AUGMENT THE CACHE, REPLACE THE ACTIVATION: A 2X2 STUDY OF INTERNAL-STATE SHARING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent systems built from frozen language models communicate through text, which discards the internal state the sender already computed and makes the receiver pay to rebuild it. Work on internal channels avoids that cost, framing the problem as *transfer*: agent A sends information to agent B. Every agent pair that must communicate needs its own transfer bridge, and every agent added to a system needs a new bridge fitted against each previous agent. We reframe this problem as sharing: an agent publishes its internal state once into a pool, and every peer reads that pool through a single mechanism fitted to no sender in particular, so adding an agent adds a publisher and no new bridge. Whereas the carrier of the internal state and the fusion mechanism are usually chosen independently in the literature, our main finding is that these two choices are coupled; the deciding factor is where the carrier enters the receiver's computation. Consumability is a property of the receiving computation rather than of the representation. We find that the key–value (KV) cache must be augmented rather than swapped in: the receiver cross-attends over a peer's cache while keeping its own, and with only a small query-side adapter trained, matches the accuracy it gets from reading the source document. Activations invert this. They are best written straight into the receiver, with no bridge, no adapter and no training, which beats exposing them alongside the receiver's own, though that advantage rests on knowing where in the receiver the state belongs. The rule is *augment the KV cache, replace the activation*, established under one protocol with a shared control ladder. One adapter, trained once, also serves publishers it never trained on, and a pool of several at once, but only those whose internal state lies close to the receiver's own.

1 INTRODUCTION AND RELATED WORK

A multi-agentic language-model system orchestrates several frozen models to plan, criticize, retrieve and divide labor, and has become a dominant pattern for assembling capable behavior from individually insufficient components (Wang et al., 2024; Du et al., 2023; Wu et al., 2024; Li et al., 2023a; Hong et al., 2024; Xi et al., 2023). Agents in every deployed system of this kind exchange information through one channel: *natural-language text*. Text is the interface because it is the only representation every model agrees on. It is also the model's *output*, a lossy projection of a much richer internal state, and whatever did not survive that projection into words is unavailable downstream.

Internal information sharing is any mechanism conveying what a model computed without requiring that model's outputs, and two objects inside a transformer are the natural carriers. The **KV cache** is the attention memory a transformer (Vaswani et al., 2017) builds over a context and consults while generating; the **residual-stream activations** are the hidden states $\mathbf{h}_\ell$ flowing between blocks. Both are computed during reading, both hold the sender's digestion of the context, and neither is an output.

Sharing, not transfer. The pairwise mechanisms of prior work on internal channels (Fu et al., 2026; Dery et al., 2026; Peng et al., 2026) can be tuned to a known sender; the shared pool we study cannot, which is the stronger requirement. The appendix tests it against a pool of eight holders whose answer-bearer is unknown (Appendix A.1) and against senders never trained on (Appendix A.2).

Carrier-level channels. The KV cache is the object the field reached for first, whether as a cross-model exchange channel (Fu et al., 2026; Dery et al., 2026; Liu et al., 2024c; Ye et al., 2025; Bian et al., 2026), as a *private* object to be shrunk (Zhang et al., 2023; Liu et al., 2023; Xiao et al., 2024;

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.

Li et al., 2024; Liu et al., 2024a; Sun et al., 2024; Hooper et al., 2024; Liu et al., 2024d), or as something to reuse across requests of the *same* model (Kwon et al., 2023; Gim et al., 2024; Jin et al., 2024). The latter two families perturb a cache the receiver itself would have produced, which makes cross-*model* reuse a different problem, and one that repeatedly fails without training. Writing into the residual stream is the parallel literature, where much of a model's behavior turns out to live in a few activation directions one can add (Todd et al., 2024; Liu et al., 2024b; Zou et al., 2023; Li et al., 2023b; Panickssery et al., 2024; Turner et al., 2023; Subramani et al., 2022; Ilharco et al., 2023; Wortsman et al., 2022; Yadav et al., 2023; Wan et al., 2024). That line establishes that *behavior* is portable. We ask whether *content* is. A mature literature on representational correspondence (Kornblith et al., 2019; Raghu et al., 2017; Moschella et al., 2023; Bansal et al., 2021; Lenc & Vedaldi, 2015; Ainsworth et al., 2023; Dravid et al., 2023; Huh et al., 2024; Jha et al., 2025) carries one caveat throughout: high representational *similarity* does not imply *interchangeability*. Work that adds information through attention sidesteps the question (Borgeaud et al., 2022; Izacard & Grave, 2021; Alayrac et al., 2022; Jaegle et al., 2021; Hu et al., 2022). In our taxonomy these are all augmentation, and their empirical success is a clue we take seriously.

Latent communication between agents. Closest to our framing is a line asking whether agents can exchange latent state in place of text. A survey of eighteen such methods organizes them on an axis of carriers (embeddings, hidden states, KV caches) matching ours, and concludes that a method's carrier and its fusion mechanism "can be chosen largely independently" (Liu, 2026). That conclusion is the one this paper contradicts, and the survey's own tabulation pairs the two opposite to our rule on both carriers. Trainable latent channels report gains on reasoning benchmarks (Yu et al., 2026), but two results temper the optimism: a controlled study finds the latent channel matches yet never exceeds text (Wenzel, 2026), and a causal audit shows aggregate accuracy does not identify the latent message's contribution, so agents can appear to communicate while the channel carries little (Zhang & Emu, 2026). Our control ladder (Section 3) exists because of the second. StateBridge is the closest positive cross-model result, aligning last-layer hidden states across *different* models with a training-free orthogonal map (Peng et al., 2026). Read our cross-family collapses against it: StateBridge hands *last-layer* states to a downstream consumer, while we inject at *intermediate* layers the receiver's remaining stack must consume.

1.1 This paper

We study both carriers under one controlled protocol, and Figure 1 states the result in one picture. Our contributions are as follows.

(1) A receiver-centric frame. We define a receiver's *readout map* and the *tolerable set* it induces (Section 2), separating two regimes the literature routinely conflates: under *substitution* a shared object replaces the receiver's own, under *augmentation* the receiver keeps its own and the shared object is exposed alongside it. That distinction is what makes the KV and activation results commensurable.

(2) A mechanism for why KV substitution fails at high fidelity. The asymmetry follows from where each object is consumed (Section 2): keys and values are read *by* the softmax of Eq. 2, a normalized competitive map, while activations are read before any softmax through a near-unit-gain path. Our measured readout gain falsifies the conditioning cliff such an argument invites, so the bound is not a matter of numerical sensitivity (Section 4.1).

(3) The scope of the substitution bound. It reproduces on seven receivers from three families (Appendix A.3), and it holds in the most favorable case physically available, where the substituted key is the nearest one the receiver ever produced itself (Section 4.1). Every cross-family *sharing* configuration we test, on either carrier, collapses to the floor.

(4) A 2x2 grid evaluation of carrier (KV cache, activation)-mechanism (replace, augment). KV augmentation clears the floor before anything is trained, so a usable channel exists without fitting an alignment map at all (Section 4.1). Activation replacement needs no training whatsoever, but it requires oracle position and layer alignment that our same-base setting supplies and a deployed system would have to establish (Section 4.3).

Our code is available at `https://anonymous.4open.science/r/atcrta2x2`.

