LATENT COMMUNICATION BETWEEN LLM AGENTS: CHANNELS, ALIGNMENT, AND THE LIMITS OF TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent LLM systems typically communicate via text, but language models may encode semantic structure that resists faithful serialisation into tokens. We compare dense latent, Sparse Autoencoder (SAE), and text channels between Llama 3.1 8B and Mistral 7B. A naive evaluation protocol that draws alignment anchors and task queries from the same concept pool creates substantial overlap between them, so we instead use alias-aware grouped train, validation, and test partitions and select all alignment hyperparameters on validation concepts only. On three strict splits, PCA-Procrustes and regularised CCA both achieve 100% mean top-1 retrieval, compared with 89.3% for full-rank Procrustes, establishing a strong low-rank cross-architecture correspondence. This geometric result does not guarantee task utility. Text exceeds the best latent result on attribute/sense tasks by 21.7 percentage points (task-clustered 95% CI [8.8, 35.0]), while its 5.4-point advantage on concept identification is unresolved (95% CI [−7.9, 18.9]); cross-lingual coverage is insufficient for a stable comparison. A leave-one-prompt-out dense-probe diagnostic reaches 81.6% over 348 predictions, versus 49.8% under label permutation, but 100% training accuracy and mostly two prompts per concept limit its interpretation. Because channel comparisons can conflate modality with transmission budget and receiver compute, we also report bits transmitted and receiver forward passes per channel, and add a continuous, alignment-free top-k lexical-distribution baseline: at 67–125 bits it is competitive with text-passage on cross-lingual identification (65.5% vs. 62.1%) but clearly weaker on attribute/sense tasks (45.0% vs. 76.7%), showing that text's advantage is not simply an artifact of a larger budget. We conclude that low-rank concept geometry transfers reliably across these architectures, but concept-identity retrieval alone does not ensure task-conditioned communication.

1 INTRODUCTION

Multi-agent LLM systems typically communicate via natural language or structured text (e.g. JSON) — an approach that is human-readable, debuggable, and compatible with any model that understands language. We test the limits of this convenience: when an agent forms a high-dimensional internal representation of a concept, it must compress that representation into a token sequence before another agent can receive it, and nothing guarantees the receiving agent's reconstruction is equivalent, or even similar. Our contribution seeks to quantify how much is lost, whether the loss matters, and whether a better alternative exists.

A recent hypothesis suggests the loss may be structurally bounded: the Platonic Representation Hypothesis (Huh et al., 2024) proposes that diverse neural networks trained on sufficient data converge toward compatible representational geometries (latent world models) independently of architecture or training objective. If true, two LLMs should develop geometrically alignable activation spaces, potentially enabling direct latent-space communication that bypasses the text bottleneck entirely.

We construct and compare three communication channels — *dense latent* (the full activation vector), *SAE-sparse* (a sparse bottleneck via pre-trained Sparse Autoencoders), and *text* (the standard serialisation baseline) — and evaluate them on concept-identification and discrimination tasks, including cross-lingual variants designed to favour the architecture-agnostic structure of the latent channel.

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.

Our results support the geometric-convergence half of the Platonic hypothesis but not the practical communication advantage it might suggest. Under strict grouped evaluation, low-rank mappings retrieve held-out concepts nearly perfectly. Text is reliably stronger on attribute/sense tasks, while the strongest concept-identification and cross-lingual comparisons remain unresolved. Text serialisation replaces 88% of the active features under our operational SAE analysis, but the augmentation evidence only shows that these features do not help the tested additive cosine decoder; it does not establish that they are universally task-irrelevant or purely surface-form features. A preliminary result on reasoning tasks (§6) motivates richer evaluations.

Our experiments follow a gated sequence designed to isolate the contribution of each component. To start, semantic concepts must be linearly separable in the sender's activation space, and the SAE feature space must faithfully capture semantic relatedness (§4.1). We then evaluate three claims: **C1**, that the SAE-sparse channel preserves concept-discriminating information lost by text; **C2**, characterising what exactly text serialisation destroys at the feature level; and **C3**, that cross-architecture alignment enables latent communication between separately trained models. Task-level evaluation tests whether these representation-level differences translate to downstream performance, and text augmentation resolves the consequential question of whether the features text destroys carry suppressed semantic content.

2 Related work and own contributions

Representational and cross-lingual convergence. The Platonic Representation Hypothesis (Huh et al., 2024) proposes that neural networks with sufficient capacity converge toward a shared statistical model of the world; Cherti et al. (2023) provide complementary evidence from CLIP models, and Gromov et al. (2024) show deep transformer layers can be pruned with minimal degradation, suggesting representational geometry stabilises before the final layer. Conneau et al. (2020) and Pires et al. (2019) document a related, narrower form of convergence within a single model: multilingual pretraining produces cross-lingually aligned representations, emerging even without an explicit alignment objective. None of this work tests the direct implications for cross-architecture *communication* quantitatively; our alignment experiment supplies a hard threshold for exactly that. If a rotation fitted on a small anchor set generalises to unseen concepts, this is direct evidence of a shared world model. Our test of the convergence claims: rather than one model representing a concept consistently across languages, we test if two *different* architectures do so, under no joint training objective whatsoever.

