channels

Four trained channels bound the comparison. PAIRFT is the per-direction ceiling of Table 1; C2C, the jointly trained shared space, and the tuned text communicator are the three alternative designs of Table 4. Table 7 records how each is trained; their results are in Appendix F.

Table 7: **Training protocol of the four reference channels.** All headline comparisons use matched effective budgets, data splits, initialization, frozen backbones, and paired evaluation; C2C step sweeps and doubled PAIRFT are sensitivity checks, with development-selected checkpoints for each baseline.

Channel	What is trained	Details and selection
PAIRFT (one per directed cell)	the target Reader R_j only, using the common initialization protocol, against the frozen foreign Writer W_i	Matched headline budget and common protocol; paired development checkpoint. The doubled-budget run in Table 30 is sensitivity only
C2C (Fu et al., 2026) (one per ordered pair)	the official stack: per-position KV fusion, one projector per receiver layer, Gumbel-annealed gates; TokenAligner on the six cross-tokenizer pairs; 269–680M parameters per pair	Matched headline budget and common protocol; 600-, 1200-, and 4800-step runs are convergence checks, with the strongest development-selected setting used for the headline result
Joint shared space (after Dery et al., 2026)	the four Writer–Reader pairs of Q06, L1, Phi, OLMo in one computation graph: each step samples four ordered paths (i, j), self-loops included, and minimizes Reader j's answer loss on Writer i's slots; no anchor, no fidelity term	Matched headline budget and common protocol; checkpoint selected on the development mean over 12 cross paths
Tuned text communicator (one per Writer)	a LoRA on the Writer, rank matched to the Writer adapter's parameter count, distilled toward its best-of-four sampled messages under a frozen Qwen3-0.6B scorer that grades the receiver's downstream answer	Matched headline budget and common protocol; messages up to 200 tokens; live T2T uses the frozen Writer's native 50–59-token message with no training

After budget matching, C2C remains the most expensive object in the study: the eighteen 4800-step projectors cost about 300 H800-hours, ten times the training of the entire nine-member hub (Appendix I). Every headline C2C number uses the dev-selected 4800-step projector; the 600- and 1200-step projectors and the unselected milestones of the 4800-step run are reported alongside it in Tables 20–21.

C SIX ZERO-SHOT READS, ONE QUESTION AT A TIME

The matrices of Section 4.2 summarize 72,000 zero-shot answers. This section shows six of them in full. Each case fixes one frozen Reader and one HotpotQA question and reports the Reader's answer under three conditions: the question alone; Half, the Reader's own five-paragraph view with the question; and MEMHUB ZS, that same view plus the 64 KB artifact an independently trained Writer of another family wrote from its own view. The Reader never sees the Writer's text; the last row of each case shows the supporting paragraph that text contained ("Writer evidence"). All outputs are greedy generations from the locked slice; question texts are reproduced verbatim from HotpotQA. The six cases span four Reader families (Llama, Qwen3 at 0.6B and 8B, OLMo, Phi) and three Writer families (Qwen3, Llama, Phi), and no Writer–Reader pair among them shares a tokenizer.

Case 1 Writer Qwen3-0.6B → Reader Llama-3.2-1B		
Question	Which English actor, comedian and singer is featured in *Chalk and Cheese*?	
Gold	Michael Patrick Smith	
Question only	John Cleese.	×
Half	Richard Cheese.	×
MEMHUB ZS	**Michael Patrick Smith**	✓
Writer evidence	*Chalk and Cheese was a short-lived sitcom television show, featuring Michael Crawford and Robin Hawdon.*	

The Reader's own view holds the biography that gives Michael Crawford's birth name; what it lacks is the sitcom's cast. The artifact carries that single fact—written by a 0.6B Qwen model, read by

a 1B Llama model with a different tokenizer—and the Reader completes the chain to the name the question asks for. Without the artifact the Reader offers a famous comedian, then a pun on the title.

Case 2 Writer Llama-3.2-1B → Reader Qwen3-0.6B		
Question	What is the are of the university at which Dick Biddle served as head football coach?	
Gold	575 acres (2.08 km^2)	
Question only	The university at which Dick Biddle served as head football coach is the University of Texas at Austin.	×
Half	The area of the university where Dick Biddle served as head football coach is not provided in the information given.	×
MEMHUB ZS	**575 acres (2.08 km^2)**	✓
Writer evidence	*Dick Biddle (born November 26, 1947) is a former American football player and coach. He served as head football coach at Colgate University from 1996 through 2013, compiling a record of 137–73.*	

Under Half the Reader states, correctly, that the information is not in front of it: the campus figure sits in its own Colgate paragraph, but nothing links Biddle to Colgate. The artifact supplies the link and nothing else; the answer, down to the unit conversion, is read off the Reader's own text. This is the division of labor the Empty control quantifies in Appendix D.6—the adapter teaches the Reader to answer tersely, the foreign artifact delivers the missing hop.

Case 3 Writer Llama-3.2-1B → Reader Qwen3-0.6B		
Question	Flynn Rider is voiced by which American actor and singer born September 29, 1980?	
Gold	Zachary Levi	
Question only	American actor and singer born September 29, 1980: Morgan Freeman.	×
Half	Frank Welker	×
MEMHUB ZS	**Zachary Levi**	✓
Writer evidence	*. . . Studios' 50th animated feature film "Tangled", its short film "Tangled Ever After", and the 2017 television series. The character is voiced by American actor Zachary Levi, who decided to audition for the role upon learning that he would also be providing the character's singing voice.*	

Two fluent wrong actors without the artifact, one of them a voice actor guessed from the domain of the question; with the artifact the Reader names the one actor whose biography in its own view carries the stated birth date.

Case 4 Writer Llama-3.2-1B → Reader OLMo-2-1B		
Question	The singer known as the "Ballad Prince" was on a romantic comedy that aired on what state-run channel from August 11 to September 30, 2010?	
Gold	SBS	
Question only	"Ballad Prince" was on "Kuume!" from August 11 to September 30, 2010.	×
Half	My Girlfriend Is a Nine-Tailed Fox	×
MEMHUB ZS	**SBS**	✓
Writer evidence	*Lee Seung-gi (born January 13, 1987) is a South Korean singer, actor, host and entertainer. Known as the "Ballad Prince", Lee has had numerous hit songs such as "Because You're My Woman", "Will You Marry Me", and "Return".*	

The Half answer is the drama's title: the Reader has the broadcast paragraph and stops one hop short, because nothing in its view resolves "Ballad Prince" to a person. The artifact resolves the epithet to Lee Seung-gi, and the Reader reads off the channel. OLMo-2-1B is the Reader with the largest zero-shot gain in the hub ($+20.9$ F1 mean, Table 1a; $+21.5$ here); this is what that gain consists of.

Case 5 Writer Phi-4-mini → Reader Qwen3-8B		
Question	Berk Communication and Marketing Group best known for representing an American professional baseball pitcher for what team?	
Gold	New York Yankees	
Question only	Los Angeles Dodgers	×
Half	Chicago Cubs	×
MEMHUB ZS	**New York Yankees**	✓
Writer evidence	*Carsten Charles "CC" Sabathia, Jr. (born July 21, 1980) is an American professional baseball pitcher for the New York Yankees.*	

An 8B Reader still needs the bridge. Its own paragraph names the agency's client; the Writer's names the client's team; neither half answers alone, and the 8B model's two unaided guesses are simply plausible teams. The artifact was written by a 3.8B model of a third family, which per Table 9 makes no measurable difference to the Reader.

Case 6 Writer Qwen3-0.6B → Reader Phi-4-mini		
Question	The Panel with striding lion was one of many that lined the way north of the gate to the inner city of what location?	
Gold	Babylon	
Question only	Thebes, Egypt	×
Half	Ishtar Gate	×
MEMHUB ZS	**Babylon**	✓
Writer evidence	*The Panel with striding lion (MA 31.13.1) is a panel of Neo-Babylonian glazed ceramic bricks or tiles dated to 604–562 B.C., now in the Metropolitan Museum of Art, New York. It was one of many that lined the Processional Way...*	

Half answers "Ishtar Gate"—the right paragraph and the wrong hop. The artifact places the panel on the Processional Way, and the Reader, whose view locates that way in Babylon, names the city.

What the six cases share. Three regularities recur, and each corresponds to a number elsewhere in the paper.

- *The question-only answers are confident and wrong.* The bridge entity is not in any Reader's parametric memory; it has to arrive from the other view.
- *The Half answers stop one hop short.* In two cases the Reader names the entity of its own supporting paragraph (the drama, the gate); in one it says outright that the information is missing; in the rest it guesses from the domain of the question. The Reader-only baseline is not failing at reading—it is missing exactly one fact.
- *The ZS answers are the gold string.* The artifact delivers the missing fact in the form the Reader can act on, and the Reader answers tersely. Across the matrix the two effects are the $+3.2$ F1 of content and the $+6.0$ F1 of adaptation that Appendix D.6 separates with the Empty control.

The Writers here range from 0.6B to 3.8B and the Readers from 0.6B to 8B; no pair shares a family or a tokenizer; and none of the five adapter pairs ever saw another member during training. The cases are the cell-level face of the 96.1% retention of Table 1a.

D CELL-LEVEL STRUCTURE OF THE LOCKED MATRICES

Three regularities organize all 828 directed cells. Retention concentrates tightly around full recovery. The interface is *writer-invariant*: once an artifact exists, the identity, family, and scale of the model that wrote it are nearly irrelevant to what any Reader extracts from it. And the interface is *run-invariant*: two hubs retrained from scratch reproduce the locked hub column for column and read its artifacts, and each other's, at within-hub level. Figure 4 shows the first two at a glance; the tables of this section give each a cell-level view.

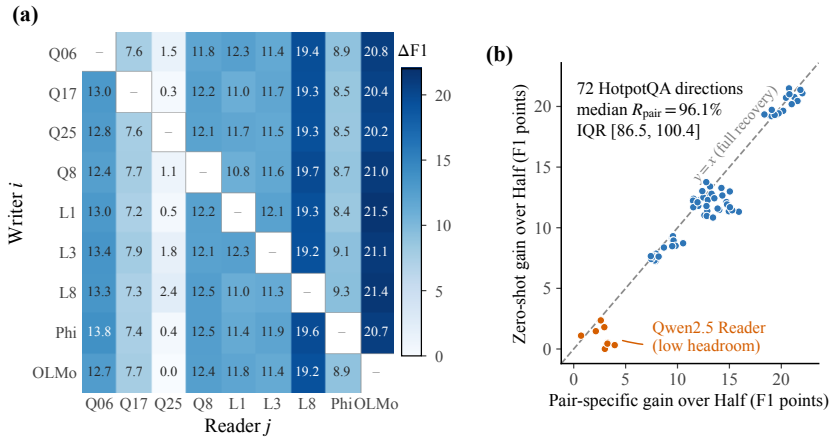


Figure 4: **Cell-level view of the locked HotpotQA matrix** ($n = 1000$ per direction). **(a)** Zero-shot gain of Writer $i \to$ Reader j over Half: every off-diagonal cell is positive, and the column structure shows that variation is governed by Reader headroom, not by Writer identity or family. **(b)** The same 72 gains against the gain of a Reader trained specifically for each foreign Writer (PAIRFT); points lie on the full-recovery diagonal across the performance range.

