The Dissociation of Geometry and Function: Why Activation-Level State Transfer Cannot Compete with Text

Jeremy McEntire[1]

April 2026

Abstract

Imagine one AI agent has just finished analyzing a complex medical case. Could it transfer its reasoning state to a second agent by copying internal neural activations—bypassing the bottleneck of explaining its findings in words? We tested this idea systematically with four approaches across 160 domain probes in four domains using Qwen 2.5-7B.

Text beats all of them.

A three-sentence summary achieves 95% domain classification accuracy and preserves 63.7% of the sender's reasoning trajectory (measured by continuation perplexity—the receiver's ability to predict the sender's specific next words). The best activation injection adds 1.9% to classification and 1.2% to trajectory alignment. Full activation injection without a text scaffold—despite preserving 4× higher representational geometry (RSA = 0.47 vs. 0.11)—carries zero reasoning trajectory information: the receiver's perplexity is indistinguishable from having no context at all. The strongest result comes from using activations to *select* input sequences rather than inject state: priming selection closes 48.9% of the expert–baseline gap, outperforming activation injection by 5.4×.

The four experiments converge on a single finding: internal representations that look geometrically similar do not produce functionally equivalent behavior. The model's internal "similarity maps" (which probes resemble which) are well-preserved by activation injection, but its actual word-by-word reasoning is not. The practical recommendation is direct: design the input sequence, do not inject activations.

1 Introduction

When two instances of the same language model process the same input under identical computational conditions (deterministic inference, same hardware precision), their internal representations are identical. When they process different inputs on the same topic, their representations diverge geometrically but may remain functionally equivalent: both classify

[1]Correspondence: jmc@cageandmirror.com

the topic correctly, both generate appropriate continuations, both maintain similar uncertainty profiles. The gap between geometric similarity and functional equivalence is the central concern of this paper.

Multi-agent AI systems face a coordination problem: when one agent has developed expertise on a complex task, how should it share that expertise with another agent? The standard approach is text—the agent summarizes what it found, and the second agent reads the summary. But text is lossy. A summary compresses a rich internal processing state into a few sentences, discarding the specific activation patterns, attention weights, and representational geometry that encoded the agent's reasoning.

The intuition behind activation-level state transfer offers an alternative. If a sender instance has processed a complex domain-specific input, perhaps its internal representation—the hidden-state vector at an intermediate layer—can be transmitted to a receiver instance more efficiently than re-processing the input through text. The sender's hidden state encodes everything the model has computed so far. Copying it should be faster and more complete than summarizing it in natural language.

We tested this intuition with four approaches, each probing a different aspect of activation-level coordination:

1. **INLP projection transfer**: Inject a 36-dimensional domain-discriminative fingerprint—extracted via Iterative Nullspace Projection (INLP; Ravfogel et al. 2020), a method that finds the directions in activation space most useful for distinguishing domains—at the sender's peak selectivity layer (Section 3).

2. **Full bandwidth transfer**: Sweep from 36 to 3,584 injected dimensions, measuring Representational Similarity Analysis (RSA; Kriegeskorte et al. 2008)—a metric that compares the full geometry of which inputs the model treats as similar—and output divergence at each bandwidth (Section 4).

3. **Continuation perplexity**: Directly measure whether the receiver can predict the sender's specific word choices, token by token—the most demanding test of reasoning trajectory alignment (Section 5).

4. **Activation-guided input selection**: Use activation-space proximity to select input sequences rather than inject activation vectors (Section 6).

Each experiment uses finer-grained metrics than the last. Domain classification (Section 3) asks "which domain?"—a coarse question that text answers at 95%. Representational Similarity Analysis (Section 4) asks "which probes are similar to which?"—capturing the

full geometry of the representation space. Continuation perplexity (Section 5) asks "can you predict the sender's specific next words?"—the operational test of reasoning alignment. Activation-guided selection (Section 6) reframes the question entirely: instead of transmitting state, use state information to program the receiver's input.

This work builds on the representation alignment literature. Centered Kernel Alignment [Kornblith et al., 2019] and Representational Similarity Analysis [Kriegeskorte et al., 2008] provide tools for comparing neural representations across conditions and models. Probing classifiers [Ravfogel et al., 2020] test whether specific information is linearly decodable from representations. Our contribution is to push beyond representational comparison to functional equivalence: not "are the representations similar?" but "does similarity in representation produce similarity in behavior?"