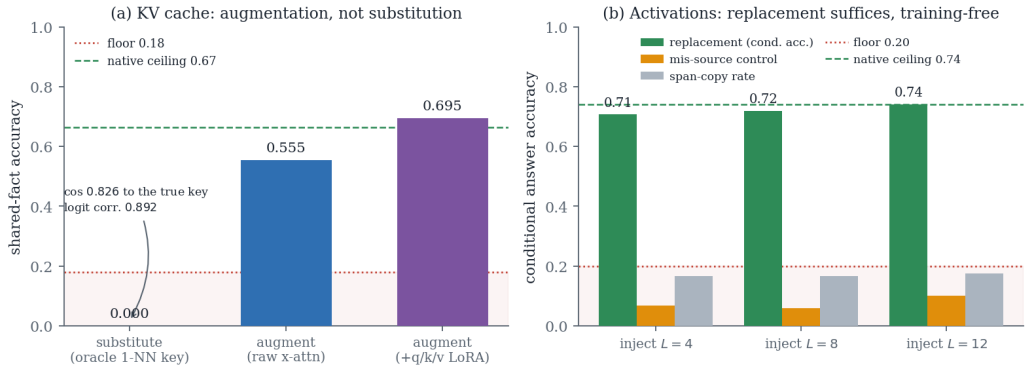


Figure 1: The two carriers take opposite mechanisms. **(a)** The KV cache cannot be substituted, even at the fidelity annotated on the bar, and must be augmented instead. **(b)** Activations invert this: direct replacement works, training-free. Floors and native references are per carrier.

2 A RECEIVER-CENTRIC ACCOUNT OF CONSUMABILITY

This section fixes the objects and defines the two regimes. It then gives the conditioning argument that predicts the asymmetry we go on to measure.

2.1 RECEIVERS, CARRIERS, AND READOUT MAPS

Fix a frozen transformer M with L layers, hidden width d, and grouped-query attention (Ainslie et al., 2023) with N_Q query heads over N_{KV} key/value heads in groups of $G = N_Q/N_{KV}$, head dimension d_h. Frozen means its weights, and so every function it applies, stay fixed throughout. Nothing below fine-tunes the receiver.

Running M on a context x produces, at every layer ℓ and position t, a residual-stream activation $\mathbf{h}_\ell(t) \in \mathbb{R}^d$ and the derived attention memory

$$\mathbf{K}_\ell(t) = \mathrm{RoPE}\Big(W_K^{(\ell)}\,\mathrm{LN}(\mathbf{h}_\ell(t))\Big), \qquad \mathbf{V}_\ell(t) = W_V^{(\ell)}\,\mathrm{LN}(\mathbf{h}_\ell(t)).$$

A *carrier* is the object a sharing mechanism transports. We consider three: the activation $\mathbf{h}$, and the cache components $\mathbf{K}$ and $\mathbf{V}$. The dependency matters. $\mathbf{K}$ and $\mathbf{V}$ are functions of $\mathbf{h}$ through the *producer's own* projections, which welds the cache to the geometry of the model that produced it in a way the activation is not.

Let $\mathcal{C}$ be the space of caches. The receiver's *readout map* $\mathcal{R} : \mathcal{C} \to \Delta(V)$ sends a cache C to the next-token distribution M produces when its attention consumes C: the composition of the attention-over-C operation, the residual stream, the remaining blocks, the final norm and the unembedding, all with M's frozen weights. $\mathcal{R}$ is *receiver-specific*: the same C fed to two receivers passes through two different maps. There is no reason $\mathcal{R}(C)$ should mean anything to a receiver other than the one whose forward pass $\mathcal{R}$ encodes.

One cache stands apart from the rest: the *native cache* $C^\star$, which M itself produces on the true context. $\mathcal{R}(C^\star)$ is the receiver reading its own work, and it defines the *native ceiling*.

For a behavioral divergence D on next-token distributions and a tolerance ϵ,

$$\mathcal{T}(M, \epsilon) \;=\; \big\{\, C \in \mathcal{C} \;:\; D\big(\mathcal{R}(C), \mathcal{R}(C^\star)\big) \le \epsilon \,\big\}. \tag{1}$$

$\mathcal{T}$ is the set of caches the receiver consumes behaviorally-equivalently to its own.

2.2 TWO REGIMES: SUBSTITUTION AND AUGMENTATION

The literature on cache sharing mixes two mechanisms that this frame separates cleanly. That separation carries most of the explanatory weight in what follows.

A *substitution* mechanism is a map g producing an object $g(\text{source})$ that is placed *in the receiver's own slots*, so the receiver attends over $g(\text{source})$ *instead of* its own state. Reconstruction, selection from a bank, quantization, and direct overwrite are all substitution. Its behavior is $\mathcal{R}(g(\text{source}))$, so Eq. 1 governs it: substitution succeeds only if g's image intersects $\mathcal{T}$.

An *augmentation* mechanism leaves the receiver's own state intact and exposes the source additionally: the receiver attends over the concatenation $[C^\star; C_{\text{src}}]$, or receives the source as an additive residual contribution. Here the receiver never applies $\mathcal{R}$ to a foreign point, so Eq. 1 does not bound it. The receiver now has strictly more to attend over than its own context alone, and augmentation can in principle exceed $\mathcal{R}(C^\star)$.

2.3 Why the carriers differ: the consumption point

What decides the outcome is *where* each carrier enters the receiver's computation.

An activation is consumed by the next block through $\mathbf{h} \mapsto \mathbf{h} + f(\mathrm{LN}(\mathbf{h}))$: a normalization, a linear map, an *additive residual*. Gain along that path is $O(1)$, so a displacement δ passes through essentially undistorted while the normalization re-projects toward the operating manifold at every layer.

A cache is consumed by exactly one operation,

$$\mathrm{Attn}(q, \mathbf{K}, \mathbf{V}) \;=\; \mathrm{softmax}\!\left(\frac{q^\top \mathbf{K}}{\sqrt{d_h}}\right)\mathbf{V}, \tag{2}$$

and the softmax is *normalized and competitive*: it scores each key against every other key, concentrating weight on whichever keys lead. A displacement leaving a key almost exactly where it was in Euclidean or cosine terms can still change the *ordering* of the $q^\top k$ scores, and once the ordering changes, weight moves to different tokens. Consumability turns on whether δ disturbs the narrow margins by which the true keys presently win. Some perturbations disturb none, so $\mathcal{T}$ is not literally zero-dimensional: a common shift in all logits, or a displacement confined to keys the query never ranks highly, leaves the readout intact. Our claim is weaker and empirical. For the substitutions we can construct, the usable neighborhood around $C^\star$ is narrow enough that no mechanism we test finds a point inside it, which we read as measured tolerance rather than the geometry of $\mathcal{T}$.

Two consequences follow. Both run against intuition, and we test both.

Proposition 1 (Fidelity does not imply consumability)**.** *Let $\phi(C)$ be any geometric fidelity of C to $C^\star$ (cosine of the key/value tensors, MSE, or correlation of induced attention logits). High ϕ does not imply $C \in \mathcal{T}$. Cosine measures Euclidean alignment while the softmax depends only on the* ranking *of $q^\top k$, and a cosine-invisible rotation can reorder that ranking. A substitution map may drive ϕ arbitrarily close to its maximum while $D(\mathcal{R}(C), \mathcal{R}(C^\star))$ stays pinned at its worst.*

Proposition 2 (The wall is not amplification, and not depth)**.** *Write κ_{off} and κ_{on} for the readout Jacobian gain under off- and on-manifold perturbation. A conditioning cliff would require $\kappa_{\text{off}} \gg \kappa_{\text{on}}$; depth-compounding would kill a carrier traversing the same L layers alongside the cache. Neither holds (Appendix A.5): the measured ratio is close to unity, and activations injected into the identical L=28 receiver transfer while caches do not. The wall belongs to the readout a carrier feeds rather than the depth it is read through.*

One analytic bound sharpens the GQA dependence. Because G query heads share each key/value slot, a perturbation there is read out through all G queries and the per-slot sensitivity constant scales as $\sqrt{2G}$; for G=6 this is $\sqrt{12} \approx 3.46$. The constant belongs to attention *shape* rather than the weights, and at $O(1)$ it is far too small to build a wall by amplification (Proposition 2).