D.1 PER-CELL RETENTION

Table 8 reports R_{pair} for every HotpotQA direction. Outside the Qwen2.5-0.5B column, all 64 cross-family and within-family cells lie in $[.71, 1.08]$ and pass both significance tests; the median is .965 and the IQR $[.891, 1.012]$. The Llama-3.2-3B column is measured against the strongest denominator in the hub—pair-specific training there exceeds the member's own self-loop by 2.6 F1 points (.540 versus .514)—and its zero-shot cells are positive under both tests in all eight directions. The Qwen2.5-0.5B Reader is the one column with no headroom to measure: its own Writer adds .009 F1 (Table 1a), so the tests rarely bind there (two positive F1 intervals, no Holm-significant EM gain).

Table 8: **Per-cell PAIRFT retention on the locked HotpotQA matrix** ($n = 1000$ per direction). Each entry is R_{pair} for Writer (row) $\to$ Reader (column). Sixty-six of 72 directions have a positive paired-bootstrap F1 interval and 64 a Holm-corrected EM gain; the remaining cells are all in the Qwen2.5 column (†: EM gain not Holm-significant; ‡: F1 interval also not positive; $–^{‡}$: denominator gain not significant, ratio undefined).

	Reader j								
Writer i	Q06	Q17	Q25	Q8	L1	L3	L8	Phi	OLMo
Q06	–	.99	.69‡	.92	.95	.75	.98	.93	.98
Q17	1.05	–	.08‡	1.06	.86	.78	1.05	.87	.95
Q25	.96	.93	–	1.01	1.01	.82	1.00	.89	.96
Q8	.92	1.03	–‡	–	.81	.84	1.03	.83	1.01
L1	.86	.92	.16‡	1.02	–	.82	.99	.96	1.04
L3	1.02	.99	.61†	1.01	1.06	–	1.01	.97	.96
L8	.93	.94	.91†	.96	.88	.71	–	.97	.98
Phi	1.08	.99	.13‡	1.00	.88	.81	.97	–	1.01
OLMo	.88	.99	.01‡	1.07	.98	.76	1.01	.94	–

D.2 WRITER INVARIANCE

Table 9 quantifies the column structure of Figure 4a and extends it to both shifted datasets. Across all 27 Reader×dataset columns, the largest spread any Writer induces is 4.3 F1 points (MuSiQue, Llama-3.2-3B) against zero-shot gains of up to 21.5 points, and the median spread is 1.5 points. An artifact written by a 0.5B model is, to within about a point of F1, as useful to every Reader as one

written by an 8B model. The anchored format equalizes Writers, and that equality is what makes the handoff economics of Appendix F possible: the cheapest Writer in the hub writes an artifact as good as the most expensive one.

Table 9: **The interface is writer-invariant.** For each Reader and dataset, the spread ($\max - \min$, F1 points) of zero-shot F1 across the eight foreign Writers, next to the Reader's mean zero-shot gain over Half on HotpotQA. Largest spread anywhere: 4.3; median: 1.5; gains: up to 21.5.

	Mean gain	ZS spread across Writers (F1 pts)		
Reader	(HotpotQA)	HotpotQA	2Wiki	MuSiQue
Qwen3-0.6B	+13.0	1.3	1.0	1.9
Qwen3-1.7B	+7.6	0.6	1.2	2.6
Qwen2.5-0.5B	+1.0	2.3	2.1	1.5
Qwen3-8B	+12.2	0.7	0.9	1.6
Llama-3.2-1B	+11.5	1.4	2.5	1.5
Llama-3.2-3B	+11.6	0.8	2.1	4.3
Llama-3.1-8B	+19.4	0.5	1.9	1.7
Phi-4-mini	+8.8	0.9	1.9	3.2
OLMo-2-1B	+20.9	1.3	1.2	1.6

D.3 Distribution shift at cell level

The frozen adapters carry the interface unchanged to 2WikiMultihopQA and MuSiQue (Table 1b): median self-retention is .985 and .970, against .976 on HotpotQA. At n=400 the paired F1 interval is the sharper instrument. On 2WikiMultihopQA it is positive in 70 of 72 directions—every Reader except Qwen2.5 at 8/8—while Holm-corrected EM significance (44/72) concentrates in the five highest-gain columns (Qwen3-8B, Llama-3.2-3B, Llama-3.1-8B, Phi, OLMo, each 8/8). On MuSiQue the two tests agree almost everywhere (64 and 60 of 72). No direction on either dataset is significantly negative under either test, and the Writer spreads of Table 9 stay within 4.3 points on both.

D.4 Held-out families

Tables 10 and 11 list all 36 edges of the two families admitted after the hub was frozen. For SmolLM2-1.7B, each of the seventeen significant edges combines a positive paired-bootstrap F1 interval with a Holm-corrected McNemar EM gain at $p_H \leq 4 \times 10^{-9}$. SmolLM2's own baselines are Half .236, Full .268, and Self .435 F1; its zero-shot Reader gain of $+.190$, earned against nine Writers it never trained with, is among the largest in the hub, and as a Writer it retains a median .963 of the incumbents' self-loop gain.

Gemma-2-2B brings a GQA configuration absent from the hub. Median self-loop retention over its 17 significant edges is .974 (IQR $[.94, 1.00]$), and as a Reader it exhibits the writer invariance of Table 9 from its first day: all nine foreign Writers land within 1.2 F1 points of one another. Its own baselines are Half .389 and Self .534 F1.

D.5 Replication hubs

Table 12 gives the Reader-level conditions of the two hubs retrained from scratch under seeds 43 and 44 with the recipe of Table 5 and 64-token target chunks throughout. Both reproduce the locked hub column for column: zero-shot F1 from eight foreign Writers matches each Reader's own self-loop to within 1.2 points in every one of the 18 columns, and the largest gains again land on OLMo and Llama-3.1-8B. The within-hub PairFT references (16 directions per hub, 600 steps) add $+0.5$ and -0.3 F1 over zero-shot reading.

The 432 cross-hub directions behind the off-diagonal cells of Table 2b are released per direction with the supplementary material. In every one of the six hub pairings the cross-hub gain over Half is positive in 64–66 of 72 directions and negative in none. The paired difference between reading the foreign hub's Writer and one's own hub's same-named Writer has median 0.0 F1 (magnitude median 0.3, never above 1.7), with intervals crossing zero in 402 of 432 directions and splitting 18

Table 10: **All 18 SmolLM2-1.7B onboarding edges** (HotpotQA, $n = 1000$ per direction). Each row is one incumbent; ZS and Δ are F1 percentages and R_{self} is retention relative to the native self-loop. Seventeen edges are significant under both tests; "–" marks the Qwen2.5 denominator. Aggregate: mean ZS and Δ, median R_{self}. Incumbent Half and Self are from the onboarding evaluation run (within 0.5 F1 of Table 1); edges with an 8B endpoint use the locked-matrix baselines.

	New Writer: SmolLM2 → member			New Reader: member → SmolLM2		
Incumbent	ZS	$\Delta \uparrow$	$R_{\text{self}} \uparrow$	ZS	$\Delta \uparrow$	$R_{\text{self}} \uparrow$
Q06	39.1	+13.4	95.3	42.7	+19.1	96.1
Q17	45.2	+7.9	106.9	42.7	+19.1	96.3
Q25	30.7	+1.1	–	42.5	+18.9	95.2
Q8	58.0	+12.2	95.4	42.9	+19.3	98.2
L1	40.5	+11.4	97.5	42.5	+18.9	95.1
L3	50.4	+11.5	91.8	42.7	+19.1	96.3
L8	63.9	+19.4	99.8	42.9	+19.2	98.0
Phi	49.4	+8.3	94.1	41.8	+18.1	91.4
OLMo	43.8	+21.0	97.2	42.8	+19.2	96.6
Aggregate	**46.8**	**+11.8**	**96.3**	**42.6**	**+19.0**	**96.3**

Table 11: **All 18 Gemma-2-2B onboarding edges** (HotpotQA, $n = 1000$ per direction; format of Table 10).

	New Writer: G2 → member			New Reader: member → G2		
Incumbent	ZS	$\Delta \uparrow$	$R_{\text{self}} \uparrow$	ZS	$\Delta \uparrow$	$R_{\text{self}} \uparrow$
Q06	39.0	+13.2	94.2	52.9	+14.0	96.8
Q17	44.1	+6.9	92.5	53.1	+14.2	97.9
Q25	30.4	+0.8	–	53.7	+14.7	101.8
Q8	58.3	+12.5	97.4	53.2	+14.2	99.2
L1	40.2	+11.1	94.9	53.0	+14.1	97.4
L3	50.3	+11.4	90.6	53.4	+14.5	100.3
L8	64.5	+20.0	103.0	53.2	+14.3	99.4
Phi	49.0	+7.9	89.5	52.5	+13.6	93.9
OLMo	44.5	+21.7	100.4	53.7	+14.8	102.1

Table 12: **Two hubs retrained from scratch reproduce the locked hub column for column** (HotpotQA, $n = 1000$ per direction; F1 %). Half is the Reader-only baseline; Self reads the hub's own Writer for that Reader; ZS averages the hub's eight foreign Writers. The locked hub (seed 42) is Table 1a.

	Seed 43			Seed 44		
Reader	Half	Self	ZS	Half	Self	ZS
Qwen3-0.6B	25.9	39.3	39.7	25.9	40.8	40.4
Qwen3-1.7B	37.0	44.0	44.3	37.0	47.3	46.7
Qwen2.5-0.5B	29.6	30.1	28.9	29.6	30.1	30.8
Qwen3-8B	45.8	56.8	57.0	45.8	56.6	56.8
Llama-3.2-1B	29.1	43.4	43.3	29.1	40.2	40.7
Llama-3.2-3B	39.2	52.8	52.9	39.2	52.1	52.2
Llama-3.1-8B	44.5	62.3	61.9	44.5	63.3	62.8
Phi-4-mini	40.7	51.5	51.5	40.7	51.9	52.3
OLMo-2-1B	23.2	43.2	42.9	23.2	47.7	47.7
Median ZS−Half gain, 72 directions	+13.7 [+10.7, +14.5]			+11.9 [+10.7, +14.8]		
Directions with positive / negative interval	64 / 0			65 / 0		
Training cost, both hubs (H800-hours)	63.7 (3.5 per member)					

to 12 between the two signs in the remainder; the cross-hub gain retains a median 0.993–1.012 of the within-hub gain across the six pairings. A Reader trained in one run reads an artifact from a hub that did not exist at the time exactly as well as one from its own.

D.6 Full-context text ceilings and gain decomposition

Table 1a places every Reader among the text conditions of Section 4.1, and Table 4a summarizes the artifact-versus-Full comparison across all three datasets. Three regularities hold everywhere.