The parent paper in this series, *Universal Entanglement in Transformer Activation Space* [McEntire, 2026a], established that concept representations in transformers are geometrically entangled—every direction in the informative subspace carries every concept. That finding motivates the present question: if concepts cannot be isolated geometrically, can they be transferred geometrically?

All experiments use Qwen 2.5-7B (28 transformer layers, $d = 3584$) with injection at layer 10 and measurement at layer 27. This is same-model transfer: the sender and receiver are the same model processing different inputs. We are not testing cross-model or cross-architecture transfer, which faces additional alignment challenges. The limitation is acknowledged upfront; same-model transfer is the best case for activation-level coordination, and if it fails here, cross-model transfer faces steeper obstacles.

2 Methods

2.1 Model and probe design

All experiments use Qwen 2.5-7B with 28 transformer layers ($d = 3584$). The injection layer is 10—the domain selectivity peak identified in earlier work in this series. The terminal measurement layer is 27.

The probe set comprises 160 domain-specific prompts: 40 each in medical, legal, code, and science domains. Probes are designed to elicit domain-specific processing—diagnostic reasoning, statutory interpretation, debugging logic, experimental design—rather than surface-level topic identification.

2.2 INLP basis computation

Iterative Nullspace Projection (INLP; Ravfogel et al. 2020) computes 36 domain-discriminative directions (9 per domain) at layer 10 via iterative ridge regression. After each iteration, the classifier's weight vector is projected out of the representation, and a new classifier is trained on the residual. The resulting 36 directions are orthonormalized via QR decomposition to obtain basis $\mathbf{Q} \in \mathbb{R}^{3584 \times 36}$.

For each probe i, the INLP projection is:

$$\mathbf{v}_i = \mathbf{Q}\mathbf{Q}^\top(\mathbf{h}_i^{(10)} - \boldsymbol{\mu}) \tag{1}$$

where $\boldsymbol{\mu}$ is the mean layer-10 activation across all probes.

2.3 Metrics

Four metrics span from coarse to fine-grained:

- **Domain classification accuracy**: Ridge classifier trained on sender terminal activations, tested on receiver terminal activations. The coarsest measure—asks only "which domain?"
- **RSA (Representational Similarity Analysis)**: Spearman correlation between pairwise cosine distance matrices at the terminal layer [Kriegeskorte et al., 2008]. Captures the full geometry of the representational space—not just domain labels but which probes are similar to which. Decomposed into between-domain RSA_{bw} and within-domain RSA_{wi}.
- **KL divergence**: Approximate $\text{KL}(P_{\text{sender}} \| Q_{\text{receiver}})$ from top-100 token probabilities. Measures functional output alignment.
- **Continuation perplexity**: Per-token cross-entropy when the receiver predicts the sender's greedy 64-token continuation. The finest measure—asks "can you predict the sender's specific word choices?"

2.4 Injection mechanism

For all injection conditions, the sender's activation vector (INLP-projected or full) is added to the receiver's hidden state at layer 10 via a forward hook:

$$\mathbf{h}_{\text{receiver}}^{(10)} \leftarrow \mathbf{h}_{\text{receiver}}^{(10)} + \alpha \cdot \mathbf{v}_i \tag{2}$$

with $\alpha = 1.0$ (the optimal value from the alpha sweep reported in Section 3).

2.5 Resource asymmetry

A note on fairness: the text condition transmits ∼150 tokens (a three-sentence summary), while activation injection transmits 36 floats (INLP) or 3,584 floats (full). In information-theoretic terms, 150 tokens at ∼16 bits per token is ∼2,400 bits; 36 floats at 32 bits is ∼1,152 bits; 3,584 floats is ∼114,688 bits. The comparison is unfair *by design*—we are asking whether any amount of activation-level information can improve on what text already provides, not whether equal-bandwidth channels perform equally. The text channel has the advantage of being processed by the full forward pass (all 28 layers), while injection enters at a single layer (layer 10) and must survive 17 subsequent layers of text-conditioned attention.

3 INLP Projection Transfer

The simplest test of activation-level coordination: extract a 36-dimensional domain fingerprint from the sender, inject it into the receiver at the same layer, and measure whether domain alignment improves.