Alignment methodology and sparse feature measurement. Lenc and Vedaldi (2015) and Bansal et al. (2021) showed internal representations of separately trained networks can be aligned by learned linear maps; Smith et al. (2017) first applied Orthogonal Procrustes to word embeddings for bilingual dictionary induction, and Conneau et al. (2018) removed the need for seed dictionaries via adversarial pre-training followed by Procrustes refinement. Our alignment is rotation-only, extending this embedding-level methodology to contextual LLM hidden states, with no shared task loss between sender and receiver and a lossless-communication rather than word-translation goal. Leveraging pre-existing geometry rather than learning a communication protocol is itself a departure from emergent-communication work (Kottur et al., 2017), where agents left to develop their own signalling produce effective but non-compositional protocols. Lester et al. (2021)'s soft-prompt steering of frozen LLMs is the closer analogue for our injection protocol. Our motivation for measuring *at the SAE feature level* is superposition (Elhage et al., 2022) — networks represent more features than dimensions via near-orthogonal encoding, which is why our SAE dictionaries are substantially overcomplete (131K features for 4,096-dimensional activations). Bricken et al. (2023), Cunningham et al. (2023), and Templeton et al. (2024) showed pre-trained SAEs factorise superposed representations into near-monosemantic, often abstract and multilingual features — the assumption our feature survival analysis relies on. We use SAEs as a measurement instrument rather than an interpretability end in itself: comparing sparse feature sets elicited by two encoding paths for the same concept isolates attenuation from outright replacement in a way dense-space cosine similarity cannot.

3 METHODS

We use Llama 3.1 8B-Instruct as the sender model and Mistral 7B (base v0.1 and instruction-tuned v0.3) as the receiver. Activations are extracted at approximately 72–75% of total depth: layer 23 of Llama and layer 24 of Mistral, further referred to as the "extraction layer". We tap the post-MLP residual stream and use the last-token hidden state in bf16 precision, yielding 4,096-dimensional vectors for both models. The SAE is release `llama-3.1-8b-instruct-andyrdt`, ID `resid_post_layer_23_trainer_1`, with $d_{\text{model}} = 4096$ and $d_{\text{SAE}} = 131072$.

Our concept dataset comprises 84 hand-curated concept pairs (165 unique labels) across four semantic categories: factual, abstract, overlap, and context-dependent. Each concept has two or more cloze-style prompts that elicit it in naturalistic context. English pairs constitute 76% of the pair dataset, supplemented by German, French, Spanish, and Mandarin Chinese examples. Separate task banks contain 165 concept-identification, 112 attribute/sense, and 107 cross-lingual rows. For strict evaluation, concepts, translations, base words, aliases, and parenthetical sense families are assigned together to deterministic 60/20/20 alignment-train, alignment-validation, and task-test partitions.

3.1 COMMUNICATION CHANNELS

Throughout this paper, *sender* is the LLM agent that holds a concept representation it wishes to communicate; *receiver* is the agent that must identify or act on that concept; *transmission* is the operation by which the sender's internal representation is converted into the format required by a given channel and delivered to the receiver. Bandwidth figures represent the minimum bits needed to faithfully represent the transmitted object, not end-to-end network throughput.

Dense latent channel. Transmits the sender's full residual-stream hidden state at layer 23 — a 4,096-dimensional bf16 vector resulting in 65,536 bits. This channel incurs zero information loss by construction and serves as the performance ceiling.

SAE-sparse channel. First compresses the sender's hidden state through a pre-trained SAE (Bricken et al., 2023) with 131,072 latent dimensions and $\sim$70 active features per forward pass. Only the indices and activation magnitudes of active features are transmitted — approximately 2,300 bits, a 28$\times$ compression relative to dense. Only the pre-trained SAE decoder needs to be shared between agents for decoding — not the full sender LLM.

Text channel. The standard inter-agent baseline. The sender serialises the concept as its name string; the Mistral receiver processes that string and extracts the resulting hidden state at layer 24. This channel requires no model alignment infrastructure.

The text channel as defined above transmits only the bare concept name (e.g., `"Photosynthesis"`), which is the minimal, most bandwidth-efficient text serialisation and the natural counterpart to the dense and SAE-sparse channels' single-vector transmission. This bare-name variant is what is reported as the "text channel" in the channel-fidelity and feature-survival results (§4.1). A richer variant that serialises a full descriptive passage instead of a bare name (*text-passage*) is introduced separately for the task-level evaluation in §3.3, where it consistently outperforms bare-name transmission by a wide margin (Table 5). This distinction matters for the C1 comparison: the 80.4% probe accuracy ceiling attributed to the text channel (detailed results in Table 1 in the Appendix) reflects the bare-name variant specifically, not an inherent ceiling of text as a modality. We return to this distinction in §5.

The channels differ in what is transmitted, but in all cases the representation used for evaluation is a 4,096-dimensional hidden-state vector, i.e. we do not evaluate actual LLM outputs against labels, but similarities of embeddings in latent space. For probe-based evaluation (§3.2) and task-level evaluation (§3.3), this vector is compared geometrically to reference vectors — no injection into a running model's forward pass occurs. The SAE-sparse channel differs from the dense channel only in that the vector passes through an encode–decode bottleneck: the sparse code (feature indices and magnitudes) is decoded back to a dense approximation via the SAE decoder matrix. The text channel is the only channel that re-encodes entirely: the concept string is processed by the receiver model from scratch, producing a new hidden state that is then used for comparison.