The Full ceiling is a distractor-taxed ceiling. Oracle beats Full by 8.4 to 24.0 F1 points for all 27 Reader×dataset pairs ($p < 10^{-5}$ in each): most of what separates any route from perfect evidence is the Reader's ability to ignore the eight irrelevant paragraphs, not access to the two relevant ones. Compressed memory pre-pays part of this tax at write time, which is why three of the four smallest Readers answer better from their own artifact than from the complete source (Self$-$Full of $+15.7$ for OLMo, $+4.6$ for Llama-1B, $+5.0$ for Qwen3-0.6B).

Shift reverses the ranking. The artifact's gap to Full is a property of the training distribution: the median ZS$-$Full gap moves from -2.7 F1 on HotpotQA to $+0.7$ on 2WikiMultihopQA and $+2.6$ on MuSiQue, where 24 of 72 directions favor the artifact significantly and one favors the text (Table 4a). A foreign model's compressed memory generalizes better than the raw text it summarizes.

The zero-shot gain decomposes into adaptation and content. Against the Empty control, task adaptation contributes a median $+6.0$ F1 over Half (positive in 64 of 72 paired intervals) and the foreign artifact's content a further $+3.2$ (positive in 72 of 72). The corresponding R_{content} is 87.0% (IQR $[75.4, 99.6]$, 95% CI $[82.2, 93.2]$) over 70 valid directions; denominator intervals that are not positive are undefined. The case studies of Appendix C show the two components at the level of single answers: the terse answer form is adaptation, the bridge entity is content.

E What the artifact carries

Three probes triangulate the semantics of a stored artifact. It behaves as an addressable record of task state: swap the artifact and the answer follows it; set it against evidence the Reader can see and the visible evidence prevails; perturb the surface form of writing or reading and nothing moves.

E.1 Addressability

Table 13 reports the held-out OpenHermes control of Section 4.5 per pair. Correct artifacts beat the next-example artifact in all nine directions (all nine after Holm correction), and under a shifted artifact the output *is* the next example's correct-artifact output in 300/300 cases for every pair: the Reader answers the artifact it is handed, not a residue of the prompt.

Table 13: **Artifact swap on held-out OpenHermes** ($n = 300$ per pair). "Correct" and "shifted" are token F1; H indicates an exact paired sign test on token-F1 wins and losses after Holm correction over the nine probes. Deltas are computed before rounding.

Writer→Reader	Correct	Shifted	Δ	H
Llama-1B→Qwen-0.6B	.0933	.0478	+.0455	yes
Qwen-0.6B→Llama-1B	.0639	.0420	+.0219	yes
Qwen-0.6B→Llama-3B	.0543	.0405	+.0139	yes
Qwen-0.6B→Llama-8B	.0657	.0398	+.0259	yes
Qwen-0.6B→OLMo-1B	.0866	.0479	+.0388	yes
Qwen-0.6B→Phi-mini	.0548	.0433	+.0115	yes
Qwen-0.6B→Qwen-1.7B	.0779	.0480	+.0299	yes
Qwen-0.6B→Qwen2.5	.0692	.0535	+.0157	yes
Qwen-0.6B→Qwen-8B	.0494	.0318	+.0176	yes

Planted decoys do not overwrite visible evidence. Replacing the answer-bearing string in the Writer view with a planted decoy while the complementary Reader view stays intact ($n = 600$) flips none of the 140 previously correct outputs to the planted answer, and 139 remain correct. The artifact steers the task; it does not overwrite evidence the Reader can verify for itself.

E.2 Composition and source length

Table 14 evaluates the concatenation of Eq. 8: the Writer view is split in half, each half is written to its own artifact, and the Reader consumes both at disjoint bands. The merged pair matches the single full-content artifact in all nine same-writer cells (median -0.2 F1, largest deficit 1.3), and merging halves written by *different* Writers—Qwen3-0.6B and Llama-1B—stays within 1.2–3.2 points of the single-writer reference across three Readers. Reading is linear in its memories; provenance does not fragment the cache.

Table 14: **Read-time composition** ($n = 300$ per cell; F1). Z_1, Z_2 are artifacts written from complementary halves of the Writer view; Z_{full} is written from the whole view. The final block merges halves written by two different Writers.

Writer(s)	Reader	Z_1	Z_2	$Z_1 \oplus Z_2$	Z_{full}
Qwen-0.6B	OLMo-1B	.3866	.3818	.4442	.4394
Qwen-0.6B	Phi-mini	.4844	.4212	.5116	.5246
Qwen-0.6B	Qwen-1.7B	.4196	.3585	.4437	.4565
Llama-1B	OLMo-1B	.3986	.3921	.4557	.4545
Llama-1B	Phi-mini	.4756	.4291	.5188	.5206
Llama-1B	Qwen-1.7B	.4160	.3742	.4426	.4559
Phi-mini	OLMo-1B	.3862	.3844	.4364	.4382
Phi-mini	Phi-mini	.4926	.4266	.5378	.5341
Phi-mini	Qwen-1.7B	.4222	.3661	.4526	.4610
Q06+L1	OLMo-1B	.3866	.3921	.4226	.4545
Q06+L1	Phi-mini	.4844	.4291	.5125	.5246
Q06+L1	Qwen-1.7B	.4196	.3742	.4401	.4565

Table 15 pads the Writer view with retrieval distractors to a controlled input length while the question stays answerable from the original half. The artifact's footprint and quality are both flat: over a $16\times$ length range (512 to 8,192 tokens) zero-shot F1 moves within 1.2 points in both instrumented directions, while re-reading the same stored text peaks at 1,024 tokens and falls by 4.8 and 4.3 points from there to 8,192 as the Reader's context fills with distractors. In the weaker-Reader direction the artifact leads stored text at every length, by 8.1 to 12.6 points with every paired interval positive; in the strong-Reader direction the two routes are level at four of five lengths. Compression concentrates exactly where reading is hard.

Table 15: **Zero-shot F1 versus source length** ($n = 300$ per point). Latent reads the 64 KB artifact written from the padded source; Stored text re-reads the padded source itself. Brackets give the paired 95% interval on Latent$-$Stored.

Direction	Length	Latent	Stored text	Δ [95% CI]
L3→Q06	512	.4104	.2978	+.113 [+.071,+.155]
L3→Q06	1024	.4056	.3242	+.081 [+.040,+.124]
L3→Q06	2048	.4071	.2930	+.114 [+.075,+.153]
L3→Q06	4096	.4062	.2800	+.126 [+.087,+.165]
L3→Q06	8192	.4009	.2761	+.125 [+.089,+.164]
Q06→L3	512	.5266	.5369	−.010 [−.041,+.020]
Q06→L3	1024	.5154	.5565	−.041 [−.073,−.010]
Q06→L3	2048	.5219	.5437	−.022 [−.053,+.010]
Q06→L3	4096	.5205	.5235	−.003 [−.035,+.029]
Q06→L3	8192	.5211	.5135	+.008 [−.026,+.041]

E.3 Storage, surface form, and load

A stored artifact is a stable object. Table 16 records the fidelity probes summarized in Section 4.5: transporting the virtual band by 128 or 1,024 positions, quantizing the artifact to 8 bits, and the worst cell of each prompt grid move F1 by -0.94 to $+1.65$ points, seven of the eight probes by less than

one point; five reads of one serialized artifact are bitwise identical. Serialized size is 65,536 bytes in fp16 and 32,896 at int8 for every Writer and every source length.

Table 16: **Representation fidelity under position shift, 8-bit quantization, and adversarial prompt perturbation** (F1 %; Δ vs. the unperturbed reference). Serialized size is 64.0 KiB in fp16 and 32.1 KiB at int8 (int8 writes cost 22.8–23.8 ms).

Probe	Endpoint	Perturbation	n	Ref. F1	Test F1	ΔF1
Position	Q06	offset 128	500	39.56	41.21	+1.65
Position	Q06	offset 1024	500	39.56	39.62	+0.06
Position	L3	offset 128	500	52.41	52.77	+0.36
Position	L3	offset 1024	500	52.41	52.44	+0.03
Quantize	Q06	8-bit	500	39.56	40.26	+0.70
Quantize	L3	8-bit	500	52.41	52.43	+0.02
Prompt	Q06→L1	worst of grid	300	44.40	43.48	−0.92
Prompt	L1→Phi	worst of grid	300	53.30	52.36	−0.94

Table 17 gives the complete prompt grids—the Writer's input format against the Reader's instruction style. Across the nine combinations of each grid F1 moves within 1.1 points: the interface keys on content, not on formatting.

Table 17: **Complete prompt-robustness grids** ($n = 300$ per cell; F1). Rows vary the Writer's input format; columns vary the Reader's instruction.

Pair	Writer format	Canonical	Brief	Polite
Qwen-0.6B→Llama-1B	plain	.4440	.4368	.4444
Qwen-0.6B→Llama-1B	bulleted	.4373	.4373	.4427
Qwen-0.6B→Llama-1B	wrapped	.4399	.4348	.4417
Llama-1B→Phi-mini	plain	.5330	.5292	.5298
Llama-1B→Phi-mini	bulleted	.5276	.5316	.5238
Llama-1B→Phi-mini	wrapped	.5338	.5342	.5236

Figure 5 shows the distractor sweep: with eight irrelevant artifacts concatenated to the correct one, the Reader keeps 87.1% (Llama-3.2-3B) and 80.3% (Qwen3-0.6B) of its single-artifact F1. Reading degrades gracefully with load, and a Reader handed several memories at once still finds the one that matters.

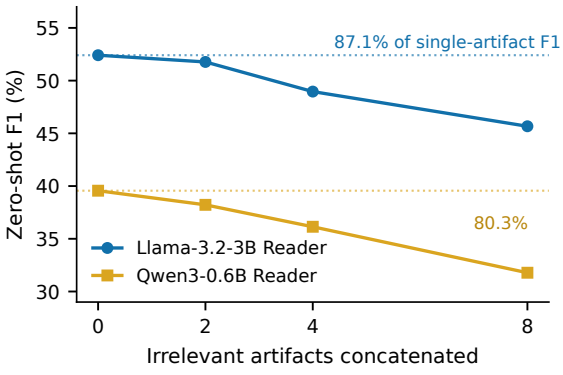


Figure 5: **Stored artifacts degrade gracefully under distractor load.** Zero-shot F1 as k irrelevant artifacts are concatenated with the correct one ($n = 500$ per point); dotted lines mark the single-artifact reference. Exact values at $k = 0, 2, 4, 8$: Qwen3-0.6B .3956/.3822/.3614/.3178; Llama-3.2-3B .5241/.5177/.4896/.4567.

Wall-clock accounting. The costs of Table 4d are medians over $n = 500$ instrumented reads per route ($n = 100$ for live decoding) on three Reader endpoints of one H800 with warm caches. *Read*

loads the serialized artifact, builds the prefix cache, and prefills the question; *re-read* prefills the ten stored paragraphs (both views) and the question; *live T2T* has the sharer decode its message (query-aware, 50–59 tokens; query-agnostic, 153–200 tokens) and the receiver prefill it. End to end, a 32-token answer from a stored artifact takes 104–105 ms at batch 1 against 1026–1213 ms for the query-aware live handoff and 2536–3725 ms for the query-agnostic one. At batch 32 the artifact route completes an answer in 31–63 ms per sample, the re-read route in 68–125 ms, and the stored-message route in 49–87 ms. Forty reads amortize a write below 1 ms each. Message lengths over the 24,300 exchanges of the 81 cells have median 56 words, IQR $[34, 103]$, and 90th percentile 143.