3.1 Text baseline: the natural language bottleneck preserves domain identity

A three-sentence summary of the sender's input, fed to the receiver, achieves 95.0% domain classification accuracy at the terminal layer (Table 1). The summary preserves enough domain-specific content—medical terminology, legal jargon, code syntax, scientific notation—that the receiver correctly identifies the domain 95% of the time with no activation-level augmentation.

Table 1: Text baseline metrics. Domain classification accuracy from text summaries alone.

Domain	Classification	INLP Cos (L10)	Terminal Cos	Entropy
Medical	0.950	0.179	0.677	1.73
Legal	0.950	0.313		
Code	0.900	0.375		
Science	1.000	0.198		
Overall	**0.950**	**0.266**	**0.677**	**1.73**

Despite 95% classification accuracy, the INLP cosine between sender and receiver at layer 10 is only 0.266—the domain information is present in both representations but encoded in *different directions*. Code has the highest INLP cosine (0.375), likely because code syntax is

preserved most faithfully in text summaries. Medical and science have the lowest (0.18–0.20), suggesting their domain-specific processing involves representations that are harder to recover from text.

3.2 INLP injection: marginal improvement over text

Table 2: INLP injection alpha sweep. Only $\alpha = 1.0$ produces measurable improvement.

α	Accuracy	Δ Acc	Terminal Cos	INLP VF	Entropy
0.01	0.950	+0.000	0.6771	0.0197	—
0.05	0.950	+0.000	0.6771	0.0198	—
0.10	0.950	+0.000	0.6772	0.0198	—
0.20	0.950	+0.000	0.6773	0.0199	—
0.50	0.950	+0.000	0.6775	0.0201	—
1.00	**0.969**	**+0.019**	**0.6781**	0.0204	—

INLP injection produces zero improvement at $\alpha \leq 0.5$ and a +1.9% improvement at $\alpha = 1.0$ (Table 2). The improvement ceiling is 5% (from 95% to 100%); INLP injection captures 38% of the available headroom. The scaling parameter α is a simple multiplicative factor rather than a principled norm-matching scheme; higher values were not tested because $\alpha > 1.0$ risks distorting the receiver's representation beyond the natural activation range. Norm-based or learned scaling might improve injection efficacy, though the continuation perplexity results in Section 5 suggest the limiting factor is not injection magnitude.

3.3 Controls: INLP outperforms full activation

Table 3: Controls at $\alpha = 1.0$. The 36-dimensional INLP projection outperforms the full 3,584-dimensional activation.

Condition	Accuracy	Δ vs Text
Text baseline (no injection)	0.950	—
INLP injection (36 dims)	**0.969**	+0.019
Full activation (3,584 dims)	0.963	+0.013
Random injection (matched norm)	0.950	+0.000

The 36-dimensional INLP projection outperforms the full 3,584-dimensional activation injection (+1.9% vs. +1.3%; Table 3). The INLP projection acts as a denoising filter: it retains domain-discriminative signal and strips domain-agnostic structure that would interfere with the receiver's representation. Random injection has zero effect, confirming the improvement is direction-specific.

3.4 Cross-domain pull matrix

Table 4: Cross-domain pull matrix at $\alpha = 1.0$. Rows: injected domain-mean INLP. Columns: true domain of target probes. Values: fraction classified as the injected domain.

	Injection				Text Baseline			
Inject ↓ / True →	Med	Leg	Code	Sci	Med	Leg	Code	Sci
Medical	0.950	0.000	0.025	0.025	0.950	0.000	0.025	0.000
Legal	0.000	0.950	0.025	0.000	0.000	0.950	0.025	0.000
Code	0.000	0.000	0.950	0.000	0.000	0.000	0.900	0.000
Science	0.075	0.050	0.125	**1.000**	0.050	0.050	0.050	1.000

Science injection is the most effective cross-domain perturbation: it raises code-probe misclassification-as-science from 5.0% to 12.5% and achieves 100% science self-recognition. Medical, legal, and code injections largely reproduce the text baseline. The asymmetry may reflect hierarchical domain overlap—science shares vocabulary with medical and code, while medical is more lexically isolated.

3.5 Rotation analysis: no shared coordinate system

Table 5: Procrustes rotation analysis of INLP encodings at layer 10.