We separately evaluate generation injection within Llama. A contextual dense vector or its SAE reconstruction replaces a blend of the last hidden state of the neutral prompt `"The answer is"` at

layer 23, after which the model greedily generates 20 tokens. Whole concept pairs are deterministically assigned to 18 development and 66 test pairs. The blend coefficient $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ is selected independently per channel on development strict-hit accuracy and frozen before test evaluation. No injection, a norm-matched random vector, and a receiver vector elicited by the bare concept name are controls. The primary scorer requires token-boundary concept matches; a looser substring-matching criterion is reported separately as a scoring-sensitivity check (§4.1).

3.2 Channel Fidelity, Feature Survival, and Cross-Architecture Alignment (C1–C3)

Linear probes and channel fidelity (C1). We train logistic regression *linear probes* on ground-truth sender representations $\{\mathbf{h}_A^{(i)}\}$ and $\{\mathbf{h}_B^{(i)}\}$ and test them on versions transmitted through a given channel, to measure how well each channel preserves concept-discriminating structure. Fitting and scoring probes on transformations of the same prompt vectors gives 100.0%/99.4%/80.4% dense/SAE/text accuracy; we treat this only as an in-sample reconstruction diagnostic, not evidence of prompt-level generalisation. As a leakage check, we additionally use leave-one-prompt-out evaluation: in each fold, PCA and L_2-regularised logistic regression are fitted only on the remaining dense contextual vectors, and the held-out prompt vectors are scored once, with a label-permutation probe fitted within each fold as a negative control. Prompt-level SAE and text caches are unavailable, so this stricter diagnostic is dense-only.

SAE feature survival (C2). To characterise text-channel information loss at the feature level, we define a *text round-trip*: we elicit a concept via a contextual cloze-style prompt and, separately, via its bare name, then compare the two SAE encodings. We average the contextual hidden vectors for a concept, encode that mean through the SAE, and call every feature with activation greater than zero active, yielding a per-concept average of $\sim$43 active features out of 131,072; this is not a strict intersection across prompts, so we avoid calling these features "stable". The bare name is SAE-encoded the same way. For each concept we compute feature survival rate, Jaccard overlap, cosine similarity in SAE space, and Pearson correlation between original and text-channel feature magnitudes.

Cross-architecture alignment (C3). Activation spaces are coordinate systems: two models may encode the same concept along different directions in their respective 4,096-dimensional spaces, but nothing forces those directions to coincide. If the two spaces are rotated versions of each other, a rotation matrix can recover the correspondence from a modest number of paired concept vectors (*anchor* concepts); withheld concepts serve as the generalisation test. We compare full-rank Orthogonal Procrustes, PCA-reduced Procrustes, regularised CCA, and SVCCA. For each anchor concept, both models produce a mean concept vector, and full Procrustes finds the rotation $\mathbf{R}^* = \arg\min_{\mathbf{R}:\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \|\mathbf{X}\mathbf{R} - \mathbf{Y}\|_F$, where $\mathbf{X}$ and $\mathbf{Y}$ are the sender and receiver anchor matrices; the fitted rotation is applied to held-out concept vectors and retrieval accuracy is measured by nearest-neighbour lookup in the receiver's space. Our primary protocol uses three alias-aware grouped 60/20/20 splits (seeds 42–44); reduced rank, regularisation, and SVCCA variance thresholds are selected on alignment-validation concepts only, test concepts remain untouched until final retrieval and task evaluation, and full-rank Procrustes is retained as an underdetermined baseline.

3.3 Task-Level Evaluation, Augmentation, and Bandwidth Protocols

Evaluation mechanism. Probe accuracy (§3.2) measures discriminability with access to labelled representations, which a receiver does not have at inference time. We therefore also score channels with a multiple-choice (MC) protocol that mimics a realistic zero-shot setting: for each task, one concept is transmitted from sender to receiver through a given channel, producing a query vector $\mathbf{h}_{\text{query}}$ in the receiver's activation space (Procrustes-mapped for latent channels; re-encoded from scratch for the text channel). The receiver separately pre-caches a reference vector for each of four candidate concept names (bare name string, last-token hidden state at layer 24) and scores all four by cosine similarity to $\mathbf{h}_{\text{query}}$; the highest-scoring candidate is the answer. No joint forward pass over query and candidates occurs.

Only task rows whose grouped concept belongs to the task-test partition are evaluated, and no test concept or alias contributes to alignment fitting or hyperparameter selection. Concept-identification distractors are the paired concept plus two same-category concepts; attribute/sense distractors are

hand-authored. The bare-name query is identical to its correct candidate on concept-identification rows and therefore yields a tautological 100%, so we omit it as a meaningful baseline there. We report both seed-wise means and pooled hits/counts because grouped splits produce unequal numbers of eligible tasks.

