Communication Faithfulness in Latent Multi-Agent LLMs: Bandwidth and Norm, Not Feature Identity

Anonymous ACL submission

Abstract

Latent communication between LLM agents is typically evaluated by the sender's sparse autoencoder (SAE) reconstruction quality, a metric independent of whether the receiver actually uses the transmitted code. We introduce *communication faithfulness*, a receiver-side measure that quantifies each SAE feature's causal effect on the receiver's gold-token probability via activation patching, and use it as a probe of the latent channel. Two phenomena emerge. **(i) Decoupling:** sender-side reconstruction quality is at best weakly aligned with receiver utility; high-FVE reconstructions routinely fail to drive the receiver, and low-FVE ones often succeed. **(ii) Bandwidth and norm, not feature identity:** when full-SAE injection hurts the receiver, the harm is recovered almost entirely by bandwidth reduction (a $K{=}10$ subset matches or exceeds full-SAE injection at every layer) and injected-norm control; among $K{=}10$ subsets, the advantage of our receiver-causal ranking over uninformed selection from the same SAE is within test-fold noise on every layer. A role/function taxonomy further shows that the two 8B backbones organize their channels differently even when these levers transfer.

1 Introduction

Large language models (LLMs) deployed as collaborating multi-agent systems can outperform single-agent chain-of-thought on reasoning, debate, and multi-hop question answering (Wei et al., 2022; Du et al., 2024; Liang et al., 2024). The canonical collaboration channel is natural language, but a recent line of work sends hidden states between agents directly: LatentMAS (Zou et al., 2025) has agents generate last-layer hidden embeddings as latent thoughts and pass them through a shared latent working memory, SDE (Tang et al., 2025) transfers natural-language tokens augmented with token-wise state-transition trajectories, and Interlat (Du et al., 2025) learns a compressed code that the receiver decodes with a small adapter. A natural way to evaluate these latent channels is to reuse a single-agent interpretability tool: a sparse autoencoder (SAE) trained on the sender's activations (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024), scored by the fraction of variance it explains (FVE).

FVE is sender-side and geometric: it scores how well the SAE reproduces h_s, with no dependence on the receiver. Nothing in the SAE training forces the directions FVE rewards to be the directions the receiver actually reads. We therefore split the question in two. *Reconstruction faithfulness* is the sender-side round-trip quality (FVE). *Communication faithfulness* is its receiver-side causal counterpart: given a candidate replacement for the sender's residual, does the receiver still produce the correct gold token once the replacement is patched into its forward pass?

Figure 1 illustrates both notions and the four-step selection pipeline. The procedure reuses standard activation-patching tools (Meng et al., 2022; Wang et al., 2023; Heimersheim and Nanda, 2024) and trains no new SAE or probe. our codes are open-sourced [1].

Our contributions are threefold:

- We introduce *communication faithfulness*, a receiver-side causal counterpart to sender-side reconstruction quality, operationalised as a per-feature activation-patching score and treated as a measurement protocol throughout.
- We document a *decoupling* phenomenon: across backbones, benchmarks, inject layers, and compression ratios, sender-side FVE carries essentially no information about whether the receiver can use the signal; high-FVE reconstructions routinely fail to drive the receiver, and low-FVE ones often succeed.

[1] https://anonymous.4open.science/r/agent_communication_SAE-D408/

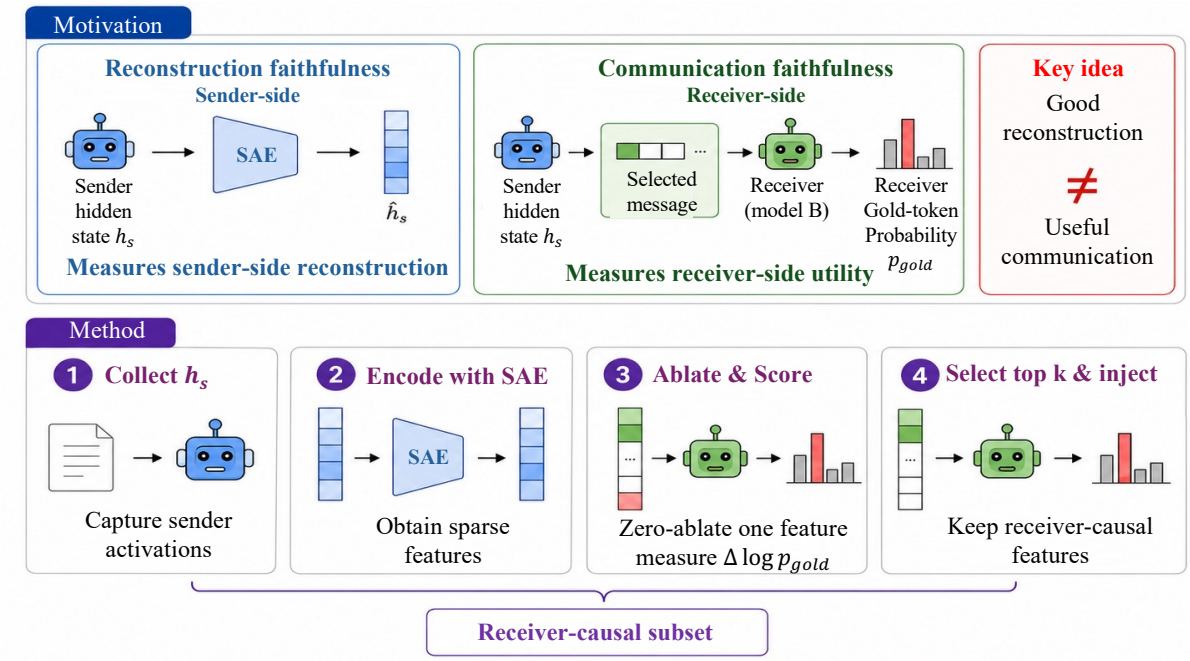


Figure 1: Overview. **Top (motivation):** reconstruction faithfulness scores only the sender-side SAE round-trip ($\hat{h}_s \approx h_s$); communication faithfulness scores the receiver's gold-token probability p_{gold} once the selected sender message is patched into its forward pass. Good reconstruction does not imply useful communication. **Bottom (probe):** the receiver-causal scoring pipeline used as a diagnostic: (1) collect sender activations h_s, (2) encode them with the SAE into sparse features, (3) zero-ablate each feature and score it by its effect on $\log p_{\text{gold}}$, and (4) use the resulting per-feature scores to decompose what drives the receiver: *bandwidth* (how many features), *norm* (how loudly), and *identity* (which ones). We find the first two dominate.

- Using this score as a probe, we find that the practical levers for latent communication are *bandwidth* and *injected norm*, not feature identity at fixed bandwidth. A $K{=}10$ subset matches or exceeds full-SAE injection at every layer, with most of the harm-layer recovery carried by the injected-norm difference, and at fixed $K{=}10$ our receiver-causal ranking over uninformed selection lies within test-fold noise.