2.4 The Asymmetry Conjecture

The frame yields a falsifiable prediction, and that prediction organizes the rest of the paper.

Conjecture 1 (Asymmetry of internal information sharing)**.** *For a frozen receiver,*

- *(C1)* ***The cache must be augmented.*** *Because the usable neighborhood around $C^\star$ is narrow, any substitution of keys and values is bounded near the transfer floor, whatever fidelity ϕ it achieves. Augmentation never asks the receiver to consume a foreign cache in place of $C^\star$, so $\mathcal{T}$ does not bound it, and it can approach or exceed the native ceiling.*

(C2) ***The activation may be replaced.*** *Because the activation is consumed through a near-unit-gain pre-softmax path, its tolerable region is a genuine neighborhood rather than a point, and direct replacement transfers. Augmenting the same carrier only tilts the receiver's own state instead of installing the holder's, and it transfers* less.

The conjecture is falsified if a cache substitution transfers well above the floor while provably outside $\mathcal{T}$; if cache augmentation with query-side adaptation fails to clear the floor; or if activation replacement fails while activation augmentation succeeds.

3 METHODOLOGY

Models. Unless noted, holder and receiver are the same frozen Qwen2.5-1.5B (Qwen Team, 2024) with L=28 layers, hidden width 1536, and grouped-query attention in groups of G=6. The cross-family holder is GPT-Neo-1.3B, mapped into the receiver's key space by a frozen per-layer SVD projection. Weights are frozen in every experiment; only bridge parameters, where present, are trained. Appendix A.4 gives the full geometry of both models.

The split-context fact-transfer probe. Every number in this paper comes from one instrument. An item is a SQuAD (Rajpurkar et al., 2016) passage whose answer appears verbatim and runs to at most four words, paired with its question, in two versions: *full*, the original passage, and *masked*, the passage with the answer span replaced by _____. The *holder* processes the full passage. The frozen *receiver* gets the masked passage, so it cannot answer from its own context. We then supply the receiver an object derived from the holder ($\mathbf{K}$, $\mathbf{V}$, or $\mathbf{h}$) and ask whether it can now answer. Prompts are `{passage}\nQuestion: {q}\nAnswer:`, truncated to 256 tokens, with 8 greedy answer tokens. Evaluation items come from the validation split, training items from train.

Anchors. Two per-receiver anchors bound every measurement. The *floor* is the receiver on the masked passage with no useful injection: it knows the question and has a fact-free context. The *native reference* is the receiver on the full passage reading its own state, the identity case $\mathcal{R}(C^\star)$. *The KV-side and activation-side probes carry different anchors and must not be compared in absolute terms.* KV-side runs score floor 0.18 / ceiling 0.665 at n=200, activation-side runs 0.20 / 0.742 at n=120. We report position within the floor–ceiling band and plot both anchors on every panel.

Metrics. Generation is scored by normalized substring match: the gold answer, lowercased and stripped of punctuation, must appear in the generation. We report **shared-fact accuracy** over all items. Because a receiver might answer by copying a visible span, we also report **conditional answer accuracy**, restricted to items the receiver answers correctly at its ceiling, beside an explicit **span-copy rate** (the answer shares a two-token contiguous bigram with the injected span *and* was produced). A transfer claim is credited only when shared accuracy clears the floor *and* conditional accuracy exceeds span-copy by a wide margin.

Controls. Four, of increasing strength, defined in Appendix A.4: a *no-op* injecting the receiver's own masked state at the same sites, a *position* control at the fact-free prefix, a *mis-source* control supplying a different item's state, and an *answer-disjoint split* for trained bridges. The last two separate genuine routing from a mechanism emitting memorised answers.

Mechanisms compared. On the KV side, *substitution* places produced or selected keys in the receiver's own slots, while *augmentation* keeps the receiver's cache and cross-attends over the concatenation with the holder's, used verbatim. Only the receiver's own projections adapt, by low-rank adapters that start at exactly the identity (Hu et al., 2022). On the activation side, *replacement* overwrites $\mathbf{h}_\ell$ at the fact region, training-free. One caveat belongs here and not later. The fact region comes from diffing the masked token sequence against the full one, and the injection layer is chosen from a swept set, so position and depth are both supplied to the mechanism. That is an oracle, and it makes the activation arm a measurement of consumability under correct alignment and not a deployable interface (Appendix A.4). Activation *augmentation* instead adds a routed convex combination of source activations to the residual stream.

Verification. Attention is re-implemented manually to permit injection, so every run must first reproduce native greedy decoding string-exactly and rebuild the native cache to within tolerance before its numbers are read (Appendix A.4).

	α=0.0	0.1	0.25	0.5	0.75	1.0
k=1	0.180	0.185	0.110	0.005	0.000	0.000
k=2	0.180	0.180	0.110	0.020	0.010	0.005
k=4	0.180	0.205	0.120	0.010	0.005	0.000
k=8	0.180	0.175	0.120	0.005	0.000	0.000
k=16	0.180	0.145	0.135	0.020	0.015	0.010

Table 1: Oracle substitution: the receiver's key is interpolated toward the k-th nearest key the receiver itself produced elsewhere, so every candidate is on its own manifold. No neighbor rank recovers a graded tolerance. Qwen2.5-1.5B, n=200, floor 0.18, ceiling 0.665.

Setting	Adapted	Shared	Control	Δ vs. raw	vs. ceiling
same model	none (raw)	0.555	0.185	—	−0.110
	q	0.600	0.155	+0.045	−0.065
	k, v	0.640	0.090	+0.085	−0.025
	q, k, v	**0.695**	0.200	+0.140	+0.030
cross family	none (raw)	0.010	0.005	—	—
	q	0.010	0.005	+0.000	—
	k, v	0.080	0.085	+0.070	—
	q, k, v	0.040	0.035	+0.030	—

Table 2: KV augmentation. The receiver keeps its own cache and cross-attends over the holder's verbatim cache; only the receiver's projections adapt. The control feeds a holder cache built on the answer-masked passage. Same-model, n=200.

4 Experimental Results and Analysis

4.1 The KV cache: substitution is floor-bounded

We first establish the bound of Conjecture 1(C1) under conditions engineered to be maximally favorable to substitution. *Synthesizing* a key risks landing off-manifold, so instead we *select* one from a bank of 6000 genuine keys the receiver itself produced on other contexts. Every candidate is then on the receiver's own key manifold by construction. For a target slot we take the k-th nearest neighbor of the true key and substitute it.

The result is categorical. Full substitution by the nearest genuine key transfers at 0.000. The tolerance is flat in neighbor rank (k=1, . . . , 16 all $\leq$ 0.01), so there is no graded neighborhood to exploit. Interpolating from the true key toward the neighbor craters: every cell at mixing weight $\alpha \geq 0.5$ sits at or below 0.02 (Table 4.1). The *identity* arm, which hands the receiver its own exact keys, reproduces the ceiling 0.665 bit-exactly, with cosine 1.000 and all 28 layers clearing the fidelity threshold. So the correct point exists, is on-manifold, and is fully consumable. The *nearest available miss* already returns nothing.

This is Proposition 1 in its sharpest form. The bank-selected key attains cosine 0.826 and attention-logit correlation 0.892 to the true key, which is high fidelity by any conventional proxy, and it transfers nothing. A broader sweep of substitution mechanisms (freeform regression, manifold-matched training objectives, column- and row-space construction, worst-case hinge losses, high-capacity per-layer MLPs, and codebook selection) spans transfer 0.00 to 0.20 against the same floor of 0.18. None clears it. The best-fidelity member of the family is also not the best-transferring one.