Channels and baselines. All three channels from §3.1 are evaluated, with text split into two variants: (1) **dense latent**, the full sender hidden state, Procrustes-mapped; (2) **SAE latent**, the SAE encode–decode reconstruction, Procrustes-mapped; (3) **text-passage**, the receiver's own encoding of a full descriptive passage (e.g. "*The process by which plants convert sunlight into glucose is called Photosynthesis*"); (4) **text-name**, the receiver's encoding of only the bare concept name; and (5) **no-comm**, a baseline where no signal is transmitted and the receiver scores candidates from its cached reference vectors alone, expected to perform at chance. Significance is assessed via one-sided binomial test against the 25% chance baseline (four uniform candidates, no prior over the answer).

Text augmentation. The task-level protocol reveals a consistent 3–10 pp gap in which text-passage outperforms the latent channel. We test whether this reflects (a) geometric distortion from Procrustes alignment, which adding aligned sender information to the receiver's text representation should partially repair, or (b) genuinely additional task-relevant signal in the receiver's native text processing, in which case augmentation gives no benefit. The augmented representation is $\mathbf{h}_{\text{aug}} = \mathbf{h}_{\text{text}} + \alpha \cdot \mathbf{R}^* \mathbf{h}_{\text{sender}}$, with $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$ and $\mathbf{R}^*$ the alignment matrix from §3.2. We test four variants of $\mathbf{h}_{\text{sender}}$: the full dense hidden state; the full SAE reconstruction; the top-32 SAE features only; and the *lost features* active in contextual but not bare-name encoding.

Bandwidth, compute, and a continuous lexical channel. Comparing channels by accuracy alone conflates modality with computational budget. For every channel we additionally report bits transmitted ($\log_2 V$ bits per transmitted vocabulary token or feature index, V the sender's vocabulary size) and the receiver's marginal compute at query time: dense and SAE-sparse channels need zero receiver forward passes per query because candidates are precomputed and reused, while text and lexical channels require a fresh forward pass every query. We also add one continuous, architecture-portable baseline, the *top-k lexical-distribution channel*: the sender's top-32 next-token probabilities after the cloze passage are decoded to plain strings and independently embedded by the receiver, exactly as for the text channel, and the query vector is their probability-weighted sum (renormalised within the top-32 set). Because the receiver only ever embeds architecture-agnostic decoded strings, no Procrustes alignment is required. For each seed and task family, $k \in \{2, 4, 8, 16, 32\}$ is selected on alignment-validation tasks (maximising accuracy, ties toward smaller k) and frozen before scoring task-test concepts.

4 Results

4.1 Channel Fidelity, Feature Survival, and Generation Injection (C1–C2)

Before testing the main claims, we verify that semantic concepts are linearly separable in activation space and that related concepts have greater SAE feature overlap than unrelated pairs; an in-sample probe reaches 100% accuracy but is retained only as a separability diagnostic.

Channel fidelity (C1). In-sample dense/SAE/text probe accuracies of 100.0%, 99.4%, and 80.4% respectively are obtained by fitting on all prompt vectors and scoring transformed versions of those same prompts; they demonstrate reconstruction fidelity but cannot establish generalisation. A dense-only leave-one-prompt-out analysis instead makes 84 concept-pair probes produce 348 held-out predictions, achieving 81.6% micro accuracy versus 49.8% under within-fold label permutation. Mean fold training accuracy remains 100%, exposing the high-dimensional overfitting risk, and expected-overlap pairs are harder than unrelated pairs (70.5% vs. 92.1% macro pair accuracy). Because most concepts have only two prompts and matched prompt-level SAE/text caches are unavailable, we do not use this diagnostic to rank channels.

Feature survival (C2). Under the operational definition in §3.2, text round-tripping replaces most active features: the mean loss rate is 88.3% and mean SAE-space cosine is 0.187 (Table 3). These measurements compare an SAE encoding of the mean contextual vector with a bare-name encoding and do not identify a prompt-wise intersection; UMAP (Figure 1) is used only as a visualisation and

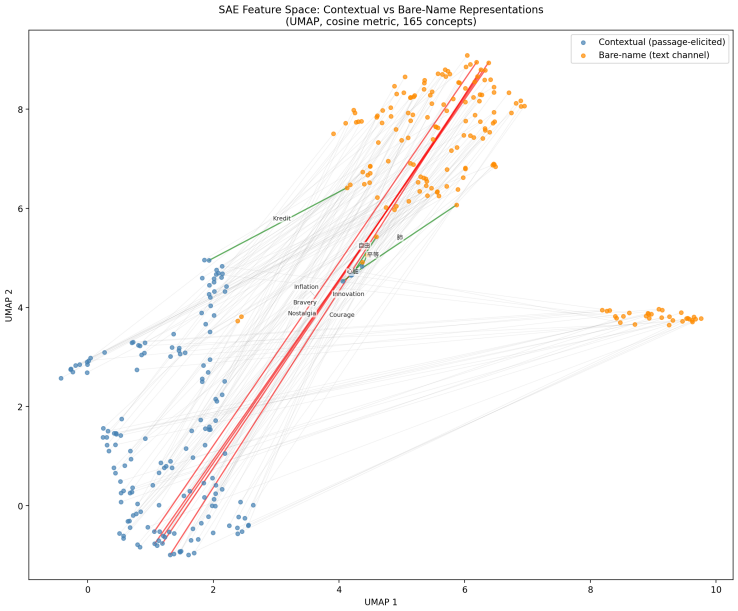


Figure 1: UMAP visualisation of SAE feature vectors; textual annotations indicate most/least displaced concepts.

does not establish distinct manifolds or feature semantics. Tokenisation, prompt structure, position, and contextual semantics are all plausible contributors to this replacement, but the present experiment does not disentangle them. Survival rate correlates modestly with the in-sample text-channel probe diagnostic (Pearson $r = 0.26$). Adding lost-feature reconstructions does not improve the tested additive geometric decoder (§4.3); this is evidence about that decoder, not proof that every lost feature is surface form or task-irrelevant in other interventions.