2 Preliminary

SAEs. A sparse autoencoder (SAE) (Huben et al., 2024; Bricken et al., 2023) is a learned dictionary that decomposes a residual-stream activation $h \in \mathbb{R}^d$ into a sparse linear combination of d_{sae} feature directions. We write the encoder–decoder round-trip as $\text{SAE}(h) \approx h$ and, for a subset $\mathcal{F}$ of feature indices, write $\text{SAE}_{\mathcal{F}}(h)$ for the same operation restricted to features in $\mathcal{F}$ during decoding. Following EleutherAI (2024), we treat each column of the decoder as a "feature" throughout this paper.

Sender–receiver protocol. We use a sender–receiver residual-injection protocol inspired by latent-space communication work such as Latent-MAS (Zou et al., 2025); our intervention is a fixed-layer residual-stream addition rather than a shared latent working memory. Two instances of the same LLM collaborate on a task. The sender sees the task with a partial context and produces a short textual summary; we capture its residual-stream activation h_s at a designated inject layer at the last prompt token. The receiver sees the task with the remaining context and the sender's summary; at the same inject layer we add the captured activation, or some candidate replacement, to the receiver's residual stream at the last prompt token. The receiver continues with greedy decoding. The candidate replacement is $\text{SAE}_{\mathcal{F}}(h_s)$ for some feature subset $\mathcal{F}$; the choice of $\mathcal{F}$ is the only quantity that varies across the methods we compare.

Datasets. We evaluate on six task families that span multi-hop QA, mathematical reasoning, science QA, and clinical multiple-choice: HotpotQA (Yang et al., 2018), Hendrycks MATH (Hendrycks et al., 2021), ARC-Easy and ARC-Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), and MedQA-USMLE (Jin et al., 2021). Each

dataset is split deterministically with seed 0 into four disjoint folds of 50 tasks: $D_{\text{discovery}}$ (candidate pool and Jacobian rank), $D_{\text{selection}}$ (per-feature causal effect ranking), $D_{\text{validation}}$ (held-out causal check), and D_{test} (locked final evaluation). N_{test} is a compute-cap prefix of the 50-task fold: 28 for the main-paper Llama HotpotQA run, 25 for the four short-form benchmarks, and as low as 20 for a few legacy Qwen MATH cells. Split hashes are in the supplementary.

Models. We use four backbones spanning two series and two scales. (i) Llama-3.1-8B-Instruct paired with the EleutherAI TopK residual-stream SAE (EleutherAI, 2024) ($d_{\text{sae}} = 131{,}072$, $k = 192$, trained on Llama-3-8B base activations; He et al. (2024) release a related family of open SAEs for Llama-3.1-8B-Base). Because this SAE was trained on Llama-3-8B base activations, we report HotpotQA and MATH runs on Llama-3-8B-Instruct (the base-matched checkpoint) and the four short-form benchmarks on Llama-3.1-8B-Instruct; the two Instruct-tuned checkpoints share the Llama-3 architecture and tokenizer, and we apply the same SAE to both as a transfer of the Llama-3-base-trained SAE to Llama-3.1 activations. (ii) Qwen3-8B-Base with the public Qwen-Scope residual-stream TopK SAE release (Deng et al., 2026) ($d_{\text{sae}} = 65{,}536$, $k = 50$). (iii) Qwen3-1.7B-Base with Qwen-Scope ($d_{\text{sae}} = 32{,}768$, $k = 50$). (iv) Llama-3.2-1B-Instruct with a public residual-stream ReLU SAE (Chanin, 2024) ($d_{\text{sae}} = 16{,}384$, expansion factor 8). No SAE is trained in this work; every checkpoint is a pinned public release. Inject layers are spread across the network: $\ell \in \{8, 12, 14, 16, 24\}$ on Llama-3 / Llama-3.1-8B, $\ell \in \{14, 16, 18\}$ on Qwen3-8B, $\ell \in \{11, 14, 18\}$ on Qwen3-1.7B, and $\ell \in \{4, 6, 8\}$ on Llama-3.2-1B (constrained to the layer range the public SAE covers).

Candidate pool. The candidate pool is the top $N_{\text{pool}} = 200$ SAE features by mean activation magnitude on $D_{\text{discovery}}$; this pool is fixed before any D_{test} access.

3 Reconstruction and Communication Faithfulness

Prior latent-communication work (Zou et al., 2025; Tang et al., 2025; Du et al., 2025) evaluates latent channels primarily through downstream task success or sender-side representation losses, leaving open whether sender-side reconstruction quality tracks the receiver-side causal effect of the transmitted signal. We isolate this sender-side geometric quantity from its receiver-side causal counterpart and analyze the relationship between the two on a locked test fold.

3.1 Two Faithfulness Notions

Reconstruction faithfulness. The geometric quality of the sender-side round-trip is

$$\mathrm{FVE}(h_s, \mathcal{F}) = 1 - \frac{\|h_s - \mathrm{SAE}_{\mathcal{F}}(h_s)\|^2}{\|h_s\|^2}. \quad (1)$$

This is the loss the SAE was trained on and the natural sender-side metric if one applies SAE interpretability to the transmitted code; it is a function of h_s and $\mathcal{F}$ alone, with no dependence on the receiver.

Communication faithfulness. The receiver-side causal quantity reads the receiver's gold-token log-probability under residual injection. For a task t with gold next token y_t, let $L(t; h)$ be the receiver's log-probability of y_t when residual h is added to its stream at the inject layer at the last prompt token. Define

$$\mathrm{CF}(h_s, \mathcal{F}) = \mathbb{E}_t\big[L\big(t;\, \mathrm{SAE}_{\mathcal{F}}(h_s)\big)\big]. \quad (2)$$

This quantity depends on the receiver's parameters, the inject layer, the rest of the receiver's context, and the chosen subset $\mathcal{F}$. Everything that might affect whether the receiver produces the right answer is in scope.

We measure the rank-correspondence between FVE and CF by the Spearman rank correlation ρ on a held-out fold, and write $|\rho|$ for its absolute magnitude (*not* a p-value).

3.2 Receiver-Readout Effective Rank

The receiver reads only a small subspace of the residual: most directions in a d-dimensional residual are not consumed by any downstream MLP block or by the unembedding. The receiver's downstream dependence on the injected h_s decomposes as $Y = U \circ \text{MLP-chain} \circ (r_\ell + h_s)$, where U is the unembedding, MLP-chain is the rest of the receiver's decoder layers, and r_ℓ is the receiver's own residual at the inject layer. Differentiating Y with respect to h_s gives the receiver's effective readout direction $J_{\text{recv}} = \partial Y / \partial h_s$. We define its effective rank as

$$r_{\text{eff}} = \big|\{i : \sigma_i(J_{\text{recv}}) > 0.01 \cdot \sigma_{\max}\}\big|. \quad (3)$$

A random-probe Jacobian SVD on $D_{\text{discovery}}$ at the LatentMAS inject layer on Llama-3-8B-Instruct returns $r_{\text{eff}} = 32$ at a 32-probe budget on every task; the singular spectrum decays smoothly past the cap, consistent with $r_{\text{eff}} \ll d = 4096$ (Appendix D). This low-rank receiver readout is a heuristic rationale for choosing K at the scale of r_{eff} rather than d_{sae}, not a sufficiency bound.