F THE LATENT ROUTE AGAINST TEXT AND AGAINST TRAINED CHANNELS

This section gives the cell-level structure behind Table 4b–c: the live and trained text routes, the 18-pair C2C comparison, and the jointly trained shared space. The training protocol of each reference is in Table 7.

F.1 LIVE TEXT HANDOFF: THE ADVANTAGE FOLLOWS WRITER SCALE

The 72-cell comparison of Section 4.5 has one organizing variable, the Writer's scale. Writing an artifact is scale-invariant—an artifact from a 0.5B Writer serves every Reader about as well as one from an 8B Writer (Table 9)—but decoding a good intermediate text message is not. Table 18 makes the split explicit on the 56 decode-matched directions, where both routes decode greedily: Writers at or below 1.7B lose to their own artifacts in all 28 cells, by a mean of 12.0 F1 points, while Writers at 3B and above roughly match their text messages (mean $+3.0$ points, 17 of 28 cells won). Even in the large-Writer half, no cell's paired F1 interval favors T2T.

Table 18: **The latent route's advantage over a live text handoff is governed by Writer scale.** Direct-minus-T2T F1 by Writer over the 56 decode-matched directions ($n = 300$ per cell; seven Reader cells per row). CI^+ counts cells whose paired F1 interval favors the direct route and H^+ counts Holm-corrected EM wins; no cell's interval favors T2T.

Writer	Mean ΔF1	Median ΔF1	Wins	CI^+	H^+
Qwen3-0.6B	$+.132$	$+.126$	7/7	7	7
Llama-3.2-1B	$+.114$	$+.107$	7/7	7	5
OLMo-2-1B	$+.116$	$+.112$	7/7	7	6
Qwen3-1.7B	$+.116$	$+.110$	7/7	5	4
Llama-3.2-3B	$+.021$	$+.010$	5/7	2	2
Phi-4-mini	$+.055$	$+.059$	5/7	3	2
Llama-3.1-8B	$+.004$	$-.002$	3/7	1	1
Qwen3-8B	$+.040$	$+.022$	4/7	2	2

The Reader side carries the complementary asymmetry: Llama-3.1-8B and OLMo-2-1B gain the most from consuming artifacts instead of text ($+.147$ and $+.135$ mean F1 over their columns of the same subset), and these are the Readers with the largest zero-shot headroom in the locked matrix. The two margins jointly identify where the latent route pays most: small or cheap Writers feeding capable Readers, exactly the configuration in which serializing state through tokens costs the most quality and, per Table 4d, the most time. Over the full 72-cell matrix, including the two Qwen2.5 endpoints whose default T2T configuration applies repetition penalty 1.1, the paired F1 interval favors the direct route in 42 cells and T2T in 3, and Holm-corrected EM favors the direct route in 36 and T2T in none.

F.2 A TEXT COMMUNICATOR TRAINED FOR THE JOB

The live route uses each frozen Writer's native messages; the tuned control of Table 7 trains the sender at the memory adapters' own budget. Table 19 shows the outcome on six cross-family directions of the locked slice ($n = 1000$): the zero-shot artifact stays ahead by a median $+7.1$ F1 ($+2.0$ to $+11.8$), with a positive paired interval and a Holm-corrected EM win in five of six directions and no reversal. The latent advantage of the main text is a property of the medium, not of an untrained sender.

Table 19: **The zero-shot artifact against a text communicator trained at the members' own budget** ($n = 1000$ per direction; F1 %). Half and ZS are re-evaluated within this comparison run on the same n=1000 slice under the protocol of Section 4.1, paired by example with the channel outputs; they differ from the canonical cells of Table 1 by at most 0.5 F1. Brackets give the paired 95% interval on ZS−Tuned; EM flips count examples that only the direct route (D) or only the text route (T) answers exactly, with the Holm-corrected McNemar p over the six directions.

Direction	Half	Tuned T2T	MEMHUB ZS	ZS−Tuned [95% CI]	EM flips D / T	Holm p
Qwen-0.6B→Llama-1B	29.1	34.7	41.6	+6.9 [+3.4, +10.4]	218 / 112	1.7×10^{-8}
Qwen-0.6B→Phi-mini	41.1	43.6	49.4	+5.8 [+2.3, +9.3]	234 / 137	1.1×10^{-6}
Qwen-0.6B→Qwen-1.7B	37.3	42.7	44.8	+2.0 [−1.8, +5.7]	190 / 159	1.1×10^{-1}
Llama-1B→Qwen-0.6B	25.8	26.7	38.5	+11.8 [+9.4, +14.3]	154 / 35	3.1×10^{-18}
Llama-1B→Phi-mini	41.1	42.3	49.6	+7.3 [+5.2, +9.5]	140 / 39	7.3×10^{-14}
Llama-1B→Qwen-1.7B	37.3	35.3	44.3	+8.9 [+6.5, +11.3]	139 / 44	4.5×10^{-12}
Median				+7.1 (CI$^+$ 5, CI$^-$ 0)		Holm$^+$ 5 / 0

F.3 PAIR-TRAINED C2C AT THREE BUDGETS

Table 20 lists all 18 C2C pairs. C2C clears Half in all 18 (median $+9.4$ F1), including every cross-tokenizer pair, and the zero-shot artifact leads it by a median $+4.3$, significantly in 10 of 18 and never the reverse (Holm-corrected EM 10 / 0). The margin holds against the smaller projectors: $+3.9$ (13 / 0) against the 1200-step projector and $+5.5$ (17 / 0) against the 600-step one. C2C is below PAIRFT in 17 of 18 pairs (median 4.6 F1; 5.1 on shared, 4.5 on disjoint tokenizers). C2C spends a projector and a training run on every ordered pair; the artifact spends nothing on the pair.

Table 20: **All 18 C2C pairs** ($n = 1000$, paired by example; F1 %). The 600-, 1200-, and 4800-step settings are a convergence sweep; the headline comparison uses the matched effective budget and the strongest development-selected setting. C2C is trained from its official components at batch 32, with its TokenAligner on the cross-tokenizer pairs (×tok); Half, PAIRFT, and ZS are the locked-matrix cells of Table 1. Brackets give the paired 95% interval against the dev-selected 4800-step projector.

Sharer→Receiver	Half	$C2C_{600}$	$C2C_{1200}$	$C2C_{4800}$	PAIRFT	ZS	$C2C_{4800}$−Half [95% CI]	ZS−$C2C_{4800}$ [95% CI]
Q06→Q17	37.0	39.3	42.6	40.8	44.7	44.7	+3.8 [+1.4, +6.1]	+3.9 [+1.4, +6.1]
Q06→Q8	45.8	47.2	45.4	51.1	58.6	57.6	+5.3 [+3.2, +7.6]	+6.5 [+4.4, +8.6]
Q17→Q06	25.9	25.6	36.1	30.1	38.3	38.9	+4.3 [+2.0, +6.5]	+8.8 [+6.5, +11.1]
Q17→Q8	45.8	48.7	49.6	51.5	57.3	58.0	+5.7 [+3.4, +8.0]	+6.5 [+4.4, +8.7]
Q8→Q06	25.9	33.9	34.5	31.6	39.4	38.3	+5.7 [+3.3, +8.3]	+6.7 [+4.1, +9.2]
Q8→Q17	37.0	35.8	38.4	42.9	44.5	44.7	+5.8 [+3.4, +8.2]	+1.8 [−0.5, +4.1]
L1→L3	39.2	48.9	49.6	50.6	53.9	51.3	+11.3 [+8.9, +13.8]	+0.8 [−1.4, +2.9]
L1→L8	44.5	56.4	56.0	55.8	64.1	63.8	+11.3 [+8.8, +13.9]	+8.0 [+5.9, +10.2]
L3→L1	29.1	39.7	35.7	41.8	40.7	41.4	+12.7 [+10.2, +15.3]	−0.4 [−2.7, +2.0]
L3→L8	44.5	55.9	54.9	59.0	63.4	63.7	+14.5 [+12.0, +17.0]	+4.6 [+2.5, +6.7]
L8→L1	29.1	35.6	36.9	41.1	41.7	40.2	+12.0 [+9.5, +14.5]	−1.0 [−3.2, +1.2]
L8→L3	39.2	46.8	49.0	49.2	55.1	50.5	+10.0 [+7.5, +12.5]	+1.3 [−0.9, +3.5]
L1→Q06 (×tok)	25.9	33.3	36.5	36.9	40.9	38.9	+11.0 [+8.3, +13.6]	+2.0 [−0.6, +4.3]
Q06→L1 (×tok)	29.1	31.9	30.8	34.5	42.0	41.4	+5.4 [+3.3, +7.7]	+6.9 [+4.6, +9.1]
Q17→L3 (×tok)	39.2	46.3	50.7	50.0	54.3	50.9	+10.8 [+8.5, +13.0]	+0.9 [−1.1, +3.0]
L3→Q17 (×tok)	37.0	39.6	40.8	40.3	45.0	44.9	+3.2 [+0.8, +5.7]	+4.6 [+2.3, +6.9]
Phi→OLMo (×tok)	23.2	35.9	35.0	35.4	43.7	43.9	+12.2 [+10.0, +14.5]	+8.5 [+6.3, +10.7]
OLMo→Phi (×tok)	40.7	45.1	49.6	49.5	50.3	49.7	+8.8 [+6.3, +11.4]	+0.1 [−2.0, +2.4]
Median, 18 pairs							+9.4 (CI$^+$ 18, CI$^-$ 0)	+4.3 (CI$^+$ 10, CI$^-$ 0)

These runs are a convergence sensitivity analysis around the matched headline setting; they do not define separate fairness conditions. The extra budget buys little (Table 21). Per pair the three budgets cost a median 2.1, 4.1, and about 16 H800-hours (529M projector parameters; 269–680M across pairs), against 3.5 hours for a member's complete Writer–Reader pair. These are implementation-cost comparisons: at 1200 steps C2C matches a member in compute, and at 4800 steps it matches a member in samples at roughly five times the cost. The unselected test-slice trajectory rises $+2.7$ F1 from step 1200 to 2400 (14 of 18 paired intervals positive, 2 negative) and moves -0.6 from 2400 to 4800 (2 / 3); the dev-selected 4800-step projector adds a median $+0.4$ over the independently trained 1200-step one (6 / 2) and $+2.7$ over the 600-step one (11 / 1). Reading the final 4800-step projector instead of the dev-selected one leaves the comparison unchanged (median $+4.9$ in the artifact's favor, 11 / 0).

Table 21: **C2C plateaus after 2400 steps.** Locked-slice F1 (%, $n = 1000$) of the current projector at five milestones of the 4800-step run, with no checkpoint selection; the 1200-step run and the zero-shot artifact are shown for reference. Lower block: median paired increment between milestones over the 18 pairs, with CI^+ / CI^- counts.