Held-out generation injection. Dense and SAE injection both select $\alpha = 0.75$ on 18 development pairs. On 66 untouched test pairs, dense injection yields 26 strict concept hits (39.4%) and SAE reconstruction yields 24 (36.4%); no injection, norm-matched random injection, and the bare-name receiver-vector control each yield zero. The channels share 23 hits (three dense-only, one SAE-only); their 3.0-point paired difference is unresolved (exact McNemar $p = 0.625$; paired-bootstrap 95% CI $[-3.0, 9.1]$). This decoder responds to contextual concept vectors, but the available sample does not establish a detectable SAE compression penalty. Scoring the same 84 SAE-only generations by substring match instead of token-boundary match raises the count from 26/84 (31.0%) to 33/84 (39.3%) — e.g. crediting `Heart` inside `heartwarming` — so we treat token-boundary matching as primary and report substring matching only as a scoring-sensitivity check.

4.2 Cross-Architecture Alignment and Task-Level Evaluation (C3)

Under a naive protocol without grouped splits, many task queries reuse an alignment anchor: 60.6% of concept-identification rows, 69.6% of attribute/sense rows, and 57.9% of cross-lingual rows. We therefore treat such non-grouped task evaluation as transductive and base our primary claims on the strict grouped splits. Across seeds 42–44, validation-selected PCA-Procrustes and regularised CCA each achieve 100% mean top-1 and top-5 retrieval on untouched task-test concepts; SVCCA reaches 98.8%/100%, and full-rank Procrustes reaches 89.3%/99.0% (mean cosine 0.911, 0.951, 0.969, and 0.472 respectively; selected reduced ranks range from 8 to 64). These results confirm that the useful cross-architecture correspondence is low rank and that fitting an unconstrained 4,096-dimensional rotation from roughly 100 anchors is an unnecessarily weak baseline.

Task-level evaluation. Strict task coverage varies by grouped split: concept identification contributes 28, 34, and 30 rows across seeds, attribute/sense 16, 20, and 24. On concept identification, pooled text-passage accuracy is 64.1% (59/92) versus 58.7% (54/92) for the best latent channel, SAE-CCA (5.4-point difference, 95% CI $[-7.9, 18.9]$, unresolved). On attribute/sense, text reaches

76.7% (46/60) versus 55.0% (33/60) for SAE full-Procrustes (21.7-point difference, 95% CI [8.8, 35.0]); every latent comparison on this family excludes zero in favour of text (exact McNemar tests, Holm-adjusted over eight latent comparisons). Near-perfect concept retrieval therefore does not ensure that a mapped query preserves the task-conditioned geometry needed by the MC decoder. Cross-lingual coverage is too sparse for a stable comparison: seed 42 has no eligible test row, seeds 43/44 contribute 17 and 12; pooled text accuracy is 62.1% versus 55.2% for dense PCA-Procrustes (6.9-point difference, 95% CI [−17.2, 31.0]; unweighted seed means 59.1%/59.3%). We treat this apparent parity as inconclusive rather than evidence of a latent advantage.

4.3 Augmentation, Bandwidth, and a Continuous Baseline

On the English attribute/sense subset, all four augmentation variants, including the lost-feature variant, equal the 76.8% text-passage baseline at $\alpha = 0.01$ and degrade at larger coefficients. Strict, validation-selected dense and full-SAE controls (raw addition, norm-matched addition, spherical interpolation, no augmentation, plus shuffled/norm-matched random sender vectors) never beat native receiver text on an untouched test split; the only non-text validation selection falls from 71.4% to 64.3% on test, matching its shuffled and random controls, pointing to projection/selection instability rather than sender-specific gain.

A targeted sender-name SAE cache (384 task rows, 180 unique concept names) decomposes each concept's passage-elicited sparse code into lost, surviving, and gained features relative to the bare-name encoding, plus a matched-random control of the same count and coefficient multiset as the lost set. A lost-feature mixing rule is selected on alignment-validation tasks per seed and frozen before scoring every subset on task-test concepts. Pooled across seeds 42–44: concept identification gives text/lost/matched-random all 59/92 (64.1%), surviving 60/92 (65.2%), gained 61/92 (66.3%); attribute/sense gives every channel an identical 46/60 (76.7%); cross-lingual gives lost 19/29 (65.5%) versus text/surviving/gained/matched-random all 18/29 (62.1%). No feature subset differs from its matched-random control by more than one hit, so lost features behave like noise of matched count and magnitude under this decoder rather than carrying detectable suppressed semantics.