3.3 Decoupling Result

We test the rank correlation between FVE and CF on D_{test} across four backbones (Llama-3-8B-Instruct / Llama-3.1-8B-Instruct, Qwen3-8B, Llama-3.2-1B-Instruct, Qwen3-1.7B-Base), six benchmarks, fourteen inject layers, and four compression ratios, totaling 438 cells. For each cell we compute the Spearman ρ between $\text{FVE}(h_s)$ and the receiver's gold-token log-probability under $\text{SAE}_{\mathcal{F}}(h_s)$, with 500-bootstrap 95% confidence intervals, 500-permutation paired null p-values, and Benjamini–Hochberg FDR correction at $\alpha = 0.05$ (Benjamini and Hochberg, 1995). Per-cell point estimates and 95% CIs at the three core-cell layers are in Table 1; the per-method log-prob gain at the same ratio is in Figure 2.

We find no strong positive correlation between FVE and the receiver's gold-token log-prob. We adopt $|\rho| < 0.35$ as an operational weak-correlation cutoff (Cohen's conventional benchmarks are 0.10, 0.30, 0.50 for small / medium / large effects (Cohen, 1988); our cutoff sits at or just above medium). The vast majority of cells sit below this cutoff (Table 3) and the matrix-wide mean is near zero. The largest aggregate correlations sit on the multiple-choice ARC benchmarks ($\bar{\rho}$ down to -0.74 on Qwen3-1.7B and -0.52 on Llama-3.1-8B), where the receiver's answer logits are prompt-dominated and the candidate set is narrow, so both faithfulness signals are small in absolute terms and the rank correspondence reads stronger than the underlying signal would justify. We emphasise that, given the small per-cell N_{test} and the wide bootstrap CIs of Table 1, our finding is best read as *absence of a strong positive correlation under the protocol we test* rather than as positive evidence of a near-zero correlation; the two readings have the same operational implication (a sender-side FVE proxy is not a tight predictor of the receiver-side causal effect in our setup), but the second is a stronger statistical claim that our test folds do not support.

The same picture holds across (model, benchmark) pairs. The same weak rank correlation reappears in every (model, benchmark) pair we test (Table 3 in §5), consistent with the receiver's Jacobian being low rank ($r_{\text{eff}} \geq 32$ against $d = 4096$, §3.2): a sender-side metric that scores how well the SAE reproduces all of h_s has limited bearing on receiver behaviour.

4 Receiver-Causal Scoring and Bandwidth Decomposition

Standard SAE feature-importance scores rank by activation magnitude or by single-feature contribution to reconstruction quality. We define a receiver-side counterpart: each feature's measured effect on the receiver's gold-token log-probability under activation patching. We treat this score as a measurement protocol, using it both to rank features for compact transmission and to decompose what actually drives the receiver into bandwidth, injected norm, and feature-identity contributions.

4.1 Selection Procedure

Marginal effect. For each candidate feature j in a pool of size N_{pool},

$$\text{effect}_j = \mathbb{E}_i \Big| L_{\text{g}}\big(i;\, \text{SAE}(h_s^i)\big) - L_{\text{g}}\big(i;\, \text{SAE}_{-j}(h_s^i)\big) \Big|, \quad (4)$$

where L_{g} is the receiver's gold-token log-probability and SAE_{-j} denotes the SAE round-trip with feature j zeroed before decoding.[2] Marginal greedy sorts candidates by descending effect_j and takes the top $K = \lceil r \cdot N_{\text{pool}} \rceil$ features at compression ratio r.

Redundancy-aware sequential greedy. Marginal scoring is suboptimal when SAE features overlap in decoder space: top-K marginal effects can pick K near-parallel directions that encode the same concept, wasting bandwidth on duplicate signal. We replace the sort with sequential greedy on a redundancy-penalised score. Let $\mathcal{F}_0^\star = \{\arg\max_j \text{effect}_j\}$. For $k = 1, \ldots, K-1$,

$$\mathcal{F}_{k+1}^\star = \mathcal{F}_k^\star \cup \Big\{ \arg\max_{j \notin \mathcal{F}_k^\star} \text{effect}_j \cdot \big(1 - \rho_{j,\mathcal{F}_k^\star}\big)^\alpha \Big\}, \quad (5)$$

[2] The absolute value ranks by causal informativeness rather than net helpfulness; the signed split is captured by the role taxonomy (§5), and a sign-aware variant is left for future work.

method	ℓ=12	ℓ=14	ℓ=16
greedy receiver-causal (Ours)	$+0.11$ $[-0.29, +0.52]$	-0.09 $[-0.45, +0.35]$	$+0.08$ $[-0.26, +0.46]$
Top-FVE	$+0.08$ $[-0.36, +0.51]$	-0.15 $[-0.51, +0.31]$	-0.04 $[-0.36, +0.34]$
random (within pool)	$+0.09$ $[-0.35, +0.53]$	-0.06 $[-0.42, +0.37]$	-0.05 $[-0.38, +0.34]$
freq-matched non-causal	$+0.03$ $[-0.38, +0.46]$	-0.12 $[-0.49, +0.32]$	$+0.09$ $[-0.26, +0.45]$

Table 1: Per-method Spearman $\rho(\mathrm{FVE}, L_{\mathrm{g}})$ on D_{test} at $r = 5\%$ with 500-bootstrap 95% CIs.

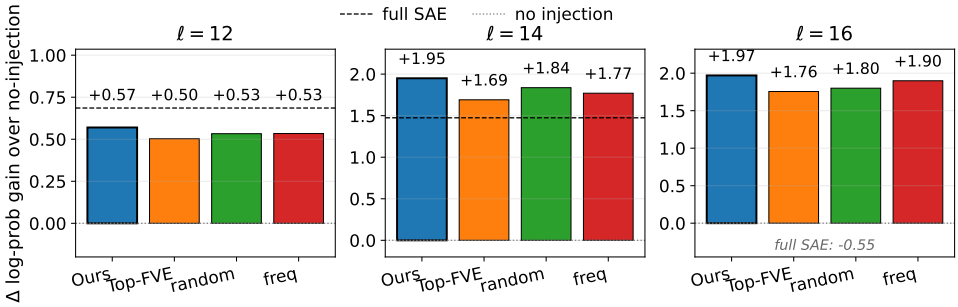


Figure 2: Per-method gold-token log-prob gain over no-injection at the headline ratio $r = 5\%$ ($K = 10$ features), one panel per core-cell inject layer. Dashed black: full-SAE injection.

where $\rho_{j,\mathcal{F}} = \max_{i \in \mathcal{F}} \left|\cos\big(W_{\mathrm{dec}}[j], W_{\mathrm{dec}}[i]\big)\right|$ is the largest absolute cosine similarity between feature j's decoder direction and any already-selected feature, and $\alpha \geq 0$ controls the redundancy penalty ($\alpha = 0$ recovers marginal greedy, $\alpha \to \infty$ enforces hard orthogonality). We use $\alpha = 1$ throughout.[3]

The marginal version is our default method; the sequential redundancy-aware version is an exploratory extension evaluated on the same locked test fold.