4.2 The KV cache: augmentation clears the ceiling

We change only the mechanism. The receiver keeps its own real cache and cross-attends over the concatenation with the holder's verbatim cache; only query-side parameters adapt. Table 4.2 reports the ladder.

Inject layer	Shared fact	Cond. acc.	Span-copy	Mismatch ctrl	No-op ctrl
L=4	0.533	0.708	0.167	0.067	0.200
L=8	0.550	0.719	0.167	0.058	0.200
L=12	0.567	**0.742**	0.175	0.100	0.200
L=16	0.567	0.708	0.167	0.408	0.200
L=20	0.575	0.708	0.183	0.425	0.200
all of the above	0.517	0.685	0.142	0.067	0.200

Table 3: Activation replacement, training-free. The receiver's hidden states at the fact region are overwritten with the holder's. Qwen2.5-1.5B, n=120 (cond n=89), floor 0.20, ceiling 0.742. The mismatch control is clean at $L \leq 12$ but not at $L \geq 16$ (shaded reading in the text).

We have three readings. First, *augmentation escapes the substitution bound immediately*: the untrained bridge, with zero trainable parameters, already lands several times clear of the floor, which no substitution mechanism we tested could approach. The mechanism does the work here, not the optimization. Second, the adaptation ladder is monotone, $q, k, v \succ k, v \succ q \succ$ raw, so adapting the receiver's own projections is what converts available signal into answers. Third, the best arm meets the receiver's native reference, and we are deliberately careful about what that licenses. At n=200 the binomial standard error on the difference is 0.047, so 0.695 versus 0.665 is $z \approx 0.6$ and *not* statistically distinguishable from a tie. The defensible claim is that query-side-adapted augmentation *reaches* the native reference; we cannot claim it provably exceeds it, since establishing that needs multiple seeds we have not run.

Is this lexical retrieval? The holder's cache contains the answer passage verbatim and the receiver cross-attends over it, so a high score could in principle come from locating and copying answer tokens instead of consuming the holder's reading. We close this by porting the hardened activation-side ladder across, on an answer-disjoint evaluation where no evaluation answer appears in the bridge's training data. Conditional accuracy clears span-copy by 0.519 ($z \approx 11$), the reverse of what contamination predicts, and the mis-source arm, handed a *different* item's answer-bearing passage, sits below floor. Widening the pool sharpens this: span-copy *falls* while conditional accuracy persists, though more distractors would give a copying mechanism more to copy. Two caveats. Copying is not eliminated, only shown insufficient as an explanation, since a span-copy rate of 0.366 at N=1 means some answers do arrive as lifted spans. And the mis-source arm, clean at N=1, rises above floor in the larger pools, where the trained bridge recovers some answers whichever source carries them; there we stand behind the control-subtracted lift and not the raw accuracy.

Which senders, and how many. Two sweeps bound the word *compatible*. Facing a pool of N holders of which one carries the answer, raw attention drops below floor early while the trained query-side bridge keeps a positive control-subtracted lift out to N=8 (Appendix A.1); varying the publisher instead, a bridge trained on the receiver's own state still transfers at cosine 0.9994 to the receiver's cache and is indistinguishable from its control by 0.9933 (Appendix A.2, Figure 4). Compatibility is narrow and measurable.

4.3 Activations: replacement transfers, and beats augmenting

We switch carrier and keep the probe. Two things are given to this mechanism and should be kept in view while reading the numbers: the fact region, obtained by diffing masked against full, and the injection layer, taken from a swept set. What follows measures how consumable a replaced activation is when it lands in the right place, and not whether an agent could find that place. Overwriting the receiver's residual stream at the fact region with the holder's activations clears the floor at every depth probed and peaks at L=12, where conditional accuracy is numerically indistinguishable from the receiver's native reference (Table 4.3). This is training-free: no cross-attention, no bridge, no trained parameters anywhere. The no-op control sits at *exactly* the floor at every layer, and the position control likewise, so the lift is not "something was overwritten here". Conditional accuracy exceeds span-copy roughly fourfold, which rules out regurgitation.

A caveat we must not bury. The mismatch control is clean at $L \leq 12$ (0.058–0.100, below floor) but *not* at $L \geq 16$, where it rises to 0.408–0.425. Injecting the holder's fact at the *wrong* positions lifts the

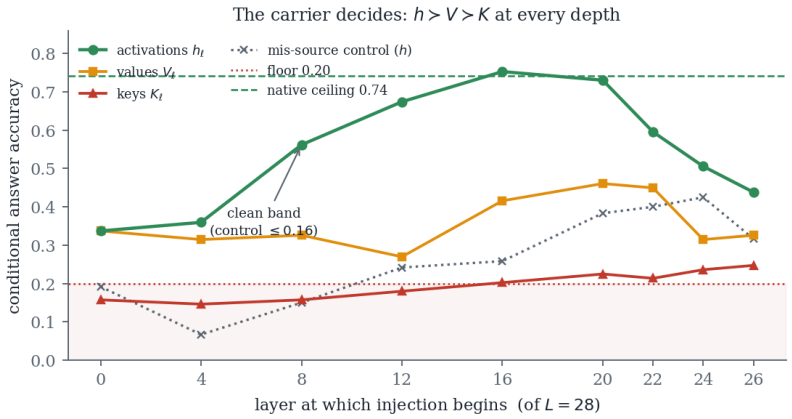


Figure 2: Carrier ordering across injection depth: activations dominate values, which dominate keys, and keys never leave the floor band. The clean operating band is shallower than the peak (Section 4.3).

receiver $\approx 2\times$ above floor at exactly the depths where replacement performs best. The no-op and position controls stay pinned everywhere, so the core claim survives. We read the clean result off the $L\in\{4, 8, 12\}$ band and treat the deeper layers as suggestive only.

Replacement beats augmentation on this carrier. We ran the same carrier through an augmentation mechanism instead, adding a routed convex combination of three holders' activations to the receiver's residual stream, again training-free. Its best cell reaches 0.438, clear of the floor yet far short of direct replacement. The dose response is monotone and gentle-is-better (α=$0.5 \succ 1 \succ 2 \succ 4$ at every depth). In line with a robust finding across this program, the router is *inert*: an oracle per-token router that sees the correct source earns nothing over the uniform one. Augmenting only *tilts* the receiver's own state, while replacement *installs* the holder's. Mechanism preference inverts between the two carriers, which is Conjecture 1(C2).

The carrier ordering. Holding the protocol fixed and varying only which object is shared gives a depth-stable ordering, activations over values over keys, spanning conditional accuracy 0.753 down to 0.247 (Figure 2). Keys are dead at every depth we probe. The ordering follows the consumption-point argument exactly: the key is read through the query dot-product and softmax, the value is averaged *after* the softmax and so is more forgiving, and the activation is read before either. One qualification. The value carrier's partial transfer appears only in the suffix regime (injecting at every layer from ℓ onward); at a single layer, $\mathbf{K}$, $\mathbf{V}$ and both together are flat-dead (lift $\leq +0.025$).

Why the two carriers differ. Two alternative explanations would account for the asymmetry without any appeal to the consumption point, and Appendix A.5 rules out both. The wall is not built by amplification: the measured readout gain is $O(1)$ on both sides of the manifold, far too small to turn a tolerable perturbation into a fatal one. It is not built by depth either, since activations injected into the same receiver through the same stack transfer while caches do not. What remains is where each object is read, and the measured tolerances scale with it: the activation carrier has roughly $20\times$ more geometric headroom than the key carrier.