Bandwidth and compute. Table 6 reports bits transmitted and receiver compute alongside pooled test accuracy for every channel. The mean transmitted passage is 15.8 sender tokens (267.5 bits); the mean bare name is 4.0 tokens (67.6 bits). Dense (65,536 bits) and SAE-sparse (2,300 bits) channels need zero receiver forward passes per query because candidates are precomputed and reused, but this bandwidth is one to two orders of magnitude larger than the transmitted text and still does not exceed text-passage accuracy on any task family. The top-k lexical channel selects $k \in \{2, 4\}$ on nearly every seed and task family (66.6–124.8 bits, comparable to or smaller than the bare-name channel and 4–20× smaller than text-passage), needing only k single-token forward passes rather than one ∼16-token pass. Pooled across seeds it reaches 56.5% on concept identification (vs. 64.1% text; CI [−7.1, 22.4], unresolved), 45.0% on attribute/sense (vs. 76.7%; CI [16.7, 47.6], text stronger), and 65.5% on cross-lingual tasks (vs. 62.1%; CI [−24.1, 20.7], unresolved, nominally favouring the lexical channel). No channel dominates every axis: text-passage is strongest and most bandwidth-efficient on attribute/sense; the lexical channel matches or nominally exceeds text on cross-lingual identification at a fraction of the bandwidth and compute; dense/SAE are advantageous only where many queries reuse the same precomputed candidates.

5 Discussion

Representational convergence between Llama 3.1 8B-Instruct and Mistral 7B is substantial: low-rank maps recover held-out concept identity almost perfectly under alias-aware splits. The strict task results separate this geometric fact from communication utility. Better alignment eliminates the retrieval deficit but does not eliminate the task deficit, showing that concept identity is not sufficient for task-conditioned communication.

5.1 The Honest Negative

On the two strict task families with adequate coverage, text-passage outperforms every evaluated latent mapping. Cross-lingual results remain inconclusive because one split contributes no test rows.

The result is informative because reduced-rank retrieval is already near perfect: the remaining task gap cannot be dismissed as failed concept matching.

The correct reading is narrower than "text is sufficient" or "latent communication does not help." Under this model pair, grouped split protocol, cosine decoder, and task suite, near-perfect concept retrieval does not guarantee task-conditioned communication. Text is demonstrably stronger on attribute/sense tasks; the best concept-identification comparison and the underpowered cross-lingual comparison remain unresolved. Matched lost/surviving/gained/random feature-subset controls (§4.3) show that lost features are not distinguishable from noise of matched count and magnitude under this decoder, but this does not prove that lost features are universally irrelevant under other decoders or interventions, and the sparse prompt design does not support broad probe-generalisation claims.

This comparison could still be confounded by unequal transmission budgets and receiver compute: text-passage carries roughly 268 bits and one $\sim$16-token forward pass, while dense and SAE latent channels carry far more bits (65,536 and 2,300 respectively) yet incur no marginal receiver compute at query time (§4.3). The top-k lexical channel isolates modality from budget: at bandwidth comparable to or below the bare-name channel, it is competitive with text-passage on cross-lingual identification but clearly weaker on attribute/sense. Text's advantage on attribute/sense therefore appears to reflect something about the modality itself, not merely a larger transmission or compute budget.

5.2 Limitations

Two conditions not met by the current experiments could in principle yield a practical latent advantage: tasks requiring information that resists faithful text serialisation entirely, and cross-modal communication with a model that has no text channel to fall back on. We take up both directions in §6. A third condition, however, is already realised here rather than merely hypothetical: in bandwidth- or interpretability-constrained settings, the SAE channel's 28$\times$ compression over dense at 99.4% fidelity, with named and interpretable feature activations, has practical value independent of whether it exceeds text accuracy.

Both sender and receiver are 8B-parameter models; larger models may develop richer latent structures in which contextual representations encode a greater proportion of task-relevant information. There are only very few pretrained SAEs available for such larger models, so for this initial investigation we chose to restrict ourselves to small models.

Full-rank Orthogonal Procrustes is underdetermined at the available anchor counts. Reduced-rank methods resolve concept retrieval but not task performance, suggesting that the geometric decoder needs task-conditioned information beyond concept identity. By contrast, held-out intra-model generation injection confirms that contextual dense and SAE vectors can affect output far above matched controls, although neither channel exceeds 40% and their difference is unresolved. This does not establish a ceiling for learned nonlinear channels, richer anchor supervision, or improved generative receiver interfaces.

Lastly, most probe concepts have only two prompts, and strict grouped splits leave small, unequal task subsets; the cross-lingual estimate uses only two nonempty seeds. The inventory includes English and multilingual concepts but remains small and hand-curated. All task answers are text-expressible, and candidate scoring is geometric rather than generative. Structural injection barriers in instruction-tuned models provide an additional deployment limitation (Appendix A.4). The receiver-compute figures in §4.3 are token-count proxies, not wall-clock measurements: the top-k lexical channel's k forward passes are unbatched in our implementation and would be markedly cheaper if batched.