Candidate pool and K. We restrict j to the top-200 features by mean absolute activation on $D_{\mathrm{discovery}}$ and set $K = 10$ at the headline compression $r = 5\%$; both are pragmatic choices. Scoring all $d_{\mathrm{sae}} \in [16\mathrm{K}, 131\mathrm{K}]$ features via Eq. 4 is prohibitive, and the pool is shared by every baseline so the comparison is fair within it. The headline pattern is stable across the pool and K choices we tested: pool size in $\{100, 200, 400\}$ (Appendix F, Table 12), $K \in \{5, \ldots, 50\}$ across two independent sweeps (Tables 12 and 7), and random K=10 drawn from the full SAE dictionary with no activation filter (Table 11). Low-frequency but highly causal features remain outside the pool by construction, which we treat as a limitation.

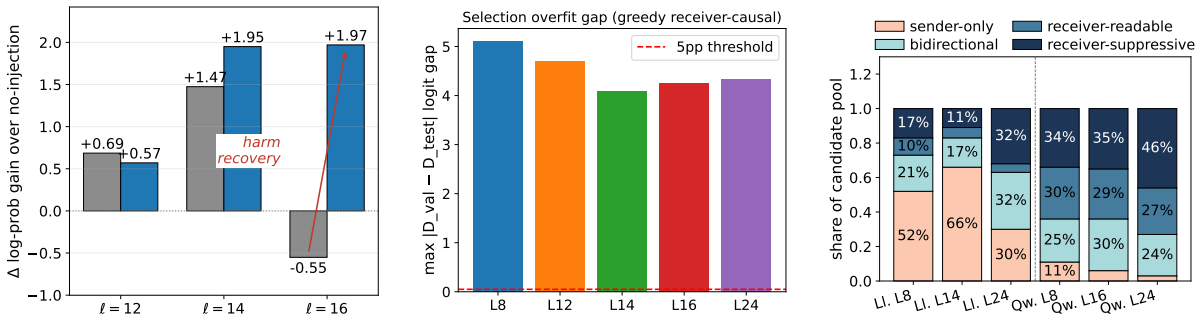
Negative baselines. We compare receiver-causal selection against four baselines. Three draw from the same candidate pool: *Top-FVE*, the natural FVE-driven proxy, picks the top-K features by single-feature contribution to $\mathrm{FVE}(h_s)$; *random within pool* samples a uniformly random size-K subset from the pool; *frequency-matched non-causal* keeps the highest-activation features that fall below the median causal effect on $D_{\mathrm{selection}}$ (matches activation magnitude while contradicting the receiver-causal signal). A fourth baseline, *norm-matched full SAE*, isolates direction selection from injected magnitude: we scale the full-SAE reconstruction so its decoded norm matches the receiver-causal K-feature subset's, then inject. All four baselines reuse the same SAE, ratios, inject layer, and test fold; the first three additionally reuse the pool.

[3]The cosine matrix is computed once over fixed SAE decoder rows; the sequential pick costs $O(K \cdot N_{\mathrm{pool}})$ table lookups with no extra forward passes.

4.2 Results

Table 2 reports the receiver's gold-token log-prob gain over the no-injection baseline at $K = 10$, with full-SAE injection as the natural reconstructed reference. The SAE-level bandwidth reduction spans roughly three orders of magnitude across the four SAEs. Figure 3a shows the bandwidth–quality curve across compression ratios.

Headline reading: bandwidth and norm dominate over identity. Table 2 shows the $K = 10$ subset matching or beating the full-SAE injection at every layer. A five-layer six-method ablation in Appendix F separates this gain into three components by adding two controls: *norm-matched full SAE* (full reconstruction rescaled per task to



(a) Headline comparison. (b) Selection-overfit gap. (c) Role distribution.

Figure 3: Result-side panels referenced across §4–§5. **(a)** Headline subset-vs-full on HotpotQA (Llama-3-8B + EleutherAI SAE): full SAE (gray) vs. $K = 10$ receiver-causal subset (blue). **(b)** $D_{\text{val}} - D_{\text{test}}$ overfit gap across compression ratios; dashed red at the 5-log-prob budget. **(c)** Per-(model, layer) share of the top-100 candidate features across the four roles; 8-cell breakdown in Table 4.

ℓ	full SAE	greedy r-causal	Δ(ours − full)
8	−1.40	+0.12	+1.52
12	+0.69	+0.57	−0.12
14	+1.47	+1.95	+0.48
16	−0.55	+1.97	+2.52
24	−0.74	+0.24	+0.98
mean	−0.11	+0.97	+1.08

Table 2: Receiver gold-token log-prob gain over no-injection at $r = 5\%$ on D_{test} (HotpotQA, Llama-3-8B + EleutherAI SAE). Short-form benchmark runs use Llama-3.1-8B (Table 3).

Ours's $K = 10$ decoded norm) and uninformed $K = 10$ selection from the full SAE dictionary. *(i) Magnitude carries most of the full-SAE harm*: at the harm-recovery layers, scaling the full-SAE reconstruction down to Ours's $K = 10$ norm moves the gain from $\{-1.45, -1.59, -1.19\}$ to $\{+0.01, +0.42, -0.11\}$ at $\ell = 8, 16, 24$. *(ii) $K = 10$ concentration adds to this at mid depth*: paired Ours − norm-matched full SAE is $+0.43, +1.15, +1.48$ at $\ell = 12, 14, 16$ (95% CIs exclude zero) and within noise at $\ell = 8, 24$. *(iii) Among $K = 10$ subsets, no detectable ranking advantage at this N*: paired Ours − random_unrestricted 95% CI crosses zero at every layer. Receiver-causal selection picks features no worse than random at $N_{\text{test}} = 25$; a larger held-out fold is needed to detect a positive ranking advantage if one exists.

Selection transfers within the held-out margin. We adopt a 5 log-prob selection-overfit budget: the gap between gold-token log-prob on $D_{\text{validation}}$ and on D_{test} should be smaller than the smallest method advantage in Table 2. The observed gap stays below this budget at every compression ratio (Figure 3b). Doubling the test fold at the Latent-MAS inject layer leaves every claim in the negative zone on both HotpotQA and MATH.

5 Cross-Model and Cross-Task Behavior

The decoupling and the minimal-subset advantage are reported above on Llama-3.1-8B HotpotQA. We now ask how the protocol behaves on the other (model, benchmark) corners. The same data split, the same candidate pool, and the same selection procedure are used; only the model, benchmark, and inject layer change.

5.1 Experimental Settings

Tasks and backbones. We evaluate the full 4×6 matrix of {Llama-3-8B-Instruct, Qwen3-8B, Qwen3-1.7B-Base, Llama-3.2-1B-Instruct} $\times$ {HotpotQA, MATH, ARC-Easy, ARC-Challenge, GSM8K, MedQA-USMLE}, where the Llama-3-8B checkpoint pairs with EleutherAI SAE for HotpotQA / MATH and the Llama-3.1-8B-Instruct checkpoint pairs with the same SAE on the four short-form benchmarks. For each (model, benchmark) cell we report the mean Spearman ρ between FVE and the receiver's gold-token log-probability, and the receiver's real end-task pass rate under no-injection, under the full-SAE injection, and under the $K = 10$ receiver-causal subset at $r = 5\%$.