Where it breaks. Two boundaries limit every positive result above, and Appendix A.6 reports both. Cross-family sharing collapses on both carriers: every arm sits at or below its own control, and a frozen projection of foreign keys is dead before any bridge trains, so the bottleneck is representational alignment and not the receiver's adaptability. Separately, the trained arm that first looked strongest on the activation carrier failed a mis-source control, and the arm the controls credit instead scores lower but survives them. A trained sharing mechanism means nothing until it is paired with a mis-source arm and an answer-disjoint split.

4.4 Summary of the two regimes

Table 4.4 collects the comparison the paper exists to make. Read by mechanism, the two carriers come out near-mirror images. The cache is dead under substitution and works under augmentation. The activation works under replacement, and works *less* under augmentation.

Mechanism	KV cache ($\mathbf{K}, \mathbf{V}$)	Activations ($\mathbf{h}$)
Substitution / replacement	0.000 *at floor, dead*	$\mathbf{0.742}$ *at ceiling*
Augmentation	$\mathbf{0.695}$ *at ceiling*	0.438 *above floor, below replacement*
Winning mechanism	**augment**	**replace**
Tolerable displacement	$\cos \approx 0.995$	$\cos \approx 0.90$
Consumption point	through the softmax	before the softmax
Cross-family	collapses	collapses

Table 4: The asymmetry. Best result per (carrier, mechanism) cell, each against its own probe's anchors. Cache numbers are shared-fact accuracy at n=200 (floor 0.18, ceiling 0.665); activation numbers are conditional accuracy at n=120 (floor 0.20, ceiling 0.742). Cells are not comparable across columns in absolute terms, only relative to their anchors.

5 Conclusion and Further Work

This paper asked whether internal state, requiring no model output, can serve as the channel between frozen agents in place of text. The answer depends less on *what* is shared than on how it is consumed. One carrier behaves oppositely under the two mechanisms: the KV cache cannot be substituted, because fidelity and consumability come apart under the softmax that reads it, and the same cache can be augmented. The activation inverts the rule. What unifies them is where the representation enters the receiver's computation. *Augment the cache; replace the activation.*

Limitations. Our conclusions are restricted to a controlled latent fact-transfer setting. Only the substitution bound is architecture-general (Appendix A.3); every *positive* result uses one model family, so we do not establish that the positive half of our rule holds across architectures, scales or attention shapes, and the mechanism we propose is architectural. The analytic constant $\sqrt{2G}$ is validated at a single group size, which the seven-receiver sweep varies only incidentally alongside depth, width and tokeniser. Cross-family transfer collapses on both carriers under every alignment mechanism we test, and the held-out-sender sweep (Appendix A.2) confines the usable range to a narrow band around the receiver, so our rule holds in a compatibility regime only.

The holder's cache contains the answer verbatim, a standing confound on every cache-side number. The hardened run of Section 4.1 clears the bar, but it shows only that copying does not explain the result, and it covers one arm.

Activation replacement is given oracle position and oracle layer, neither of which a deployed agent knows, so that result speaks only to replacement *correctly aligned*. The pool of Appendix A.1 is controlled multi-source retrieval and not a multi-agent workflow. Every run uses a single seed, so we claim that query-side augmentation *reaches* the native reference and never that it exceeds it. The accounting in Section 1 concerns the number of fitted alignment mechanisms, and says nothing about receiver-side computation, bandwidth, or the memory held by the pool.

Further work. *Cross-family alignment* is the blocking problem, and the activation carrier's wider geometric headroom makes it the more promising substrate. The informative experiment compares against closed-form alignment of last-layer states (Peng et al., 2026), which succeeds cross-model where our intermediate-layer injection fails; separating the alignment method from the injection depth would tell us whether the wall is real. Our account also predicts that fragility tracks the sharpness of the consuming readout and not depth, which existing evidence confounds, so varying attention sharpness and depth *independently*, through GQA group size and attention temperature, would test the consumption-point mechanism. Multiple seeds would settle whether query-side augmentation exceeds the native reference.

Nothing blocks internal communication between frozen agents in principle. What blocks it is the framing of *substituting* one agent's state for another's, which recent cache-level messaging proposals adopt (Fu et al., 2026; Dery et al., 2026). Agents do not need to replace one another's internal state; they need to attend over it, or, where the carrier is read before the softmax, to receive it directly.

USE OF AI TOOLS

We disclose the use of generative AI in this work, following the categories in the ICLR policy.

A LLM assistant edited and implemented some experiment code, including the manual attention forward pass that permits injection, the metric and control harnesses, and the plotting scripts; it assisted in analyzing and interpreting results; it gave feedback on experimental design, and proposed a few controls, including the mis-source arm and the answer-disjoint split whose application demoted one of our own earlier numbers; and it edited sections of this manuscript and compressed them to the page limit. It also searched for and summarized related work, and helped structure the paper.

We did not use generative AI to generate synthetic data, and the theoretical framing of the readout map and the tolerable set is our own, though an assistant helped state it formally and sharpened two claims we had originally written too strongly.

Every number in this paper comes from a logged run, and the numbers reported here were checked against the raw result files rather than against any assistant's summary of them. Verification gates on the experiment code (parity against native decoding, cache rebuild to 10^{-4}, and a bit-identical zero-initialized bridge) are themselves automated checks on assistant-written code, and we read their output before reading any result. We reviewed all AI-assisted work and we take responsibility for the final content of this paper, including its text, its claims, and the artifacts produced with AI assistance.

ETHICS STATEMENT

This work studies information transfer between frozen language models on a factual question-answering probe built from SQuAD, a public dataset. It involves no human subjects, no personal data, and no model training beyond small adapter bridges. One dual-use consideration: mechanisms that move information between models without passing through text also bypass text-level monitoring, so inter-agent channels of the kind studied here would need their own auditing provisions if deployed. Our results bear on this only indirectly, and suggest such channels are considerably harder to build than they appear.

REPRODUCIBILITY STATEMENT

Anonymous source for every experiment in this paper is included in the supplementary material, one script per claim, with a table mapping each script to the section and float it produces. Section 3 specifies models, the probe construction, both anchor pairs, all four controls, and the two verification gates every run passes, with the fine-grained geometry and the gate thresholds in Appendix A.4. Section 4 reports each number against the anchors of its own probe, and the figures are generated directly from the experiment result files by a single script included with the source, so every plotted value is traceable to the run that produced it. Where a result is a smoke test, contaminated, or statistically indistinguishable from its baseline, we say so inline: the cross-family collapses, the mismatch-control failure at $L \geq 16$, the contaminated trained-router arm, and the non-significant margin over the native reference.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.13245.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.04836.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language

model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2204.14198.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.07682.

Zhuohang Bian, Feiyang Wu, Chengrui Zhang, Hangcheng Dong, Yun Liang, and Youwei Zhuo. TokenDance: Scaling multi-agent LLM serving via collective KV cache sharing. *arXiv preprint arXiv:2604.03143*, 2026.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2112.04426.

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.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2306.09346.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

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 (ICLR)*, 2026. arXiv:2510.03215.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024. arXiv:2311.04934.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.00352.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2401.18079.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021. arXiv:2007.01282.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.03206.

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

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. RAGCache: Efficient knowledge caching for retrieval-augmented generation. *arXiv preprint arXiv:2404.12457*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1905.00414.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023. arXiv:2309.06180.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. arXiv:1411.5908.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a. arXiv:2303.17760.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv:2306.03341; spotlight.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.14469.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. MiniCache: KV cache compression in depth dimension for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2405.14366.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in-context learning more effective and controllable through latent space steering. In *International Conference on Machine Learning (ICML)*, 2024b. arXiv:2311.06668.