6 Outlook and Conclusion

The central open question this paper cannot answer is whether a language model encodes knowledge in its latent space that it cannot faithfully transmit through text. A preliminary Chain-of-Thought (CoT) experiment on BIG-Bench Hard reasoning tasks is encouraging: injecting a sender's pre-answer hidden state recovers 33.1% accuracy on tasks where CoT succeeds but direct inference fails

— far above a 0.6% no-communication baseline, though well below the 96.2% the full text CoT trace achieves on the same tasks (full setup in Appendix A.2). Latent states do encode reasoning information absent from the direct answer, but text serialisation of that reasoning conveys it far more effectively than latent injection does. Demonstrating a latent advantage on reasoning tasks would require either tasks where CoT serialisation itself loses information, or substantially improved injection methods — both beyond our current experimental setup.

A more immediate and equally rigorous test of the Platonic hypothesis is cross-modal alignment between a pure vision model and a pure language model on shared medical concepts: since neither model would ever have shared a modality or a training objective, any resulting geometric compatibility could not be attributed to joint training and would be purely emergent. We outline a candidate protocol using a text-free medical imaging foundation model in Appendix A.3.

We find strong low-rank representational convergence between architecturally distinct models, but no corresponding latent-channel advantage on the strict tasks. Text is reliably stronger on attribute/sense tasks, while uncertainty prevents ranking the best channels on concept-identification and cross-lingual tasks. The key open question is therefore not only whether spaces can be aligned, but which task-conditioned state must be transmitted and how a receiver should use it. Larger prompt families, matched continuous baselines, norm-aware interventions, and communication-specific reasoning tasks are the next tests.

REFERENCES

Y. Bansal, P. Nakkiran, and B. Barak. Revisiting model similarity. In *NeurIPS 2021*, 2021.

T. Bricken, A. Templeton, J. Marcus, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR 2023*, 2023.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. In *ICLR 2018*, 2018.

A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov. Emerging cross-lingual structure in pretrained language models. In *ACL 2020*, 2020.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.

M. Huh, B. Cheung, T. Wang, and P. Isola. The platonic representation hypothesis. In *ICML 2024*, 2024.

S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *EMNLP 2017*, 2017.

K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR 2015*, 2015.

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP 2021*, 2021.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *ACL 2019*, 2019.

R. Schäfer, T. Nicke, H. Höfener, A. Lange, D. Merhof, F. Feuerhake, V. Schulz, J. Lotz, and F. Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 2024.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR 2017*, 2017.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

A Appendix

A.1 Prerequisite Gate Details

Logistic regression probes trained on layer 15 activations achieve 100% in-sample binary classification accuracy on all 84 concept pairs. This establishes separability of the observed points but not prompt-level generalisation. Activation steering with the fitted directions produces concept shifts, but the layer-15 intervention is separate from the layer-23 channel evaluation. The leave-one-prompt-out layer-23 diagnostic is reported in §4.1.

At the SAE feature level, semantically related concept pairs activate significantly more overlapping features than unrelated pairs: $\Delta\cos = +0.319$ at layer 23, and consistently positive at all seven tested layers (3, 7, 11, 15, 19, 23, 27), excluding a random effect.

A.2 Preliminary Chain-of-Thought Latent Injection

On a 27B-parameter model (Gemma 3), CoT improves accuracy from 70.7% (direct) to 87.2% (with reasoning) on BIG-Bench Hard logical reasoning tasks, identifying 157 "reverse-gap" tasks where CoT succeeds but direct inference fails. Injecting the sender's pre-answer hidden state into the receiver recovers 33.1% accuracy on these tasks, against a 0.6% no-communication baseline and a 96.2% text-CoT ceiling on the same tasks (a 63 pp gap).

A.3 Cross-Modal Medical Alignment: Candidate Protocol

We identify MedicalMultitaskModeling (M3/UMedPT; (Schäfer et al., 2024)) as a promising candidate vision model — a foundational medical imaging framework trained via supervised multi-task learning with no text objective. The alignment protocol would follow §3.2, adapted for the dimensionality mismatch between the vision encoder's feature space and the language model's residual stream. A positive retrieval result — given a pathology image, retrieve the correct text concept — would be the strongest available evidence for cross-modal representational convergence, since the two models would never have shared a modality, let alone a training objective.

A.4 Structural Injection Barriers in Instruction-Tuned Models

Beyond the geometric alignment question addressed in §4.2, practical deployment of latent injection faces a second, largely independent obstacle in instruction-tuned models. Pilot experiments with Gemma 3 27B-IT reveal that latent injection into models with structured prompt templates faces architectural obstacles beyond geometric alignment. The BOS token position acts as an attention sink, accumulating a hidden-state norm 35$\times$ larger than regular positions by layer 20. Injecting sender vectors at this position produces $\leq$3.3% accuracy regardless of blending coefficient, because the sender signal constitutes $<$7% of the blended vector's energy even at $\alpha = 0.70$. Norm-rescaling recovers partial signal (10%) but remains far below the only viable alternative: last-token injection at layer 20, which achieves 20% on an 85-concept evaluation — above the no-communication baseline (McNemar $p < 0.001$) but not significantly above chance. More fundamentally, embedding-level prepend of source tokens before the chat template markers fails completely (3.3%), despite the same tokens achieving 96.7% when placed *within* the template. Instruction-tuning couples semantic extraction tightly to template structure; tokens outside the expected format are treated as positional noise.