Role-by-Functional taxonomy. To probe how the two 8B backbones use their candidate pools (the taxonomy is reported only at the 8B scale where SAE dictionary sizes match), we partition each pool by two axes. The *role* axis classifies a feature by the per-task mean, variability, and

Benchmark	Llama-3 8B†		Qwen3-8B		Qwen3-1.7B		Llama-3.2-1B	
	$\bar{\rho}$	$p_{\mathrm{f}}/p_{\mathrm{F}}/p_{\mathrm{O}}$	$\bar{\rho}$	$p_{\mathrm{f}}/p_{\mathrm{F}}/p_{\mathrm{O}}$	$\bar{\rho}$	$p_{\mathrm{f}}/p_{\mathrm{F}}/p_{\mathrm{O}}$	$\bar{\rho}$	$p_{\mathrm{f}}/p_{\mathrm{F}}/p_{\mathrm{O}}$
HotpotQA	$-.02$	.09/.11/.11	$-.17$	.32/.25/.25	$-.08$	.00/.00/.00	$-.03$	.00/.04/.03
MATH	$+.08$	.00/.00/.00	$+.14$	.18/.23/.18	$+.26$	.04/.03/.02	$-.21$	.04/.03/.02
ARC-Easy	$-.52$	.72/.38/.**76**	$-.26$	.96/.85/.84	$-.73$	.31/.22/.28	$-.27$	.34/.15/.**49**
ARC-Challenge	$-.26$	.72/.35/.**73**	$+.06$	.88/.73/.72	$-.74$	.36/.25/.33	$-.11$	.12/.09/.**20**
GSM8K	$+.02$	.16/.36/.23	$-.14$	.24/.29/.07	$-.26$	.12/.17/.13	$+.13$	.12/.21/.11
MedQA	$+.18$	.28/.29/.34	$+.10$	.64/.37/.**53**	$-.20$	.24/.21/.21	$-.03$	.24/.12/.**28**

Table 3: Per-(model, benchmark) means on D_{test}. $\bar{\rho}$: mean Spearman correlation between FVE and the receiver's gold-token log-probability. $p_{\mathrm{f}}/p_{\mathrm{F}}/p_{\mathrm{O}}$: pass rate at $r = 5\%$ under no-injection / full-SAE / $K = 10$ receiver-causal subset; leading zeros dropped. Bold marks cells where $p_{\mathrm{O}} > p_{\mathrm{F}}$ and $p_{\mathrm{F}} < p_{\mathrm{f}}$ (subset beats full SAE and recovers a full-SAE harm). †Llama-3-8B-Instruct on HotpotQA/MATH; Llama-3.1-8B-Instruct on the four short-form benchmarks.

sign-consistency of its signed causal effect on the receiver's gold-token log-probability into four levels (*sender-only*, *receiver-readable*, *receiver-suppressive*, and *bidirectional*), and the *function* axis classifies by whether HotpotQA or MATH gives the larger mean $|e_j^{\mathrm{sgn}}|$ (*context-binding* vs. *task-reasoning*). This yields an 8-cell breakdown per (model, layer) point. Full thresholds, decision rules, and a five-setting sensitivity sweep are in Appendix G; the mid-depth Llama-vs-Qwen gap in sender-only share survives every threshold setting we tried, while the shallow and deep gaps are threshold-dependent.

5.2 Results

Decoupling holds in every (model, benchmark) pair across two scales. Table 3 reports $\bar{\rho}$ per (model, benchmark) for four backbones spanning two series and two scales. Twenty-one of the twenty-four cells stay below our 0.35 operational weak-correlation cutoff; the three exceptions sit on the multiple-choice ARC benchmarks for Llama-3.1-8B and Qwen3-1.7B, where (as discussed in §3.3) the small candidate set and prompt-dominated logits compress both faithfulness signals so that the rank correspondence reads stronger than the underlying signal would justify. Decoupling strengthens at the smaller scale on the easier non-math benchmarks, in the direction expected when a smaller backbone's SAE captures more sender-side variance that the receiver does not read.

The $K = 10$ advantage holds across models, except on math. The pattern that motivated the small subset (full SAE harms the receiver, $K = 10$ at the right total norm restores it, §4.2) replicates on Llama-3.2-1B MedQA across every layer we test, ruling out a Llama-3.1-8B layer-16 artifact. The one boundary it does not cross is mathematical reasoning: MATH and GSM8K are where the full SAE wins across most backbones (Table 3). The simplest reading is that math depends on residual content beyond what ten features at the controlled norm can carry; we treat this as a conjecture.

The two 8B backbones organize their channels differently at mid depth. Figure 3c shows the role distribution of the top-100 candidate features per (model/layer). At mid depth, Llama-3.1-8B's pool is dominated by sender-only features that the SAE reconstructs cleanly but the receiver never reads; Qwen3-8B at the same normalized depth carries almost no sender-only band and routes most of its pool through features the receiver consumes or actively suppresses. To decide when two role distributions differ meaningfully we adopt a 15% gap on any cell of the 8-cell taxonomy as the cross-model consistency threshold; below this, the two distributions agree within sampling noise at the per-cell test size we use. By this criterion the two pools differ at the shallow and mid layer pairs and sit on the boundary at the deep pair (Table 4). The functional partition is model-specific, while the receiver-causal subset itself transfers across models (Table 3).

Redundancy-aware selection helps where the SAE pool is dense. The sequential-greedy variant from §4.1 improves on the marginal version when SAE features overlap in decoder space. Across the Qwen3 (benchmark, layer, ratio) cells from the four short-form benchmarks, the redundancy-aware version raises the receiver's gold-token log-prob on average and wins about half the cells. The effect is sharply K-dependent: it helps most at the smallest compression and washes out at the largest, where the budget is wide enough

		snd-only		bidir.		rcv-rd.		rcv-sp.	
pair	series	ctx	tsk	ctx	tsk	ctx	tsk	ctx	tsk
L8 vs L8	Ll.	.23	.29	.10	.11	.04	.06	.09	.08
	Qw.	.04	.07	.17	.08	.15	.15	.21	.13
	$\lvert\Delta\rvert$	.19	**.22**	.07	.03	.11	.09	.12	.05
L14 vs L16	Ll.	.40	.26	.14	.03	.01	.05	.05	.06
	Qw.	.04	.02	.17	.13	.15	.14	.13	.22
	$\lvert\Delta\rvert$	**.36**	.24	.03	.10	.14	.09	.08	.16
L24 vs L24	Ll.	.14	.16	.15	.18	.02	.03	.11	.21
	Qw.	.02	.01	.21	.03	.13	.14	.20	.26
	$\lvert\Delta\rvert$	.12	**.15**	.06	**.15**	.11	.11	.09	.05

Table 4: Role $\times$ Functional taxonomy of the top-100 candidate features on Llama-3.1-8B (Ll.) and Qwen3-8B (Qw.) at three layer pairs. Columns: role $\times$ function (*ctx*: context-binding, *tsk*: task-reasoning); cells are pool proportions in $[0, 1]$ with leading zeros dropped (*.23* = 0.23). Bold marks the row-max $\lvert\Delta\rvert$.

that the redundancy penalty no longer changes which features get selected. The gain on Llama is much smaller, consistent with a property of the SAE rather than of the procedure: EleutherAI SAE's larger per-token sparsity budget already places near-orthogonal features at the top of the activation pool, while Qwen-Scope's tighter budget admits more decoder-direction overlap.