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

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, and Esha Choukse. DroidSpeak: KV cache sharing for cross-LLM communication and multi-LLM serving. *arXiv preprint arXiv:2411.02820*, 2024c.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.17118.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024d. arXiv:2402.02750.

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 (ICLR)*, 2023. arXiv:2209.15430; notable top-5%.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

Yanwen Peng, Delvin Ce Zhang, Xi Wang, and Nikolaos Aletras. StateBridge: Training-free hidden-state alignment for latent communication in LLM multi-agent systems. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2608.13317.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.05806.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2205.05124.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.05254.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2401.10491.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Markus Wenzel. Latent communication between language model agents: Channels, alignment, and the limits of text. *arXiv preprint arXiv:2607.14103*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2203.05482.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2308.08155.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.01708.

Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, and Yiran Chen. KVCOMM: Online cross-context KV-cache communication for efficient LLM-based multi-agent systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.12872.

Ye Yu, Heming Liu, Haibo Jin, Xiaopeng Yuan, Peng Kuang, and Haohan Wang. Learning to communicate: Toward end-to-end optimization of multi-agent language systems. *arXiv preprint arXiv:2604.21794*, 2026.

Huixiang Zhang and Mahzabeen Emu. Do latent channels actually communicate? a causal audit of latent multi-agent LLM. *arXiv preprint arXiv:2607.26773*, 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14048.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix

Conditioning measurement. The readout Jacobian gain κ is estimated over 40 probe contexts, perturbation magnitudes $\delta \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$, 8 random directions per probe and 3 seeds, with an on-/off-manifold tolerance of 0.1. Mean gains are $\kappa_{\text{off}} = 0.506$ and $\kappa_{\text{on}} = 0.343$, a ratio of $1.475\times$. The analytic Jacobian bound predicts the per-layer measured gain at Pearson correlation 0.999, with governing constant $\sqrt{2G} = \sqrt{12} \approx 3.46$ for G=6. Per-layer gains are heterogeneous, small early (around 0.11 at layer 1) and rising to roughly 0.72 by layer 6, but nowhere exhibit a cliff.

Bridge training. Query-side bridges are hand-rolled rank-16 LoRA on the receiver's own q, k and v projections at all 28 layers, with $\Delta = (hA)B$, $A \sim 0.01 \cdot \mathcal{N}(0, 1)$ and B initialized to zero, so the untrained module is exactly the identity. Training minimizes cross-entropy of the gold answer's first token at a single reserved decision step, through the frozen receiver, with Adam at learning rate 2×10^{-4}, gradient clipping 1.0, and four epochs over 200 training items; a keep-best guard evaluates a 32-item slice each epoch. We note as a limitation that this slice is small enough that its scores do not track the 200-item evaluation, so checkpoint selection is noisy. A clean version of the adaptation ladder needs a proper validation split.

Substitution sweep. For completeness, the substitution mechanisms surveyed in Section 4.1 and their transfer against the floor of 0.18: freeform regression 0.00; manifold-matched training objective 0.01; by-construction column-space synthesis 0.01; row-space target 0.085; worst-case hinge 0.08; high-capacity per-layer MLP 0.20; codebook selection 0.00; oracle nearest-genuine-key 0.000. The identity arm reproduces the ceiling 0.665 exactly. Fidelity within this family is uncorrelated with transfer: the best-fidelity member (codebook selection, cosine 0.880, attention-logit correlation 0.928) transfers at zero.

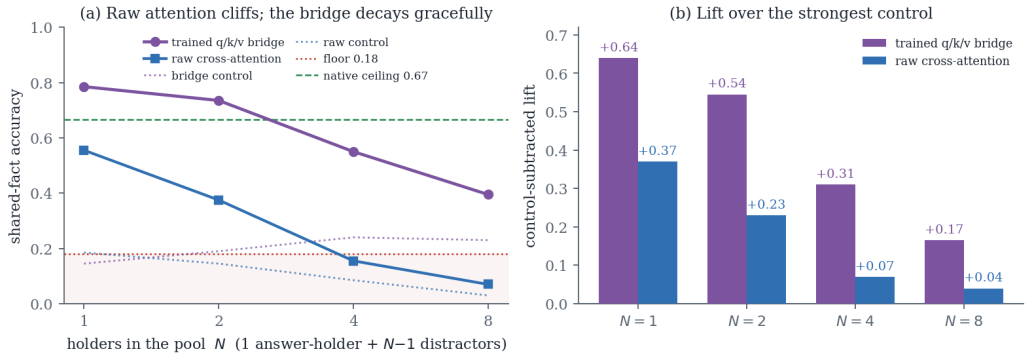


Figure 3: Multi-source KV augmentation over a pool of N holders, one of which read the answer-bearing passage. **(a)** Raw cross-attention collapses below the floor by $N{=}4$; the trained query-side bridge decays gracefully and remains above floor at $N{=}8$. Dotted lines are the corresponding controls, which themselves lift above floor at large N. **(b)** Lift over the strongest control at each pool size: positive throughout, eroding with N.

A.1 Multi-source sharing: a pool of holders

A pairwise result is not yet a multi-source one. We study the primitive a multi-agent internal channel would need: sender-agnostic retrieval of task-relevant state from a shared pool. Concretely, we place the receiver in front of a pool of N holders, of which exactly one has read the answer-bearing passage and the other $N{-}1$ have read distractor passages, and let it cross-attend over the union of their caches. The bridge is trained once at pool size 4 and evaluated at every N, so $N \in \{1, 2, 8\}$ also tests generalization across pool size.

Figure 3 shows a clean dissociation. Raw cross-attention *cliffs*: $0.555 \to 0.375 \to 0.155 \to 0.070$, falling below the floor of 0.18 by $N{=}4$. Left to itself, the receiver cannot locate the answer-holder among distractors. The trained query-side bridge *decays gracefully*: $0.785 \to 0.735 \to 0.550 \to 0.395$, still more than twice the floor at $N{=}8$. The query side is doing selection work that raw attention cannot do.

Honesty requires the control-subtracted view, and we lead with it. Both controls (pool with the answer-holder also masked; pool of pure distractors) sit cleanly below floor at $N \leq 2$, but climb to ≈ 0.23–0.24 at $N \geq 4$. With a large pool and no answer present, the trained receiver still answers about a quarter of the time by leaning on its own masked prefix and priors. Subtracting the strongest control gives lifts of $+0.640$, $+0.545$, $+0.310$ and $+0.165$ at $N{=}1, 2, 4, 8$. The lift is positive at every pool size, and eroding. That is the substantive finding, and it is smaller than the raw numbers suggest.

A.2 Held-out senders: how many agents one bridge serves

The pool of Section A.1 varies what each holder read, not who read it. Every holder there is the receiver's own weights over a different passage, so that experiment establishes invariance to pool size and to the answer-holder's slot. What the sharing argument needs is stronger: a mechanism that does not care which agent published the state. We test that here.