Domain	Raw Cosine	Procrustes-Corrected
Medical	0.179	0.315
Legal	0.313	0.437
Code	0.375	0.377
Science	0.198	0.306
Overall	**0.266**	**0.359**

Procrustes alignment improves the overall INLP cosine from 0.266 to only 0.359, with a residual of 1.18 exceeding the signal norm (Table 5). The rotation matrix has determinant -1 (an improper rotation/reflection). There is no consistent rotation between sender and receiver INLP encodings. The transformation from original probe to text summary produces content-dependent rotations in the INLP subspace—different summaries rotate the domain encoding differently. A fixed 36×36 correction matrix cannot solve the alignment problem.

4 Full Activation Injection: Bandwidth vs. Fidelity

If 36 dimensions are not enough, would more help? This section sweeps injection bandwidth from 0 to 3,584 dimensions while measuring representational similarity and output divergence—finer-grained metrics than domain classification.

4.1 Eight transmission conditions

Eight conditions span the bandwidth range: text only (0 injected dimensions), text + INLP (36d), text + PCA at 50/100/200 dimensions, text + full activation (3,584d), BOS token + full activation (no text), and BOS only (control). All use $\alpha = 1.0$ at layer 10.

4.2 RSA: text determines terminal geometry

Table 6: RSA and KL divergence across transmission conditions.

Condition	Dims	RSA	RSA_{bw}	RSA_{wi}	KL	Cos
Text only	0	0.107	0.063	0.112	7.35	0.677
Text + INLP	36	0.116	0.069	0.116	7.35	0.678
Text + PCA-50	50	0.116	0.067	0.126	7.30	0.684
Text + PCA-100	100	0.115	0.067	0.125	7.30	0.685
Text + PCA-200	200	0.114	0.065	0.124	7.30	0.685
Text + Full	3,584	0.114	0.065	0.124	7.30	0.685
BOS + Full	3,584	**0.466**	**0.403**	**0.469**	**6.97**	0.398
BOS only	0	0.000	0.000	0.000	6.98	0.398

All text-based conditions cluster at RSA $\approx$ 0.11, regardless of injection bandwidth (Table 6). Injection does not improve representational fidelity when text is present. The text representation determines the terminal-layer geometry; injected signal is overwritten by 17 subsequent layers of text-conditioned attention.

BOS + Full achieves RSA = 0.47—4$\times$ higher than any text condition. Without text to overwrite the injected signal, the full activation propagates and preserves the sender's representational geometry. BOS alone produces RSA = 0: a single token without injection generates a fixed representation with no probe-specific structure.

The bandwidth ceiling is approximately 100 PCA dimensions. PCA-100, PCA-200, and Full produce identical results (RSA 0.114–0.115, KL 7.30). This is consistent with the effective dimensionality $d_{\text{eff}} \approx 20$ established earlier in this series: activation energy is concentrated in a low-dimensional subspace.

4.3 Domain reversal: text and activations preserve different structures

Table 7: Within-domain RSA by domain and condition.

Condition	Medical	Legal	Code	Science
Text only	0.121	0.234	0.108	−0.015
Text + INLP	0.129	0.238	0.107	−0.013
Text + PCA-50	0.142	0.258	0.109	−0.006
Text + Full	0.140	0.258	0.106	−0.008
BOS + Full	0.398	0.355	**0.485**	**0.638**

This is the most interesting finding in the bandwidth experiments. Text-based conditions preserve legal domain structure best ($\text{RSA}_{\text{wi}} = 0.23$) and destroy science structure entirely ($\text{RSA}_{\text{wi}} \approx 0$). BOS + Full reverses the hierarchy: science (0.64) $\gg$ code (0.48) $>$ medical (0.40) $>$ legal (0.35).

The two modalities carry complementary domain-specific information. Text carries lexical and semantic structure—strong for legal, where distinctive vocabulary ("whereas," "pursuant to," "the Court finds") survives summarization. Activations carry computational and geometric structure—strong for science, where mathematical reasoning and experimental logic patterns are encoded in activation geometry rather than surface vocabulary.

4.4 KL divergence: geometric preservation does not imply functional alignment

Despite 4× higher RSA, BOS + Full has KL divergence (6.97 nats) only marginally better than text conditions (7.30–7.35 nats). The forward pass from layer 10 to the output head is a many-to-one mapping that dilutes the geometric advantage: many different layer-10 geometries produce similar output distributions. Geometric preservation and functional alignment are dissociated.