This finding is orthogonal to the Procrustes tax discussed in §5: even a perfectly aligned sender vector would face this injection barrier in any model with structured prompt templates — which includes all current instruction-tuned and chat-finetuned architectures. Practical deployment of latent injection in multi-agent systems must account for these template-level constraints independently

of alignment quality, and we flag this as a prerequisite engineering problem for any future work pursuing the dense or SAE-sparse channels in instruction-tuned deployment settings.

A.5 Additional Tables and Figures

Table 1: In-sample reconstruction diagnostic. The same prompt units contribute to fitting and channel scoring, so these values must not be interpreted as held-out generalisation.

Channel	Probe Acc.	Bandwidth	Compression
Dense	$100.0 \pm 0.0\%$	65,536 b	$1\times$
SAE-sparse	$99.4 \pm 3.8\%$	$\sim$2,300 b	$28\times$
Text	$80.4 \pm 24.4\%$	$\sim$53 b	$1{,}236\times$

Table 2: Dense leave-one-prompt-out probe diagnostic over 84 concept pairs. Most concepts have only two prompts; matched prompt-level SAE and text vectors were unavailable.

Metric	Accuracy	n predictions
Dense held-out prompts	81.6%	348
Label-permutation control	49.8%	348
Mean fold training accuracy	100.0%	—

Table 3: SAE feature survival after text round-trip (165 concepts). The experiment shows that text round-tripping (§3.2) does not attenuate the original features, but replaces them.

Metric	Mean	Std
Feature survival rate	11.7%	$\pm$6.3%
Feature loss rate	88.3%	$\pm$6.3%
Jaccard overlap	5.8%	$\pm$2.6%
SAE-space cosine	0.187	$\pm$0.116
Magnitude correlation	0.326	$\pm$0.587

Table 4: Strict grouped inductive retrieval, mean over seeds 42–44. Hyperparameters for reduced-rank methods are selected on alignment-validation concepts; task-test concepts and aliases are untouched.

Mapping	Mean cosine	Top-1	Top-5
Full Procrustes	0.472	89.3%	99.0%
PCA-Procrustes	0.911	100.0%	100.0%
Regularised CCA	0.951	100.0%	100.0%
SVCCA	0.969	98.8%	100.0%

Table 5: Strict grouped task evaluation pooled across seeds 42–44. "Best latent" is selected descriptively from the evaluated mappings; method names are shown in parentheses.

Task family	Text-passage	Best latent	n
Concept identification	64.1%	58.7% (SAE-CCA)	92
Attribute/sense	76.7%	55.0% (SAE-full)	60
Cross-lingual	62.1%	55.2% (dense-PCA)	29

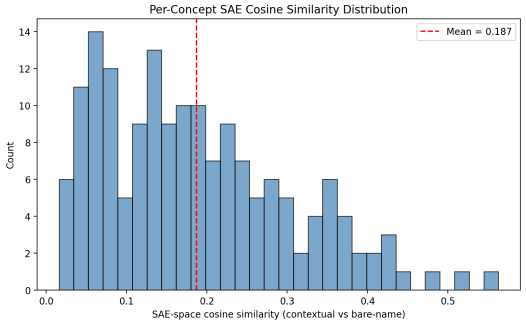


Figure 2: SAE feature cosine similarity distribution through text round-trip.

Table 6: Bandwidth, receiver compute, and pooled strict test accuracy per channel. Latent rows use PCA-Procrustes specifically (not the best method per task, cf. Table 5). Lexical bandwidth is the per-task mean of the validation-selected k.

Channel	Bits	Fwd. passes/query	Concept-ID	Attr/Sense	Cross-ling.
Dense (PCA-Procrustes)	65,536	0	54.3%	41.7%	55.2%
SAE (PCA-Procrustes)	2,300	0	54.3%	40.0%	51.7%
Text-passage	267.5	1 ($\sim$16 tok.)	64.1%	76.7%	62.1%
Text-name	67.6	1 ($\sim$4 tok.)	100.0%*	—	—
Top-k lexical	66.6–124.8	$k \in \{2, 4, 8\}$ tok.	56.5%	45.0%	65.5%

*Tautological on concept identification; omitted as a meaningful baseline (§3.3).

Table 7: Summary of experimental gates and claims.

Claim / Gate	Result	Key Number
Dense held-out probing (§4.1)	Diagnostic	81.6% vs. 49.8% permuted
SAE overlap (§4.1)	Confirmed	Δcos = +0.319
Feature loss (§4.1)	Characterised	88.3% destroyed
Sparse reconstruction	In-sample	99.4% probe diagnostic
Low-rank cross-arch (§4.2)	Confirmed	100% strict top-1
Strict concept task	Unresolved	$\Delta = +5.4$ pp, CI [-7.9, 18.9]
Strict attribute/sense	Text stronger	$\Delta = +21.7$ pp, CI [8.8, 35.0]
Strict cross-lingual	Inconclusive	29 rows, 2 nonempty seeds
Text augmentation	Null under decoder	0 pp at $\alpha = 0.01$
Lost-feature causal control	Indistinguishable from random	$\leq$1 hit vs. matched-random
Channel bandwidth/compute (§4.3)	Pareto, no dominant channel	SAE 2,300b vs. text 267.5b
Top-k lexical channel (§4.3)	Competitive on cross-lingual	65.5% vs. 62.1% text, CI [-24.1, 20.7]