6 Related Work

Latent inter-agent communication and multi-agent interpretability. A growing line of work replaces token-level messages between collaborating LLM agents with continuous representations: LatentMAS (Zou et al., 2025) passes last-layer latent thoughts through a shared working memory, SDE (Tang et al., 2025) augments natural-language tokens with state-transition trajectories, and Interlat (Du et al., 2025) learns a compressed code with a receiver-side adapter. All three are evaluated by end-task accuracy or by sender-side reconstruction (e.g. FVE of an SAE on the sender's residual), leaving open whether the directions a sender-side decomposition rewards are the directions the receiver actually reads, which is the gap we close. Concurrent multi-agent interpretability is complementary: Lee et al. (2025) outline an evaluation/alignment agenda, and Tailor (2025) study covert channels and collusion; neither measures receiver-side, feature-level causal effects on a held-out fold.

Sparse autoencoders and activation patching. We build on two standard single-agent threads. Sparse autoencoders decompose the residual stream into an approximately monosemantic dictionary (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024; Marks et al., 2025), with pre-trained suites available for our backbones (EleutherAI, 2024; He et al., 2024; Deng et al., 2026; Chanin, 2024), so we train no new SAE. Activation patching localises the representations causally responsible for a prediction (Meng et al., 2022; Wang et al., 2023; Heimersheim and Nanda, 2024). *Our modifications to this machinery are twofold: (i) we score each SAE feature by the drop in the receiver's gold-token log-probability under single-feature zero-ablation before patching into the receiver's residual, yielding a per-feature receiver-causal rank; and (ii) we estimate the rank of the receiver's effective readout via a random-probe Jacobian SVD around the patched activation.* To our knowledge, this is the first measurement of per-feature receiver-causal effects in an inter-agent latent channel.

7 Conclusion

We split reconstruction faithfulness (sender-side, geometric) from communication faithfulness (receiver-side, causal), and rank SAE features by their effect on the receiver's gold-token log-probability. Four phenomena emerge: *(P1) Decoupling*: sender-side FVE carries essentially no information about whether the receiver can use the signal; *(P2) Low-rank readout*: a small receiver-causal subset matches or exceeds full-SAE injection, and adding more features yields flat returns; *(P3) Bandwidth and norm, not identity, drive the receiver*: full-SAE injection often drops the receiver below the no-injection floor, and the $K{=}10$ subset repairs the harm mostly via the injected norm rather than the identity of the chosen features; at fixed bandwidth, our receiver-causal ranking is not detectably better than uninformed selection on our test folds; *(P4) Architecture-portable, mechanism-specific*: both phenomena hold for all four backbones (Llama-3 and Qwen3 at 1B and 8B), but the features that carry the signal differ between Llama and Qwen at every depth-matched layer pair. These suggest that latent-communication systems should report a receiver-side causal metric alongside FVE, control injection magnitude before attributing harm to wrong directions, and transmit at a bandwidth on the order of the receiver's effective readout rank.

Limitations

The same mechanism that transmits a 10-feature subset also supports steganographic compression in a low-bandwidth latent channel; we did not optimise for covert-channel capacity or evaluate against adversarial monitors, and leave that axis to follow-up work. ChatGPT was used to assist with polishing the manuscript.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

David Chanin. 2024. ReLU residual-stream sparse autoencoders for Llama-3.2-1B. Hugging Face repository: `chanind/sae-llama-3.2-1b-res`.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. 2026. Qwen-Scope: Turning sparse features into development tools for large language models. *Preprint*, arXiv:2605.11887. Hugging Face checkpoint: `https://huggingface.co/Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_50`; residual-stream TopK SAE for Qwen3-8B-Base, $d_{\text{sae}} = 65{,}536$, $k = 50$.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR. ArXiv:2305.14325.

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao, Xiaoyong Zhu, Yu Cheng, Bo Zheng, Wei Chen, and Haochao Ying. 2025. Enabling agents to communicate entirely in latent space. *Preprint*, arXiv:2511.09149. Accepted to ACL 2026.

EleutherAI. 2024. Open-source residual sparse autoencoders for Llama-3-8B (sae-llama-3-8b-32x), k=192, d_{sae}=131,072. Hugging Face repository: EleutherAI/sae-llama-3-8b-32x.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *Preprint*, arXiv:2410.20526.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *Preprint*, arXiv:2404.15255.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jae Hee Lee, Anne Lauscher, and Stefano V. Albrecht. 2025. Towards ethical multi-agent systems of large language models: A mechanistic interpretability perspective. *Preprint*, arXiv:2512.04691. Accepted to LaMAS 2026@AAAI'26.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics. ArXiv:2305.19118.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse

feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Om Tailor. 2025. Audit the whisper: Detecting steganographic collusion in multi-agent LLMs. *Preprint*, arXiv:2510.04303.

Yichen Tang, Weihang Su, Yujia Zhou, Yiqun Liu, Min Zhang, Shaoping Ma, and Qingyao Ai. 2025. Augmenting multi-agent communication with state delta trajectory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10219–10240, Suzhou, China. Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

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

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. 2025. Latent collaboration in multi-agent systems. *Preprint*, arXiv:2511.20639.

A Per-Method Spearman ρ Across All Five Core Layers

Table 1 in the main paper reports the per-method Spearman correlation at three core-cell inject layers ($\ell = 12, 14, 16$). Table 5 below extends that breakdown to all five core layers used in §4 ($\ell \in \{8, 12, 14, 16, 24\}$). Every point estimate satisfies $|\rho| \leq 0.15$ and every 95% CI spans both sides of 0 at all five layers; the matrix-wide weak-correlation conclusion is therefore not a property of the three core-cell layers in particular.

method	ℓ=8	ℓ=12	ℓ=14	ℓ=16	ℓ=24
greedy (Ours)	+0.15	+0.11	−0.09	+0.08	−0.00
Top-FVE	+0.15	+0.08	−0.15	−0.04	+0.02
random	+0.03	+0.09	−0.06	−0.05	−0.04
freq	+0.13	+0.03	−0.12	+0.09	−0.01

Table 5: Per-method Spearman $\rho(\mathrm{FVE}, L_{\mathrm{g}})$ on D_{test} at $r = 5\%$ across all five core inject layers on Llama-3-8B-Instruct + EleutherAI SAE (HotpotQA). 500-bootstrap 95% CIs (suppressed for compactness) span both sides of 0 in every cell; full intervals are in `phase_E/results_layer{8,12,14,16,24}.json`.

B Compression-Ratio Sweep Across All Five Core Layers

The headline numbers in Table 2 fix the compression ratio at $r = 5\%$ ($K = 10$ features out of 200 candidates). To check that the subset's advantage over the full-SAE injection is not a property of one particular K, Table 6 sweeps r across the four operating points we evaluated, at each of the five core inject layers on Llama-3-8B-Instruct + EleutherAI SAE (HotpotQA). The Ours-minus-full gap is roughly ratio-independent at every layer: it tracks the full-SAE harm-recovery offset at the shallow / mid / deep harm layers ($\ell \in \{8, 16, 24\}$) and stays small but non-negative at the cooperative layers ($\ell \in \{12, 14\}$). This is consistent with the receiver-readout-rank reading of §3.2: once the budget K exceeds the receiver's effective readout rank, adding more features mostly pads receiver-irrelevant directions.