	4800-step run, projector at step						
Sharer→Receiver	600	1200	2400	3600	4800	$C2C_{1200}$	ZS
Q06→Q17	33.9	35.8	40.5	40.1	39.5	42.6	44.7
Q06→Q8	47.5	37.2	49.7	46.3	49.0	45.4	57.6
Q17→Q06	23.8	9.4	23.6	20.5	23.2	36.1	38.9
Q17→Q8	30.2	44.5	51.5	48.4	45.5	49.6	58.0
Q8→Q06	33.6	29.3	31.6	28.9	31.0	34.5	38.3
Q8→Q17	40.4	40.0	42.6	42.9	43.4	38.4	44.7
L1→L3	48.5	46.9	49.5	50.6	50.5	49.6	51.3
L1→L8	51.1	55.9	58.8	58.0	58.2	56.0	63.8
L3→L1	40.0	39.3	41.5	41.3	41.2	35.7	41.4
L3→L8	47.3	53.5	58.6	58.9	57.0	54.9	63.7
L8→L1	38.5	37.7	42.0	38.6	39.4	36.9	40.2
L8→L3	47.5	49.4	44.9	44.8	43.6	49.0	50.5
L1→Q06 (×tok)	34.8	35.0	38.3	36.0	37.2	36.5	38.9
Q06→L1 (×tok)	30.8	33.6	25.6	32.0	31.9	30.8	41.4
Q17→L3 (×tok)	45.4	49.0	51.2	51.5	52.0	50.7	50.9
L3→Q17 (×tok)	37.3	36.5	39.2	40.0	40.2	40.8	44.9
Phi→OLMo (×tok)	35.4	33.5	35.9	40.5	39.5	35.0	43.9
OLMo→Phi (×tok)	43.3	47.7	49.3	47.8	48.5	49.6	49.7
Median, 18 pairs	39.3	38.5	42.3	42.1	42.3	41.7	44.8

Paired increment between milestones (median [IQR] F1 points; CI^+ / CI^- of 18)		
Step 1200 − 600	−0.1 [−1.4, +3.4]	6 / 3
Step 2400 − 1200	+2.7 [+2.2, +4.6]	14 / 2
Step 4800 − 2400	−0.6 [−1.1, +0.8]	2 / 3
Step 4800 − 1200	+2.6 [+1.7, +3.7]	11 / 1

F.4 JOINTLY TRAINED SHARED SPACE

Table 22 compares the artifact with the four-member space trained in one computation graph (Table 7) on all 12 foreign directions of the locked slice, $n = 1000$, paired by example. Joint training leads in every direction by 0.5 to 2.5 F1—median $+1.5$, significant in 5 of 12—and on the self-loops by $+1.8$ (2 of 4 significant), against gains over Half of $+14.0$ and $+12.6$. Those 1.5 points are the entire price of never coupling members: a fifth member joins the joint space by training through frozen incumbents' adapters— Dery et al. (2026, §5) show that a subset of incumbents suffices and that the newcomer then translates zero-shot to the held-out ones—so some incumbents' parameters and activations must be available and enter the new member's computation graph, whereas the hub admits one with a single adapter pair trained against the anchor alone (Section 3.4). Both designs store one adapter pair per model; they differ in what joining requires.

Table 22: **Independent training against the anchor versus joint training of four members in one graph** ($n = 1000$ per direction; F1 %). Half and ZS are re-evaluated within this comparison run on the same n=1000 slice under the protocol of Section 4.1, paired by example with the channel outputs; they differ from the canonical cells of Table 1 by at most 0.5 F1. Brackets give the paired 95% interval.

Direction	Half	MEMHUB ZS	Joint ZS	Joint−MEMHUB [95% CI]	Joint−Half [95% CI]	ZS−Half [95% CI]
L1→Q06	25.8	38.5	39.7	+1.2 [−0.8, +3.1]	+14.0 [+11.5, +16.4]	+12.8 [+10.3, +15.2]
Phi→Q06	25.8	39.3	39.7	+0.5 [−1.5, +2.4]	+14.0 [+11.5, +16.4]	+13.5 [+11.0, +15.9]
OLMo→Q06	25.8	38.7	39.7	+1.0 [−0.9, +2.9]	+14.0 [+11.4, +16.4]	+12.9 [+10.4, +15.4]
Q06→L1	29.1	41.6	42.9	+1.3 [−0.4, +3.1]	+13.8 [+11.4, +16.3]	+12.5 [+10.2, +14.9]
Phi→L1	29.1	40.4	42.9	+2.5 [+0.8, +4.3]	+13.7 [+11.4, +16.2]	+11.3 [+8.9, +13.8]
OLMo→L1	29.1	41.0	43.5	+2.5 [+0.7, +4.3]	+14.3 [+12.0, +16.8]	+11.9 [+9.4, +14.3]
Q06→Phi	41.1	49.4	51.9	+2.5 [+0.8, +4.2]	+10.7 [+8.5, +13.2]	+8.3 [+6.2, +10.5]
L1→Phi	41.1	49.6	51.1	+1.5 [−0.2, +3.4]	+10.0 [+7.7, +12.3]	+8.5 [+6.4, +10.7]
OLMo→Phi	41.1	49.7	51.6	+1.9 [+0.1, +3.5]	+10.5 [+8.2, +12.8]	+8.6 [+6.5, +10.9]
Q06→OLMo	22.8	44.1	45.6	+1.5 [−0.2, +3.4]	+22.8 [+20.2, +25.5]	+21.3 [+18.6, +23.9]
L1→OLMo	22.8	44.5	46.3	+1.8 [+0.1, +3.4]	+23.5 [+20.9, +26.2]	+21.7 [+19.2, +24.5]
Phi→OLMo	22.8	44.0	45.5	+1.5 [−0.3, +3.3]	+22.7 [+20.1, +25.4]	+21.2 [+18.6, +23.8]
Median				+1.5 [+1.3, +2.0]	+14.0	+12.6

G MECHANISM: SENSITIVITY, PORTABILITY, AND REPLICATION

The studies here bound the anchor analysis of Section 4.4 from four directions: exact sweep values with alignment diagnostics, the factorial replicated on a second pair and seed, portability of the coordinates across embedders, judges, and corpora, and budget and seed replications of the headline references. Each is a consequence of Propositions 1–2 (Appendix A) or a measurement of the basis they describe.

G.1 ANCHOR SWEEP AND ALIGNMENT DIAGNOSTICS

Table 23 gives the exact values behind Figure 3. Anchoring two of 32 slots lifts zero-shot transfer from $.004/.000$ to $.464/.435$ F1 and free-slot alignment from $.005$ to $.888$; sixteen anchors reach the same plateau ($.468/.445$). The anchor weight is likewise unconstraining: λ between $.25$ and 2 moves F1 within 4.4 points.

Table 23: **Exact anchor-sweep values for Figure 3** ($n = 300$ per direction; Q06/L1 pair, judge loss disabled; free-slot top-1 alignment over 150 shared probe texts; checkpoints trained separately from those of Table 3a). The $m_a = 0$ column is the no-anchor control.

	m_a=0	m_a=2	m_a=4	m_a=8	m_a=16	m_a=8, λ=.25	m_a=8, λ=2
ZS F1, Q06→L1	.004	.464	.473	.458	.468	.473	.502
ZS F1, L1→Q06	.000	.435	.406	.430	.445	.425	.446
Free-slot top-1 alignment	.005	.888	.904	.858	.874	.748	.903

The diagnostics of Table 24 separate two rates. Alignment of the anchored slots improves smoothly with m_a (cosine .878 at m_a=2 to .928 at m_a=16), while free-slot retrieval saturates at once: .888 top-1 at m_a=2 and flat thereafter. The unconstrained capacity aligns as a body the moment any basis is pinned—the signature of a gauge fixed once, not of content taught slot by slot. Masking completes the picture: zeroing either slot group at read time collapses F1 below .014 at every setting, so anchored and free slots both carry task state, in one shared frame.

Table 24: **Raw-alignment and slot-masking diagnostics for the anchor sweep.** Alignment (cosine and top-1 retrieval) uses 150 shared texts; masking evaluates Q06→L1 memory-only F1 on the first 100 with all slots, with the free slots zeroed, and with the anchored (A) slots zeroed. At $m_a = 0$ the two groups denote the same fixed slot partition, applied for diagnostics only.

Setting	A cos	A top-1	Free cos	Free top-1	Full	Zero free	Zero A
$m_a = 0$	-.0046	.0025	-.0063	.0061	.0072	.0019	.0058
$m_a = 2$	.8776	1.0000	.6659	.8878	.4495	.0076	.0083
$m_a = 4$	.9000	1.0000	.7504	.9036	.4408	.0131	.0105
$m_a = 8$	.9190	.9967	.7843	.8578	.4261	.0087	.0135
$m_a = 16$	.9277	.9992	.8141	.8737	.4233	.0069	.0064
$m_a = 8, \lambda = .25$	.9291	.9642	.8343	.7478	.4163	.0122	.0109
$m_a = 8, \lambda = 2$	.9091	.9992	.7932	.9025	.4623	.0090	.0107

G.2 FACTORIAL REPLICATION AND DIAGNOSTICS

Tables 25 and 26 complete Table 3a. Both single-constraint conditions transfer on a second cross-family pair (Llama-1B/OLMo-1B) and a second seed of the development pair, eight of eight gain intervals positive; geometric pinning retains .84–.99 of its own self-loop F1, judge pinning .40–.93.

The diagnostics of Table 26 dissociate the two mechanisms quantitatively, as Remark 2 states. Geometric pinning aligns bases in raw coordinates (free-slot top-1 $.85$); judge pinning leaves raw alignment at noise ($.02$) and only weak centered structure ($.16$), yet transfers at the same strength; and zeroing either slot group at read time collapses every interoperable condition, so both constraint types organize—rather than carry—the full 32-slot code. Appendix J.4 reads the same three conditions with linear label probes fit on other members: geometric pinning passes, judge pinning and no anchor sit at the label prior.

Table 25: **Factorial replication on a second cross-family pair and a second seed** ($n = 300$ per direction). Gains are ZS−Half with paired 95% intervals; ZS/Self is retention against the condition's own self-loop.

Condition	Pair	Direction	ZS	Gain [95% CI]	ZS/Self
Geometric	L1/OLMo (s42)	L1→OLMo	.450	+.182 [+.135,+.227]	.990
Geometric	L1/OLMo (s42)	OLMo→L1	.467	+.118 [+.075,+.168]	.982
Judge	L1/OLMo (s42)	L1→OLMo	.408	+.140 [+.093,+.185]	.699
Judge	L1/OLMo (s42)	OLMo→L1	.395	+.046 [+.002,+.090]	.398
Geometric	Q06/L1 (s43)	L1→Q06	.393	+.127 [+.085,+.169]	.964
Geometric	Q06/L1 (s43)	Q06→L1	.433	+.084 [+.043,+.125]	.844
Judge	Q06/L1 (s43)	L1→Q06	.438	+.172 [+.130,+.213]	.932
Judge	Q06/L1 (s43)	Q06→L1	.416	+.066 [+.026,+.107]	.541

Table 26: **Complete factorial alignment and masking diagnostics** (Q06/L1, seed 42). Alignment is cross-writer top-1 retrieval over 150 shared texts, in raw and centered coordinates; the last three columns give Q06→L1 memory-only F1 with all slots and with one group zeroed.