Senders that differ only in weights. Holders are drawn from four Qwen2.5-1.5B checkpoints: the receiver itself, the instruction-tuned model, the code model and the math model. All four share the receiver's depth, width and head layout, and their tokenisers are bit-identical, so the comparison isolates sender identity. This matters because a cross-family holder needs a projection between incompatible key spaces, and Section A.6 shows that the projection is what fails. Here there is no projection to blame. The math checkpoint ships $\theta_{\mathrm{RoPE}}{=}10^4$ against the receiver's 10^6; keys are stored after rotation, so we force the receiver's θ on every sender and a frequency mismatch cannot masquerade as a sender effect. Each sender must clear three gates before its numbers are read: the receiver's exact cache shape, finite values, and measurable distinctness from the receiver. The three

sender	cosine	rel. diff.	raw lift	bridge shared	bridge lift	95% interval
self (seen)	—	—	$+0.070$	0.550	$+0.310$	$[+0.219, +0.401]$
instruct (held out)	0.9994	0.033	$+0.070$	0.550	$+0.325$	$[+0.235, +0.415]$
coder (held out)	0.9933	0.170	$+0.000$	0.130	$+0.030$	$[-0.032, +0.092]$
math (held out)	0.9810	0.433	$+0.000$	0.100	-0.005	$[-0.064, +0.054]$

Table 5: Held-out senders. One rank-16 query-side bridge, trained on pools built from the receiver's own weights only, evaluated against senders it never saw. Senders are Qwen2.5-1.5B checkpoints with the receiver's geometry and a bit-identical tokeniser, ordered by the cosine between their cache and the receiver's. We report relative Frobenius difference alongside cosine: the two rank the senders identically, so the ordering is not an artifact of one metric, though two metrics over three checkpoints cannot settle which is the right notion of compatibility. Lift is shared accuracy minus the stronger of the two controls; intervals are 95% normal-approximation intervals on that difference. Pool size $N{=}4$, $n{=}200$, floor 0.18, native ceiling 0.665. The *self* row reproduces the $N{=}4$ point of Figure 3 exactly.

held-out senders sit at cosine 0.9994, 0.9933 and 0.9810 to the receiver's own cache, which gives a graded axis rather than a binary contrast.

One bridge is trained, on pools built from the receiver's own weights alone, and then evaluated against each sender unchanged. Two cells pin the baseline: raw attention scores 0.155 and the trained bridge 0.550, with controls at $0.085/0.045$ and $0.200/0.240$. Both reproduce the $N{=}4$ column of Section A.1 on all four values, so sender identity is the only quantity that varies between cells.

The nearest sender transfers; the next does not. Table A.2 gives the result. At cosine 0.9994 the bridge reaches 0.550 against a strongest control of 0.225, a control-subtracted lift of $+0.325$ with a 95% interval of $[+0.235, +0.415]$. That equals what the bridge achieves on the sender it trained on, and it is earned rather than inherited: raw attention on the same sender reaches only 0.145, below the floor of 0.18. The adapter is doing sender-agnostic work. One step further out the effect is gone. At cosine 0.9933 the lift is $+0.030$ with an interval of $[-0.032, +0.092]$, and at 0.9810 it is -0.005. Neither is distinguishable from its own control, and both sit below the floor in absolute terms.

Where the radius comes from. The receiver's tolerance for a displaced key is narrow and is not our finding: it is the tolerable set of Section 2, and the displacement at which keys stop being readable has been characterized independently. What this experiment adds is that the same narrowness governs *who* may join a pool, not only *what* may be substituted into one. Augmentation lets the receiver keep its own cache, so it escapes the bound for a cache the receiver could have produced; it does not repeal the bound for a cache produced by different weights. The radius is therefore a property we inherit, and the contribution here is measuring where it falls on the sender axis.

What this buys. The sharing argument of Section 1 needs one mechanism that serves every publisher, and a pairwise bridge per ordered pair is what it replaces. The measurement says how far that replacement goes. Within a narrow band of the receiver, one bridge covers senders it never trained on, so such a pool needs one fitted bridge rather than one per pair. Outside the band it covers nothing, and those agents fall back to text or to a fitted pair. Sharing is therefore a real alternative to transfer for a population of closely related agents, which is the common case when a deployment scales one base model into differently tuned roles, and it is not yet an alternative for a heterogeneous population. We state the bounded version because it tells a system designer which agents may share a pool.

A.3 Does the wall depend on the architecture?

Everything above holds one receiver fixed, which leaves open whether the substitution bound is a property of attention or a property of Qwen2.5-1.5B. The mechanism we propose is architectural, so this is a question the paper has to answer rather than defer. We repeat the oracle-substitution probe of Section 4.1 on seven receivers spanning three families, six depths and five attention geometries, re-deriving each receiver's own floor and ceiling so that every row is read against its own anchors.

Table A.3 gives the result. On every receiver the identity arm reproduces that receiver's ceiling exactly, so the manual injection path is faithful on all seven architectures. And on every receiver the nearest genuine key transfers at or below 0.008, against floors ranging from 0.000 to 0.350. The

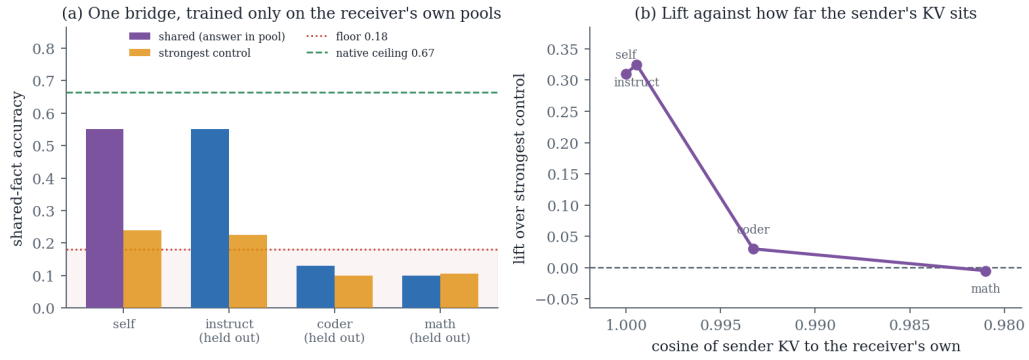


Figure 4: Held-out senders. **(a)** The bridge, trained on the receiver's own pools only, reaches the same accuracy on the nearest held-out sender as on the sender it trained on, and falls to its control on the two more distant senders. **(b)** Control-subtracted lift against the cosine between the sender's cache and the receiver's own. The effect is present at 0.9994 and absent by 0.9933, which measures where the effect falls on the sender axis. Three checkpoints locate a transition, not a radius.

receiver	L	G	floor	identity	nearest genuine
Pythia-70M	6	1	0.000	0.025	0.000
GPT-2	12	1	0.033	0.233	0.000
Pythia-410M	24	1	0.067	0.375	0.000
Pythia-1.4B	24	1	0.092	0.517	0.000
Qwen2.5-0.5B	24	7	0.092	0.633	0.008
Qwen2.5-1.5B	28	6	0.200	0.742	0.000
Qwen2.5-3B	36	8	0.350	0.792	0.000

Table 6: The substitution bound across architectures. Each receiver is probed with its own bank of genuine keys and its own anchors. *Identity* hands the receiver its own exact keys and recovers that receiver's ceiling in every case, which verifies the harness on each architecture. *Nearest genuine* substitutes the closest key the receiver itself produced elsewhere. Depth spans L=6 to L=36 and the grouped-query factor spans G=1 (multi-head) to G=8. n=120.

bound is not a Qwen property. It appears at six layers and at thirty-six, in three model families, and under both multi-head attention (G=1) and grouped-query attention up to G=8.

Two qualifications matter. First, the variation in G here is incidental rather than controlled: these checkpoints differ in depth, width, training data and tokeniser at the same time, so the table shows that the bound survives architectural variation without isolating G as the variable responsible. The clean experiment for the $\sqrt{2G}$ prediction of Section 2 would vary group size within one architecture, and we have not run it. Second, this sweep probes substitution only. It establishes that the failure mode generalizes; it does not establish that the positive half of our rule, augmentation on the cache and replacement on the activation, generalizes with it. The receivers here are also probed in isolation rather than paired, so nothing in this table speaks to cross-family sharing, which Section A.6 reports separately as a collapse.