Per-method K sensitivity at $\ell = 14$. Table 6 shows the Ours-minus-full gap is K-stable, but not whether the other K-feature methods co-move. Table 7 reports all four K-feature methods at $K \in \{10, 20, 30, 50\}$ at the LatentMAS-default layer on the seed-7 fold ($N_{\mathrm{test}} = 50$). Across the four-fold K range every method stays within 0.6 log-prob of every other; no method collapses and no method pulls away.

ℓ	full SAE	r=5%	r=10%	r=15%	r=25%
8	-1.40	$+1.52$	$+1.47$	$+1.50$	$+1.49$
12	$+0.69$	-0.16	-0.18	-0.16	-0.17
14	$+1.47$	$+0.29$	$+0.29$	$+0.27$	$+0.23$
16	-0.55	$+2.32$	$+2.27$	$+2.28$	$+2.32$
24	-0.74	$+0.98$	$+1.18$	$+1.17$	$+1.24$

Table 6: Compression-ratio sweep on Llama-3-8B-Instruct + EleutherAI SAE (HotpotQA, D_{test}). *full SAE* column: full-SAE injection's log-prob gain over no-injection. r=$\cdot\cdot$ columns: receiver-causal subset's gain *minus* the full-SAE gain at compression r ($K = 10, 20, 30, 50$ from left to right). Positive entries: subset beats full SAE; the gap is approximately ratio-independent at every layer.

method	K=10	K=20	K=30	K=50
Ours (greedy receiver-causal)	$+1.57$	$+1.36$	$+1.28$	$+1.23$
Top-FVE	$+1.33$	$+1.35$	$+1.31$	$+1.32$
random (within pool)	$+1.80$	$+1.75$	$+1.74$	$+1.62$
freq-matched non-causal	$+1.61$	$+1.67$	$+1.56$	$+1.48$
full SAE	$+0.27$ (independent of K)			

Table 7: K sensitivity at $\ell = 14$ on Llama-3-8B-Instruct + EleutherAI SAE (HotpotQA, D_{test}, $N_{\text{test}} = 50$, seed 7). Mean receiver gold-token log-prob gain over no-injection. Combined with the pool-size sweep (Table 12, $K \in \{5, 10, 20\}$) and the unrestricted-random control (Table 11), the headline subset-vs-full pattern is stable for $K \in \{5, \ldots, 50\}$ and across pool restrictions, so $K = 10$ is not a cherry-picked operating point.

C Redundancy-Aware vs. Marginal Greedy

Section 5 reports that the sequential, redundancy-aware variant of the selection (Eq. 5) improves on the marginal version by $+0.11$ receiver log-prob on average across the 48 Qwen3-8B short-form cells, with a much smaller gain on Llama-3.1-8B. Table 8 reports the per-cell breakdown that backs those numbers: for every (benchmark, layer, ratio) cell across the four short-form benchmarks (ARC-Easy, ARC-Challenge, GSM8K, MedQA), we evaluate both selectors on the same locked test fold and compare gold-token log-prob.

D Receiver-Readout Effective Rank per Task

The heuristic motivation in §3.2 relies on the receiver's Jacobian being low-rank at the inject layer. We measure $r_{\text{eff}} = |\{i : \sigma_i(J_{\text{recv}}) > 0.01 \cdot \sigma_{\max}\}|$ on five $D_{\text{discovery}}$ tasks per layer at a 32-probe budget, summarised in Table 9. At every layer except the shallowest, the probe budget saturates ($r_{\text{eff}} =$

backbone (SAE)	cells	DPP wins	mean Δ
Qwen3-8B (Qwen-Scope)	48	$25/48\,(52\%)$	$+0.11$
Llama-3.1-8B (EleutherAI)	80	$26/80\,(32\%)$	$+0.01$

Table 8: Redundancy-aware sequential selection vs. marginal greedy on the four short-form benchmarks. *cells*: number of (benchmark, layer, ratio) cells. *DPP wins*: cells where the sequential variant has a higher receiver gold-token log-prob. *mean* Δ: mean log-prob gain of the sequential variant over marginal greedy (cell-averaged). The redundancy penalty helps on Qwen-Scope (tighter sparsity budget, more decoder-direction overlap) and washes out on EleutherAI SAE (wider per-token sparsity already places near-orthogonal features at the top of the activation pool).

32) and the recorded spectrum decays smoothly past the cap, so the reported numbers are probe-cap lower bounds consistent with $r_{\text{eff}} \ll d = 4096$.

ℓ	r_{eff} per task	$\sigma_{\max}$ range
8	$23, 17, 27, 22, 12$	121–208
12	$32, 32, 32, 32, 32$	50–67
14	$32, 32, 32, 32, 32$	46–82
16	$32, 32, 32, 32, 32$	28–63
24	$32, 32, 32, 32, 32$	17–18

Table 9: Receiver-readout effective rank r_{eff} at five core inject layers on Llama-3-8B-Instruct. Each layer is measured on five $D_{\text{discovery}}$ tasks with a 32-probe budget; values at 32 are probe-cap-saturated lower bounds. The shallowest layer is below the cap on every task, the remaining four are all at the cap, and the recorded singular spectrum decays smoothly past the budget, consistent with $r_{\text{eff}} \ll d = 4096$ throughout.

E Seed Robustness

Table 10 re-runs the LatentMAS-default inject layer ($\ell = 14$) at the headline ratio $r = 5\%$ under four different split seeds, on HotpotQA and MATH with Llama-3-8B + EleutherAI SAE. Only the train / validation / test split changes; the candidate pool, the receiver-causal selection procedure, and N_{test} are held fixed. The receiver-causal subset's advantage over the full-SAE injection stays positive on every (seed, task) cell, and the rank correlation between FVE and the receiver's gold-token log-probability stays well below the $|\rho| = 0.35$ weak-correlation cutoff throughout.

F Bandwidth, Norm, and Selection: A Five-Layer Decomposition

The headline subset-vs-full comparison in Table 2 mixes three confounded effects: how many fea-

seed	HotpotQA ρ	HotpotQA $\Delta_{\mathrm{g-f}}$	MATH ρ	MATH $\Delta_{\mathrm{g-f}}$
7	$+0.13$	$+1.30$	$+0.19$	$+0.21$
42	$+0.13$	$+1.30$	$+0.15$	$+0.39$
123	$+0.13$	$+1.30$	$+0.25$	$+0.52$
2024	$+0.13$	$+1.30$	$+0.22$	$+0.48$
mean	$+0.13$	$+1.30$	$+0.20$	$+0.40$

Table 10: Seed robustness at $\ell = 14$, $r = 5\%$ on Llama-3-8B-Instruct + EleutherAI SAE. ρ: Spearman rank correlation between FVE and the receiver's gold-token log-probability on D_{test}. $\Delta_{\mathrm{g-f}}$: receiver-causal subset ($K = 10$) gain over the full-SAE injection, in log-prob units. HotpotQA results are identical across the four seeds because the seed-induced reshuffle leaves the realised $N_{\mathrm{test}} = 28$ fold and the resulting candidate pool unchanged on this benchmark.

tures are transmitted ($K = 10$ versus d_{sae}), how much total magnitude is injected, and which specific features carry the signal. To separate them we add two controls and run the five-way comparison at every core inject layer ($\ell \in \{8, 12, 14, 16, 24\}$) on Llama-3-8B-Instruct + EleutherAI SAE, HotpotQA, D_{test}, $N_{\mathrm{test}} = 25$, seed 7.