	Raw top-1		Centered top-1		Memory-only F1		
Condition	Anchored	Free	Anchored	Free	Full	Zero anch.	Zero free
Neither	.0029	.0042	.0183	.0158	.0000	.0000	.0000
Geometric only	.9871	.8521	.9879	.9754	.4229	.0158	.0082
Judge only	.0196	.0233	.1492	.1638	.3600	.0079	.0122
Both	.9817	.5275	.9825	.9225	.4492	.0037	.0143

G.3 Portable coordinates

Nothing in the interface depends on the anchor and the members sharing a family, a corpus, or a calibration step. Table 27 rebuilds the entire anchor from non-Qwen parts—BGE-M3 (Chen et al., 2024) as the embedder behind Eq. 1 and a Llama-3.2-1B no-anchor Reader as the judge—and retrains Qwen3-0.6B, Llama-1B, and OLMo-1B against it. The six cross-directions hold at production level (median paired difference $+1.4$ F1); five of six paired intervals cross zero and the sixth, L1→Q06, is 3.5 points lower.

Table 27: **Full anchor substitution** ($n = 300$ per direction). Brackets give the paired 95% interval on (non-Qwen anchor)−(production anchor) zero-shot F1.

Direction	Non-Qwen anchor	Production anchor	Δ [95% CI]
Q06→L1	.4547	.4421	+.013 [−.023,+.048]
L1→Q06	.3990	.4344	−.035 [−.070,−.002]
Q06→OLMo	.4830	.4662	+.017 [−.017,+.052]
OLMo→Q06	.3944	.4133	−.019 [−.055,+.016]
L1→OLMo	.4867	.4723	+.014 [−.019,+.048]
OLMo→L1	.4469	.4303	+.017 [−.017,+.050]

Table 28 and Figure 6 quantify the coordinate diagnosis from the converse direction. Fifty paired unlabeled texts and a single Procrustes map make the unanchored development pair readable (0 to $.448/.393$ F1), and a further 950 texts add nothing; anchored pairs start within $.014$ of their plateau at zero calibration texts and stay there. The anchor is exactly this calibration, precomputed—paid once, publicly, before any member exists.

Compatibility does not come from shared training examples either. Training the Qwen/Llama development pair on non-overlapping corpus halves (25,000 HotpotQA and 12,500 OpenHermes examples each) preserves bidirectional zero-shot readout and both anchored- and free-slot retrieval at the level of the shared-data control (Table 29).

Table 28: **Complete Procrustes calibration spectra.** Columns give the number of paired unlabeled texts used to estimate one centered, scaled orthogonal map. $n = 300$ except for unanchored Qwen-0.6B→Llama-1B ($n = 200$).

Regime	Direction	0	50	200	1000
Unanchored	Llama-1B→Qwen-0.6B	.0000	.4477	.4482	.4470
Unanchored	Qwen-0.6B→Llama-1B	.0007	.3925	.3925	.3932
Anchored	Llama-1B→Llama-3B	.5127	.5213	.5187	.5252
Anchored	Qwen-0.6B→Phi-mini	.5271	.5224	.5254	.5132

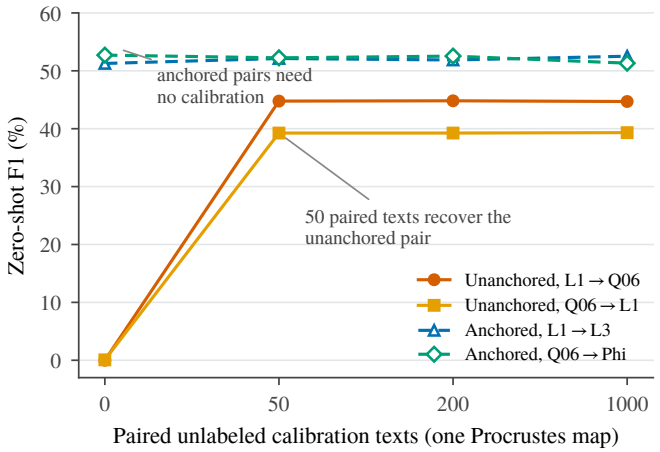


Figure 6: **Anchoring makes cross-model readout calibration-free.** Zero-shot F1 as a function of the number of paired unlabeled texts used to estimate one Procrustes map between slot bases. The unanchored Qwen/Llama development pair is unreadable without calibration and is recovered by a single linear map from 50 paired texts; the anchored Llama-1B→Llama-3B and Qwen→Phi pairs are already at plateau with zero texts and move by at most .014 thereafter.

Table 29: **Shared versus disjoint member corpora.** Task performance uses $n = 300$ per direction. Alignment uses 150 common probe texts and reports raw top-1 cross-writer retrieval for the 16 anchored (A) and 16 free slots; the final three columns of the lower panel are F1 for the Qwen→Llama probe with all slots or with one slot group zeroed.

Regime	Direction	Half	Self	ZS	ZS−Self
Shared	Qwen→Llama	.3313	.4630	.4588	−.0042
Shared	Llama→Qwen	.2657	.4295	.4210	−.0085
Disjoint	Qwen→Llama	.3313	.4856	.4805	−.0051
Disjoint	Llama→Qwen	.2657	.4295	.4257	−.0038

Regime	Raw A top-1	Raw free top-1	Full slots	Zero free	Zero A
Shared	.9646	.5246	.4239	.0040	.0076
Disjoint	.9771	.5083	.4375	.0116	.0069

G.4 Reference budget and seed replication

The 96.1% headline is measured against a converged reference under the matched protocol above. Doubling the PairFT budget from 600 to 1200 steps is a robustness check and moves the reference by at most .0139 F1 on four representative cross-family cells and the implied median retention from .935 to .945 (Table 30; Half, ZS, and both references there are from that run, within 0.5 F1 of Table 1); the pair-trained references of the two replication hubs confirm it from the other side, adding $+0.5$ and -0.3 F1 over zero-shot reading (median over 16 directions per hub; significantly positive in 3 and 2, significantly negative in 0 and 3).

Re-training the development pair under three seeds and both target chunkings reproduces ZS/Self between .982 and 1.038 in every run (Table 31), and the full-hub replications of Table 2b extend the

Table 30: **PAIRFT reference at doubled training budget** ($n = 1000$). The last column is 1200-step minus 600-step PAIRFT F1 with a paired-bootstrap interval.

Writer	Reader	Half	ZS	600	1200	R_{600}	R_{1200}	Δ [95% CI]
Qwen-0.6B	Llama-1B	.2912	.4162	.4187	.4197	.981	.973	+.0010 [−.0131, +.0143]
Llama-1B	Qwen-0.6B	.2576	.3854	.4094	.4096	.842	.841	+.0002 [−.0146, +.0156]
Llama-3B	Qwen-1.7B	.3728	.4453	.4543	.4497	.890	.943	−.0046 [−.0154, +.0067]
Phi-mini	OLMo-1B	.2280	.4398	.4378	.4517	1.009	.947	+.0139 [+.0021, +.0256]

same result to all nine members and to reading across chunkings: 28 edges of the locked matrix and 112 of the 144 directions between the locked hub and each replication hub cross chunkings.

Table 31: **Replication of the Qwen/Llama development pair across three seeds and both geometric-target chunkings** ($n = 500$ per cell). Zero-shot transfer matches self-transfer in every run.

Seed	Chunk	Writer	Reader	Self	ZS	ZS/Self
42	128	Llama-1B	Qwen-0.6B	.4141	.4125	.996
42	128	Qwen-0.6B	Llama-1B	.4208	.4367	1.038
43	64	Llama-1B	Qwen-0.6B	.4302	.4329	1.006
43	64	Qwen-0.6B	Llama-1B	.4182	.4173	.998
44	64	Llama-1B	Qwen-0.6B	.4231	.4154	.982
44	64	Qwen-0.6B	Llama-1B	.4426	.4413	.997

H DESIGN DECISIONS

The development studies that fixed the architecture use controlled Qwen3-0.6B self-loops on the 200-item development set; they precede the final anchor configuration and are never pooled with the locked evaluation families.

H.1 READER ROUTE AND WRITER LAYERS

R-append is the strongest partitioned-QA route, ahead of R-residual by .021 F1 and of a soft prefix by .065, and mid-to-deep Writer features lead shallow features by .041 while slightly exceeding a network-wide spread (Table 32). The two metrics trade off across routes: the configurations that maximize memory-only readout are not the ones that best support the deployment task, and the interface is selected for the latter.

Table 32: **Self-loop architecture ablations on Qwen3-0.6B.** The target metric is partitioned HotpotQA F1; memory-only F1 exposes route behavior. All rows use $\lambda_{\text{fid}} = 0$.

Factor	Setting	Partitioned	Memory-only
Reader route	R-append	.4474	.1057
	R-residual	.4267	.1269
	Soft prefix	.3825	.1750
Writer layers	Mid-to-deep default	.4474	.1057
	Spread-9	.4396	.1326
	Shallow-9	.4067	.1420

H.2 INTERFACE SIZE

Performance varies non-monotonically within .021 F1 over the tested 8–64 slot and 512–1024 dimension settings (Table 33), so all members adopt the fixed 32×1024 interface. The flat landscape is a feature: a public format should be an operational standard that members commit to cheaply, not a capacity knob each member is tempted to retune.

Table 33: **Qwen3-0.6B development sweep used to select the fixed artifact interface.** Partitioned and memory-only entries are F1; all four runs use $\lambda_{\text{fid}} = .5$.

Slots × dimension	Partitioned	Memory-only
8×1024	.3372	.0929
32×1024	.3165	.1228
64×1024	.3266	.0980
32×512	.3208	.1181

H.3 Capacity scaling and format migration

Two development studies on the strongest instrumented member (Llama-3.2-3B self-loops) measure the headroom above the fixed contract (Table 34). Reader-side capacity scales log-linearly with slot count: self-loop F1 rises $.590 \rightarrow .595 \rightarrow .609 \rightarrow .621$ over $32 \rightarrow 64 \rightarrow 128 \rightarrow 256$ slots—roughly one F1 point per doubling—meeting this member's own full-context reference ($.595$) at 64 slots. Migration between formats is graceful and directional: a 128-slot Reader reads legacy 32-slot artifacts at -1.2 F1, one further generation costs -5.7, and older Readers do not read larger artifacts (-6.6). A hub therefore raises capacity by upgrading Readers first, one generation at a time, without invalidating a single stored artifact.

Table 34: **Development capacity and migration studies** (Llama-3.2-3B self-loops; development protocol). Left: self-loop F1 versus slot count against the member's full-context reference $.595$. Right: reading across slot formats; Δ is against the Reader's native-format reference.