A.4 Methodological detail: model geometry and verification gates

Receiver and holder geometry. The frozen Qwen2.5-1.5B (Qwen Team, 2024) receiver has L=28 layers and hidden width 1536, with grouped-query attention over N_Q=12 query heads and N_{KV}=2 key/value heads (G=6), head dimension 128, and key/value width 256. The cross-family holder, GPT-Neo-1.3B, has 24 layers and key/value width 2048. We project it into the receiver's key space with a frozen per-layer SVD map fitted on a 40-item sample, aligning depth proportionally between the two stacks.

Verification. All attention is re-implemented manually to permit injection, so every run passes two gates before its numbers are read. A *parity* gate requires the manual forward to reproduce native

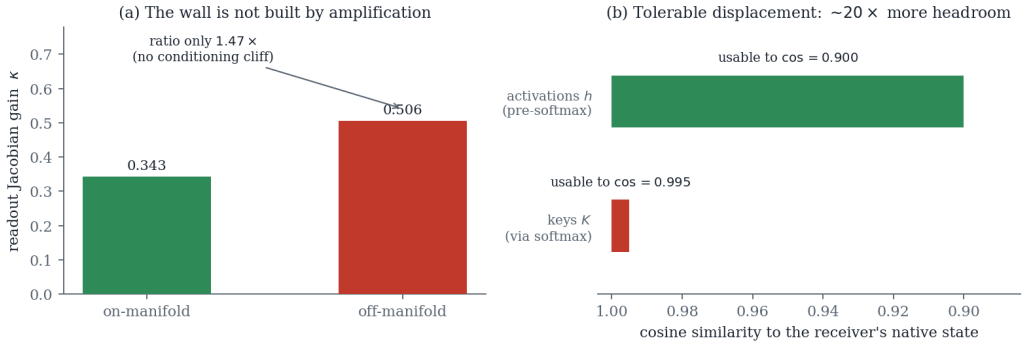


Figure 5: Why the two carriers differ. **(a)** The measured readout Jacobian gain is $O(1)$ both on and off manifold, with a ratio of only $1.47\times$: the wall is not built by amplification, so no single layer and no stacking of layers blows up a tolerable perturbation. **(b)** Tolerable displacement by carrier: activations remain usable to cosine 0.90, keys only to ≈ 0.995, about $20\times$ less headroom. Same receiver, same depth; the difference is the consumption point.

greedy decoding string-exactly from the same prefix. A *rebuild* gate requires reconstructed prefix keys and values to match the native cache (maximum absolute deviation $\leq 10^{-4}$). Trained arms additionally assert that the zero-initialized bridge scores bit-identically to the untrained arm at epoch zero.

A.5 Mechanism: the wall is at the readout, not the depth

Proposition 2 makes two claims we can measure. The first is that there is no conditioning cliff. Estimating the readout Jacobian gain over 40 probes, six perturbation magnitudes, eight directions per probe and three seeds gives $\kappa_{\text{off}} = 0.506$ against $\kappa_{\text{on}} = 0.343$, a ratio of $1.47\times$ (Figure 5a). Off-manifold displacement is modestly more damaging, not catastrophically so. Per-layer gains are $O(1)$ on both sides and no single layer explodes. An analytic Jacobian bound predicts the measured per-slot gain at Pearson correlation 0.999 with the governing constant $\sqrt{2G}$. This pins the sensitivity to the shape of attention, and it is far too small to be the wall.

The second claim is that depth is not the culprit. The control is the carrier comparison itself: activations injected into the identical 28-layer receiver, consumed through the identical stack, transfer at 0.72 while caches do not. If per-layer error compounded to unreadability, the activation would die alongside the cache. It does not. What differs is only *where* in each layer the object is consumed.

The tolerance measurements close the argument quantitatively (Figure 5b). Perturbing a shared activation and sweeping matched cosine, conditional accuracy stays near 0.70 all the way down to cosine 0.90. The KV carrier is at the floor across the same sweep, and genuine nearest-neighbor keys at cosine ≈ 0.995 already fail. That is roughly $20\times$ more geometric headroom for the activation, the budget a learned cross-model bridge would have to spend.

A.6 Where it breaks: cross-family transfer, and a control that bites

Two boundaries limit every positive result above, and both deserve to be stated plainly.

Cross-family collapses on both carriers. With a GPT-Neo-1.3B holder and a Qwen receiver, every KV augmentation arm sits at or below its own control (Table 4.2, lower block; the best is 0.080 against a control of 0.085, i.e. noise). The identity sanity arm is 0.000: the frozen SVD projection of foreign keys is dead before any bridge trains. On the activation side, all eight cross-family cells of the trained-router experiment also sit at floor, with the trained arm scoring *identically* to the uniform one, so the bridge learned nothing usable. Rank-16 adaptation of the receiver's own projections does not bridge a foreign representation. The bottleneck here is representational alignment, rather than the receiver's adaptability or the carrier.

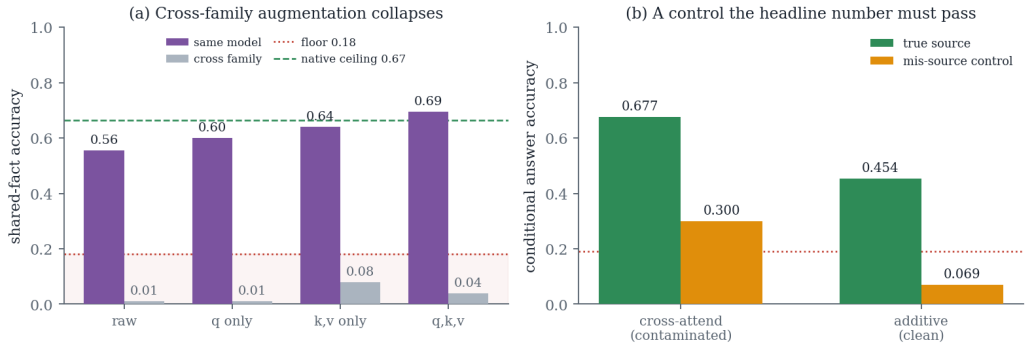


Figure 6: The boundaries. **(a)** Cross-family KV augmentation collapses at every rung of the adaptation ladder, while the same ladder within one model reaches the native ceiling. **(b)** On the activation carrier, the apparently strongest trained arm fails a mis-source control (0.300 against a floor of 0.19); the additive arm passes it (0.069) at a lower but genuine 0.454.

A trained-router positive, and the control that halved it. Training a router on the activation carrier initially produced a striking number: conditional accuracy 0.677 for a trained cross-attention consumption path, against a uniform-router baseline at floor. A mis-source control and an answer-disjoint split, run afterwards, showed this to be substantially contaminated. Feeding the router activations from a *different* item still yielded 0.300, far above the floor of 0.19, so a large part of the headline was the router emitting answers regardless of source. The same controls credit a different arm (additive consumption, router-only adaptation) with conditional accuracy 0.454 against a mis-source control of 0.069, a genuine controlled positive on unseen answers. Both LoRA-adapted arms, meanwhile, collapse to ≤ 0.115 once answers are disjoint, which reveals them as answer-vocabulary memorisation (Figure 6). We report the 0.454 as the defensible trained-router result and treat the 0.677 as an artifact. The transferable lesson is that a trained sharing mechanism must be paired with a mis-source arm and an answer-disjoint split before its number means anything.

Controls in full. *No-op*: inject the receiver's *own* masked state at the same positions and layers, separating "something was overwritten here" from "the holder's content was overwritten in." *Position*: place the holder's state at the fact-free common prefix. *Mismatch / mis-source*: supply the holder's state for a *different* item, which must collapse if the mechanism consumes source content instead of emitting memorised answers. *Answer-disjoint split*: for trained bridges, no normalized eval answer appears anywhere in train.