The controls are: *random within pool* (uniform random $K = 10$ from the top-200 activation pool); *random unrestricted* (uniform random $K = 10$ from the full SAE dictionary $d_{\mathrm{sae}} = 131{,}072$, no activation filter); and *norm-matched full SAE* (the full-SAE reconstruction scaled per task so its decoded norm equals the receiver-causal $K = 10$ subset's). Together with *Top-FVE* and *frequency-matched non-causal* this gives a six-method panel at each layer.

Reading Table 11. The decomposition supporting §4.2's three-component reading: (i) the magnitude correction is visible in the full-SAE vs. norm-matched columns at the harm layers ($\ell \in \{8, 16, 24\}$: $-1.45/-1.59/-1.19$ becomes $+0.01/+0.42/-0.11$); (ii) the concentration gain is the Ours $-$ norm-matched difference at mid depth (paired 95% CIs $+0.43\,[+0.09, +0.78]$, $+1.15\,[+0.81, +1.49]$, $+1.48\,[+1.00, +1.96]$ at $\ell = 12, 14, 16$; CIs cross zero at $\ell = 8, 24$); (iii) no detectable ranking advantage among the five $K = 10$ methods at this N (paired Ours $-$ random_unrestricted 95% CI crosses zero at every layer).

Pool-size sensitivity at $\ell = 14$. To test whether the bandwidth-effect reading depends on the specific top-200 pool, we re-run the comparison with pool sizes 100 and 400 (K scales with the pool at $r = 5\%$: $K \in \{5, 10, 20\}$). Table 12 shows that at $\ell = 14$ neither the pool size nor K in the range 5–20 moves the gain meaningfully: every method, at every pool size, stays in the $+1.83$ to $+1.97$ band against an unchanged full-SAE gain of $+1.13$, with 95% CIs overlapping fully. The receiver's insensitivity to which small subset is transmitted at the LatentMAS-default layer is robust to the pool-size choice.

G Taxonomy Threshold Sensitivity

Classifier definition. For feature j in the top-100 pool, let μ_j, σ_j be the per-task mean and standard deviation of the signed causal effect on the receiver's gold-token log-probability (the signed analogue of Eq. 4), and define $s_j = |\mu_j|/\sigma_j$. With causal threshold τ, variability threshold σ_{th}, and activation floor a_{floor}, roles are assigned in order: *sender-only* when $|\mu_j| < \tau$ and $\sigma_j < \sigma_{\mathrm{th}}$; *receiver-readable* when $\mu_j > +\tau$ and $s_j > 1$; *receiver-suppressive* when $\mu_j < -\tau$ and $s_j > 1$; *bidirectional* when $|\mu_j| < \tau$ and $\sigma_j \geq \sigma_{\mathrm{th}}$; features with max activation below a_{floor} are dropped. The function axis splits *context-binding* from *task-reasoning* by which task type (HotpotQA vs. MATH) gives the larger mean $|e_j^{\mathrm{sgn}}|$. We use $\tau = 0.02$, $\sigma_{\mathrm{th}} = 0.05$, $a_{\mathrm{floor}} = 0.05$ in the main text.

Sensitivity sweep. To check whether the "Llama-3.1-8B sender-only band vs. Qwen3-8B receiver-routed band at mid depth" interpretation is stable, we re-classify the cached per-feature causal effects across five threshold settings and report the max cross-model role-proportion gap at each of the three depth-matched layer pairs. The activation floor never moves the gap (only which features are dropped before classification), so we vary only τ and σ_{th}.

Interpretation. The qualitative claim that "Llama-3.1-8B and Qwen3-8B organise their candidate pools differently at mid depth" (and the bolded row-max $|\Delta|$ for L14 vs L16 in Table 4) is robust across all five threshold settings we tried. The sender-only share for Llama-3.1-8B at mid depth ranges from 21% (tightest $\tau, \sigma_{\mathrm{th}}$) to 85% (loosest), while Qwen3-8B's stays in the 1–23% range; sign of the Llama-vs-Qwen gap never flips. The shallow ($\ell = 8$) and deep ($\ell = 24$) layer-pair gaps are less robust: tightening either threshold collapses both pairs below the 15 pp consistency

ℓ	full SAE	full SAE (norm-mat.)	Ours	Top-FVE	rand (pool)	rand (unr.)	freq
8	-1.45	$+0.01$	$+0.22$	$+0.26$	$+0.24$	$+0.27$	$+0.28$
12	$+0.25$	$+0.16$	$+0.60$	$+0.60$	$+0.63$	$+0.62$	$+0.62$
14	$+1.13$	$+0.71$	$+1.87$	$+1.75$	$+1.95$	$+1.94$	$+1.90$
16	-1.59	$+0.42$	$+1.89$	$+1.83$	$+1.94$	$+1.88$	$+1.88$
24	-1.19	-0.11	$+0.00$	$+0.07$	$+0.00$	$+0.02$	-0.05

Table 11: Receiver gold-token log-prob gain over no-injection at $r = 5\%$ ($K = 10$ for the five K-feature methods; the norm-matched full SAE uses all d_{sae} features scaled per task to match Ours's $K = 10$ decoded norm). Llama-3-8B-Instruct + EleutherAI SAE, HotpotQA, $N_{\text{test}} = 25$, seed 7.

selection method	gain	pool=100 K=5	pool=200 K=10	pool=400 K=20
greedy receiver-causal		$+1.91$	$+1.87$	$+1.83$
Top-FVE		$+1.83$	$+1.75$	$+1.85$
random (pool-restricted)		$+1.95$	$+1.95$	$+1.97$
random (unrestricted)		$+1.94$	$+1.94$	$+1.94$
freq-matched non-causal		$+1.87$	$+1.90$	$+1.94$

Table 12: Pool-size and K sweep at $\ell = 14$, $r = 5\%$ on Llama-3-8B-Instruct + EleutherAI SAE (HotpotQA, D_{test}, $N_{\text{test}} = 25$, seed 7). Mean receiver gold-token log-prob gain over no-injection. The full-SAE baseline is the same in all columns ($+1.13$).

setting	τ	σ_{th}	L8	L14/L16	L24
default	0.020	0.050	41 pp	**60** pp	27 pp
τ halved	0.010	0.025	13 pp	**20** pp	0 pp
τ doubled	0.040	0.100	45 pp	**62** pp	62 pp
σ_{th} halved	0.020	0.025	13 pp	**21** pp	0 pp
σ_{th} doubled	0.020	0.100	42 pp	**66** pp	48 pp

Table 13: Maximum cross-model role-proportion gap (percentage points) under threshold variation, computed on the cached contrib_j tensors from `runs/taxonomy_8cell/{llama,qwen}_L*.json`. Bolded mid-depth column: the Llama-vs-Qwen gap exceeds the 15 pp cross-model consistency threshold of §5 at every setting we tried; the shallow and deep gaps drop below 15 pp under the tightest settings ($\tau = 0.01$ or $\sigma_{\text{th}} = 0.025$).

threshold, consistent with the main text's reading that the shallow gap is suggestive and the deep gap sits on the boundary.