Slots ($\times 1024$)	Self-loop F1
32	.590
64	.595
128	.609
256	.621

Artifact	Reader	ΔF1
32-slot	128-slot	-1.2
32-slot	256-slot	-5.7
128-slot	256-slot	$+3.6$
128-slot	32-slot	-6.6

H.4 Fidelity weight and the conformance rule

The per-member rule maximizes the smaller of self-loop F1 and frozen-judge read F1 (Table 35). It exists because the two can diverge sharply: at $\lambda_{\text{fid}} = .5$, Phi-4-mini's self-loop is healthy ($.474$) while the frozen judge reads its Writer at $.089$. Distilling a model's own text-conditioned distribution can pull its Writer toward a private dialect; the judge-read criterion catches this during development, and it doubles as a conformance test any prospective member can run against the public anchor before joining. The same rule keeps the smallest member interoperable: with the fidelity term the 0.5B Qwen2.5 Reader reads foreign artifacts at production level, whereas retrained without it the self-loop is unchanged ($.39$) and cross-model reading falls to $.05$.

Table 35: **Fidelity-weight selection on the 200 development examples.** Score is the minimum of self-loop F1 and frozen-judge-Writer read F1; the larger score selects the member's λ_{fid}.

Member	Self (.5)	Frozen (.5)	Score (.5)	Self (0)	Frozen (0)	Score (0)	Selected λ
Qwen-0.6B	.3223	.3220	.3220	.4180	.3639	.3639	.0
Qwen-1.7B	.4518	.3765	.3765	.4962	.4552	.4552	.0
Qwen2.5	.3164	.3152	.3152	.3300	.3073	.3073	.5
Llama-1B	.3799	.3330	.3330	.4505	.3882	.3882	.0
Llama-3B	.5303	.5309	.5303	.5638	.5178	.5178	.5
Phi-mini	.4738	.0890	.0890	.5129	.4795	.4795	.0
OLMo-1B	.2969	.3124	.2969	.4569	.4292	.4292	.0

I Resource accounting

The selected Writer–Reader adapter pairs span 51.1–115.2 million trainable parameters across the frozen backbones (Table 36); the two held-out members add 109.0 and 88.2 million. The nine

selected member runs total 29.87 H800-hours, a mean of 3.32 per member (Figure 7b); SmolLM2 onboarding costs 2.77 hours and Gemma-2-2B onboarding 5.35; the two replication hubs cost 63.7 hours for their eighteen members, 3.5 per member.

Table 36: **Trainable parameters of each selected member's combined Writer–Reader adapter** (millions). SmolLM2 and Gemma-2 are the held-out onboarding members.

Member	Role	Adapter parameters (M)
Qwen2.5-0.5B	Hub	51.1
Qwen3-0.6B	Hub	78.5
Qwen3-1.7B	Hub	88.0
Qwen3-8B	Hub	115.2
Llama-3.2-1B	Hub	66.9
Llama-3.2-3B	Hub	97.4
Llama-3.1-8B	Hub	111.0
Phi-4-mini	Hub	101.6
OLMo-2-1B	Hub	92.2
SmolLM2-1.7B	Held-out	109.0
Gemma-2-2B	Held-out	88.2

Figure 7a shows what the hub stores as it grows: one adapter pair per model against one integration per ordered pair for the pair-specific reference, so the eleventh member costs one pair where the reference would need twenty links. The bounding studies of Sections 4.2–4.5—full-context ceilings, the anchor factorial, the all-substitute anchor, the 18-pair C2C and trained text-communicator channels, the jointly trained shared space, the second onboarding, the 432 cross-hub directions, and the latency instrumentation—add roughly 550 H800-hours. The 18 sample-matched 4800-step C2C projectors alone account for about 300 of these, ten times the training of the entire nine-member hub.

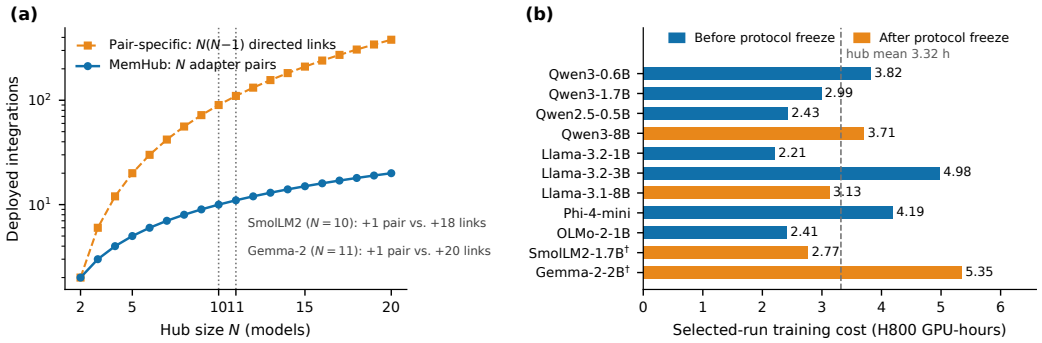


Figure 7: **Membership cost grows linearly and onboarding is cheap. (a)** Deployed integrations as the hub grows: MEMHUB stores one Writer–Reader pair per model, whereas the directional pair-specific reference stores one integration per ordered source–target pair (log scale); vertical lines mark the two measured onboarding events. **(b)** Measured H800 training hours of each member's selected run (nine hub members total 29.87 h, mean 3.32 h); orange marks runs trained after the evaluation protocol was frozen (the two 8B members and the two held-out families, marked †).

J BEYOND MULTI-HOP QA: CLASSIFICATION AND SUMMARIZATION

Every read in the main text is multi-hop QA. This appendix runs the same nine members, the same frozen Writers and artifacts, and the same statistics on two classification tasks and one summarization task. Each member's Reader is trained for the task on artifacts its own Writer has already written; nothing else changes. On all three tasks the Reader reads all eight foreign Writers zero-shot in all 72 directions, at 86–88% (classification) and 64% (summarization) of its own-Writer gain over a shuffled artifact, while the artifact of a different document from the same foreign Writer sits at the

label prior (Appendix J.2). One Reader per member, trained once on the three tasks and HotpotQA, serves all four within 2.4 points of the task-specific Readers (Appendix J.3). Without any Reader, a linear label probe fit on one Writer's artifacts reads every other Writer's at 90% retention, on the free slots as on the anchored, and fails on unanchored and judge-only members (Appendix J.4).

J.1 Protocol

Tasks. Table 37 lists the tasks and the verbatim Reader query. AG News (Zhang et al., 2015) and DBpedia-14 (Zhang et al., 2015; Lehmann et al., 2015) are 4- and 14-way classification of short articles, scored as accuracy after normalizing the greedy output and matching label aliases at word boundaries (no match counts as wrong); XSum (Narayan et al., 2018) is one-sentence summarization of BBC articles, scored by ROUGE-L (Lin, 2004). Decoding is greedy, at most 16 tokens for classification and 48 for XSum.

Conditions. *Self*, *ZS*, and *Empty* are as in Section 4.1. *Query-only* is the frozen backbone answering the query with no document and no adapter; *Full* is the frozen backbone reading the article text; *Shuffled* replaces the ZS artifact by the artifact of a different article from the same foreign Writer (a fixed derangement of the slice). Content-specific retention is $R_{\mathrm{shuf}} = (\mathrm{ZS} - \mathrm{Shuffled})/(\mathrm{Self} - \mathrm{Shuffled})$, reported alongside R_{self} (Query-only in place of Half); R_{shuf} is primary because two Llama Readers score 0.0 under Query-only.

Training. Member j writes 4,000 training articles once with its frozen Writer and trains its Reader on the cached artifacts against the task label: 400 steps, effective batch 32, AdamW at 5×10^{-5} with cosine decay and 20 warm-up steps, checkpoint selected on self-written artifacts of the pilot slice; the 27 runs total 1.02 H800-hours. The multi-task Reader of Appendix J.3 is trained the same way on the union of the three tasks and 4,000 HotpotQA examples for 800 steps, selected on the mean of the four development metrics.

Slices and tests. Each evaluation split is permuted once (seed 99); examples 0–199 form the pilot slice on which prompts, parsing, the learning rate, and checkpoints were fixed, and examples 1,000–1,999 form the confirmation slice (n=1000; 999 for XSum) reported below, evaluated once after the protocol was frozen. Intervals, McNemar tests, Holm correction over the 72 foreign directions of each task, and the reporting rule for retention follow Section 4.1. Directions with shared and mixed target chunking (44 and 28; Table 5) retain within one point of each other on every task; per-cell tables are released with the supplementary files.

Table 37: **Non-QA tasks and Reader queries.** Length is the median article length in words; under Full the article text precedes the same query.

Dataset	Task	Classes	Length	Reader query
AG News	topic classification	4	36	*Question: Which topic best describes the news article: World, Sports, Business, or Sci/Tech? Answer with a short phrase only.*
DBpedia-14	entity-type classification	14	42	*Question: Which category best describes the subject of the article: Company, Educational Institution, Artist, Athlete, Office Holder, Mean of Transportation, Building, Natural Place, Village, Animal, Plant, Album, Film, or Written Work? Answer with a short phrase only.*
XSum	one-sentence summarization	–	276	*Question: What is the article about? Summarize it in one sentence. Answer in one sentence.*

J.2 Foreign artifacts read zero-shot on three tasks

Table 38 reports the confirmation slice. On DBpedia-14 the Reader reads a foreign Writer's artifact at 80.8% accuracy against 96.1% from its own Writer, while the Shuffled artifact from the same

foreign Writer scores 8.1%, near the 7.1 prior; on AG News the numbers are 79.8, 87.7, and 26.1 against a 4-way prior of 25. Every one of the 72 foreign directions on every task has a Holm-corrected gain over Query-only and a positive paired interval over Shuffled, and Empty sits at the prior. Median content-specific retention is 88.3% on AG News and 86.2% on DBpedia-14; the median gap to Self is -7.3 and -12.2 points. The foreign Writers were trained for QA only, so this gap is that of reading a Writer that never saw the task; in QA, where every Writer was trained on the task, ZS trails Self by 0.3 F1 (Table 1a). On AG News the foreign artifact reads at the level of the frozen backbone given the full article text (median ZS$-$Full $+2.4$; 35 directions positive, 12 negative) and on DBpedia-14 1.7 points below it.

Table 38: **Foreign artifacts read zero-shot on three non-QA tasks.** Confirmation slice, n=1000 per direction (999 for XSum), nine Readers, 72 foreign directions per task; accuracy (%) for AG News and DBpedia-14, ROUGE-L (%) for XSum. Query-only, Empty, Full, and Self are Reader means; ZS and Shuffled average the eight foreign Writers. R_{shuf}, R_{self}, and ZS$-$Self are medians over the 72 directions; H^+ counts directions with a Holm-corrected gain over Query-only and $>$Shuf. those whose paired interval over Shuffled is positive.

	References			Memory reads			Retention (%)			Directions	
Task	Query-only	Empty	Full	Self	ZS	Shuffled	R_{shuf}	R_{self}	ZS$-$Self	H^+	$>$Shuf.
AG News	19.0	24.6	73.3	87.7	79.8	26.1	88.3	90.2	-7.3	72/72	72/72
DBpedia-14	4.9	6.8	82.5	96.1	80.8	8.1	86.2	86.6	-12.2	72/72	72/72
XSum	7.1	11.3	17.0	14.8	13.1	10.4	64.1	81.4	-1.5	72/72	72/72

On XSum the foreign artifact yields ROUGE-L 13.1 against 14.8 from the Reader's own Writer and 10.4 from the Shuffled artifact, positive over Shuffled in all 72 directions with median content-specific retention of 64.1%. The frozen backbone reading the full article reaches 17.0, and a 32-slot artifact written from a 276-word article is in the capacity regime that Table 34 measures for QA (one point per doubling of the slot count).

Table 39: **Per-Reader view of Table 38.** Self and ZS (mean over eight foreign Writers) in accuracy or ROUGE-L (%); R_{shuf} is the median over the Reader's eight foreign directions. Every Reader is significant in 8 of 8 directions under both tests on every task.

	AG News			DBpedia-14			XSum		
Reader	Self	ZS	R_{shuf}	Self	ZS	R_{shuf}	Self	ZS	R_{shuf}
Qwen3-0.6B	86.3	79.5	88.2	97.0	76.8	82.8	14.0	12.7	65.2
Qwen3-1.7B	87.9	81.8	91.6	97.2	80.6	85.8	14.3	13.1	67.1
Qwen2.5-0.5B	85.6	77.7	88.3	95.1	65.5	67.1	14.5	12.8	53.9
Qwen3-8B	88.5	83.3	91.6	94.0	86.4	90.0	15.0	13.8	75.3
Llama-3.2-1B	88.3	79.8	91.4	95.9	80.7	85.2	15.1	13.1	53.8
Llama-3.2-3B	87.4	78.5	86.8	95.4	82.3	85.1	14.7	12.9	64.2
Llama-3.1-8B	88.9	76.6	84.1	96.6	87.9	89.7	14.6	13.4	75.5
Phi-4-mini	88.6	80.0	87.1	96.8	83.9	86.5	14.9	13.0	60.6
OLMo-2-1B	87.8	81.2	91.2	97.0	83.5	87.5	16.1	13.0	50.7
All	**87.7**	**79.8**	**88.3**	**96.1**	**80.8**	**86.2**	**14.8**	**13.1**	**64.1**

Table 39 shows the result Reader by Reader. Every Reader, including the 0.5B Qwen2.5 that gains least on HotpotQA (Table 1a), reads all eight foreign Writers significantly on all three tasks; on classification, retention is lowest for Qwen2.5 on DBpedia-14 (67.1%) and highest for the two 8B Readers on DBpedia-14 and XSum.

J.3 One Reader per member serves four tasks

Table 40 trains a single Reader per member on the mixture of AG News, DBpedia-14, XSum, and HotpotQA and reads foreign artifacts on all four. The multi-task Reader stays within 2.0 points of the task-specific Readers on the three new tasks (ZS 77.8, 79.6, and 13.0 against 79.8, 80.8, and

13.1), with Holm-corrected gains over Query-only in 72 of 72 directions and a positive interval over Shuffled in 72, 72, and 71 of 72, and on the locked HotpotQA slice it reads self and foreign artifacts at 48.6 and 47.5 F1 against 49.7 and 49.8 for the original QA Reader. Extending the hub to a task is one Reader training run per member, after which the new readout applies to every artifact any member has written or will write.

Table 40: **One multi-task Reader per member reads foreign artifacts on four tasks.** **(a)** Non-QA tasks, format of Table 38: the task-specific Readers of Table 38 against one Reader per member trained on the four-task mixture. **(b)** HotpotQA on the first 300 examples of the locked slice, F1 (%) in the columns of Table 1a.

(a) Non-QA tasks: task-specific versus multi-task Reader

Task	Reader	Self	ZS	Shuffled	R_{shuf}	H^+	>Shuf.
AG News	task-specific	87.7	79.8	26.1	88.3	72/72	72/72
	multi-task	87.3	77.8	25.7	87.0	72/72	72/72
DBpedia-14	task-specific	96.1	80.8	8.1	86.2	72/72	72/72
	multi-task	95.2	79.6	8.0	85.2	72/72	72/72
XSum	task-specific	14.8	13.1	10.4	64.1	72/72	72/72
	multi-task	14.8	13.0	10.3	64.6	72/72	71/72

(b) HotpotQA, locked slice (n=300, 72 foreign directions)

Reader	Half	Full	Self	ZS
Original QA Reader (Table 1a)	37.5	49.0	49.7	49.8
Multi-task Reader	37.5	49.0	48.6	47.5

J.4 The label is linearly readable across Writers without a Reader

Table 41: **A linear label probe fit on one Writer reads the other eight.** Confirmation slice, accuracy (%). **(a)** Probe on all 32, the 16 anchored, or the 16 free slots: Same is the median over Writers, Cross the median and minimum over the 72 foreign directions, R the median retention. **(b)** Factorial controls: Self is the control's probe on its own artifacts, →Hub that probe on the eight hub Writers, Hub→ the hub probes on the control (medians); the label prior is 25.0 and 7.1.

(a) Cross-Writer probe transfer by slot group

	AG News				DBpedia-14			
Slots	Same	Cross	Min	R	Same	Cross	Min	R
All 32	88.0	82.2	72.2	90.5	96.6	88.3	75.0	90.4
16 anchored	87.8	82.8	72.2	92.4	96.2	88.7	78.6	91.9
16 free	87.5	81.8	70.8	90.9	95.7	84.3	61.7	86.7

(b) Factorial controls, all 32 slots

		AG News			DBpedia-14		
Condition	Member	Self	→Hub	Hub→	Self	→Hub	Hub→
Neither	Q06	81.1	24.1	24.5	87.4	8.1	6.6
	L1	82.6	24.5	25.3	85.8	6.6	7.4
Geometric only	Q06	87.5	74.3	79.6	95.9	89.0	84.3
	L1	88.3	80.5	81.6	96.5	92.5	86.4
Judge only	Q06	78.6	26.1	25.3	82.5	8.2	7.3
	L1	82.5	24.5	26.4	83.8	8.1	7.8

A linear probe removes the Reader. For each Writer i, a multinomial logistic regression is fit to the label on the artifacts of its 4,000 training articles (features centered on i's mean, one global scale, ℓ_2 penalty 1) and applied unchanged to every other Writer's confirmation-slice artifacts; retention is (cross−prior)/(same−prior) with the label prior (25.0 and 7.1); XSum has no label. The controls are the Q06/L1 factorial checkpoints of Table 3a (neither constraint, geometric only, judge only), probed in both directions against the hub.

A probe fit on one Writer reads the other eight at a median of 82.2% on AG News and 88.3% on DBpedia-14 against 88.0 and 96.6 on its own Writer (retention 90.5 and 90.4, Table 41a; weakest direction 72.2 and 75.0, Table 42). The 16 free slots alone transfer nearly as well as the 16 anchored (81.8 against 82.8 on AG News) although nothing pins them to a target: the label lies in directions the free slots share, as Proposition 2 requires when one basis change acts on all 32 slots. Raising the penalty from 0.01 to 1 lifts cross-Writer accuracy on AG News from 64.1 to 82.2 at unchanged same-Writer accuracy: the directions a probe sheds are member-specific, the ones it keeps are shared.

The controls (Table 41b) separate the two anchor terms as Table 3a does. An unanchored member reads its own artifacts linearly at 81–87%, but between it and the hub the probes sit at the prior (24.3 and 7.3 toward the hub, 24.9 and 7.0 from it): the label is encoded in a private basis. Judge only is the same (25.3 and 8.1); geometric only transfers at hub level (77.4 and 90.7). Judge pinning makes an artifact readable to one fixed nonlinear Reader (Remark 2); geometric pinning makes it readable to a linear map fit on any other member, on the free slots as on the anchored (their .53 raw retrieval in Table 26 is the stricter, instance-level test). It is also why the task-adapted Readers of Table 38 transfer: the directions a Reader learns on its own Writer's artifacts already carry the label on every other Writer's.

Table 42: **Per-cell probe transfer** (all 32 slots, penalty 1, n=1000): accuracy (%) of the probe fit on Writer i (row) on the artifacts of tested Writer j (column); the diagonal is same-Writer.

Probe i	Q06	Q17	Q25	Q8	L1	L3	L8	Phi	OLMo
AG News									
Q06	**87.6**	86.1	79.5	84.0	77.9	84.4	82.4	85.2	80.2
Q17	79.4	**88.0**	76.3	81.8	76.4	78.6	84.1	78.1	81.5
Q25	79.4	82.4	**87.4**	82.7	79.9	82.6	84.6	85.4	79.9
Q8	75.9	84.3	82.8	**88.0**	73.3	81.8	84.9	82.1	81.6
L1	85.1	87.2	81.6	84.5	**88.4**	78.6	83.9	84.9	83.2
L3	72.2	80.9	75.6	82.2	78.0	**87.3**	86.2	86.3	81.4
L8	82.2	83.9	80.8	84.5	76.9	78.9	**88.8**	83.2	79.2
Phi	82.2	84.0	81.8	83.7	73.0	84.8	85.4	**88.1**	81.7
OLMo	83.5	86.6	81.2	79.7	84.5	77.8	80.2	83.6	**88.5**
DBpedia-14									
Q06	**97.0**	95.1	88.1	87.0	85.2	80.6	79.1	88.4	92.7
Q17	91.2	**97.4**	88.4	89.5	90.6	80.9	82.5	93.1	94.3
Q25	92.3	90.0	**95.5**	89.1	86.8	78.2	83.6	89.0	90.7
Q8	94.8	95.5	88.7	**96.6**	88.2	86.8	87.8	90.6	92.8
L1	89.8	93.7	81.3	82.8	**96.2**	76.1	77.5	92.4	92.2
L3	89.4	89.7	82.2	82.8	79.3	**95.0**	84.6	90.8	86.5
L8	89.8	89.5	82.2	88.8	83.9	78.8	**95.4**	88.7	86.5
Phi	93.0	94.9	86.0	80.9	84.5	81.8	77.3	**96.8**	91.2
OLMo	93.6	95.5	85.7	79.7	86.7	77.6	75.0	91.8	**96.9**

K SCOPE OF THE INSTANTIATION

The hub studied here is one instantiation of the interface: eleven frozen backbones from seven families at 0.5B–8B, a 32×1024 slot format, and an anchor built from one embedder and one judge. Four properties of the results locate the construction beyond that instantiation. The format has measured headroom and a measured migration path—one F1 point per doubling of slot count, upgraded Readers reading legacy artifacts at -1.2 F1 (Table 34)—so capacity can be raised without invalidating stored memory. And every coordination object is public and replaceable: the embedder, the judge, and the training corpus can each be swapped for a different one with transfer unchanged (Tables 27, 29), so a hub is defined by whichever anchor its members agree to publish. The stored artifact itself is inert with respect to evidence the Reader can see—a planted decoy flips none of 140 correct answers to it (Appendix E.1)—which is the property a persistent memory shared across models has to have. And the artifact is not a QA-specific code: Readers trained for classification and summarization on their own Writers' artifacts read foreign artifacts zero-shot, one Reader per member serves four tasks, and a linear probe fit on one Writer reads the label from every other (Appendix J).