5 Continuation Perplexity

Neither domain classification nor RSA directly measures whether a receiver can follow a sender's specific reasoning chain. This section introduces a more demanding metric: continuation perplexity. The sender generates a 64-token continuation from each probe. The receiver, equipped with various coordination signals, attempts to predict those specific tokens. Low perplexity means the receiver expects the sender's word choices—a direct probe of reasoning trajectory alignment.

5.1 The metric

The KV-cache implementation is critical for a fair test. The receiver's context (text summary, with optional injection at layer 10) is processed first, producing key-value pairs that encode the coordination signal. The sender's continuation tokens are then teacher-forced through the cached attention mechanism. This ensures the injection at the last context token propagates through attention to affect all continuation token predictions.

Per-token cross-entropy is $\text{CE}_t = -\log P(x_t \mid x_{<t}, \text{context})$, averaged across the 64 continuation tokens. Perplexity = exp(mean CE).

5.2 Primary result: text preserves 63.7% of reasoning trajectory

Table 8: Continuation perplexity across coordination conditions.

Condition	PPL	CE	Δ vs Text	Note
Original context	**1.59**	0.445	—	Upper bound
Text summary	3.05	1.030	—	63.7% of gap
Text + INLP (36d)	3.05	1.029	−0.001	+0.3% over text
Text + Full (3,584d)	3.03	1.025	−0.005	+1.2% over text
BOS + Full (3,584d)	5.60	1.482	+0.452	≈ No context
No context	5.62	1.483	+0.453	Lower bound
Scrambled	1,637	7.111	—	Noise floor

The original context achieves PPL = 1.59—nearly deterministic prediction (Table 8). This makes sense: given the same input, the same model produces the same output; the residual uncertainty comes from the summary prompt framing. Text summary achieves PPL = 3.05, preserving 63.7% of the perplexity gap between the original context and no context at all.

INLP injection adds 0.3% improvement over text (CE: 1.030 → 1.029). Full activation injection adds 1.2% (CE: 1.030 → 1.025). Both improvements are real but negligible relative to the text contribution.

5.3 Full activation without text: zero reasoning trajectory information

BOS + Full achieves PPL = 5.60, indistinguishable from no context at all (PPL = 5.62). This is the most striking result across all four experiments. The bandwidth experiments showed BOS + Full achieves 4× higher RSA than text. Yet this geometric fidelity carries zero reasoning trajectory information.

The explanation: RSA measures *relative* structure (which probes are similar to which), which is invariant to absolute position in activation space. Perplexity measures *absolute* prediction: does the model predict this specific token? Activation injection preserves the geometry of the domain space but does not place the receiver in the same processing state as the sender. Text achieves the opposite: it places the receiver in a similar functional state (low perplexity) through completely different geometric means (low RSA).

5.4 Per-domain breakdown

Table 9: Per-domain continuation perplexity.

Condition	Medical	Legal	Code	Science
Text summary	2.60	3.04	3.05	2.49
BOS + Full	3.96	4.94	4.93	3.88
No context	3.97	4.95	4.93	3.88

Medical and science achieve lower text-baseline perplexity (2.5–2.6) than legal and code (3.0–3.1), suggesting their continuations are more predictable from summaries (Table 9). This is notable given the bandwidth experiments' finding that science has the *lowest* within-domain RSA for text (near zero). Text preserves science's functional trajectory despite destroying its geometric structure. The two types of preservation are independent.

BOS + Full shows identical per-domain perplexity to the no-context condition, confirming that activation injection provides no domain-specific reasoning information when text is absent.

5.5 The 36.3% gap

The gap between text summary (PPL 3.05) and original context (PPL 1.59) represents 36.3% of the total perplexity range. Single-layer activation injection closes at most 1.2% of this gap. What accounts for the remaining gap is unknown. The original context contains the full input—every token, every syntactic structure, every implicit association—and the summary compresses this to three sentences. The lost information likely includes specific chains of reasoning that the model would follow, token-level associations that prime particular continuations, and implicit context that shapes generation strategy. Characterizing this gap more precisely is an open problem.

6 Activation-Guided Input Selection

The preceding three experiments treated coordination as a transmission problem: how much of the sender's internal state can be copied into the receiver? This section reframes coordination as a programming problem: what input sequence, when processed by the receiver's *entire* forward pass, produces the closest approximation to the sender's processing state?

6.1 Experimental design

Five agents with different contextual priming histories (∼150 tokens of multi-turn conversational context: medical, legal, code, science, and neutral) process each of the 160 domain probes. For each probe, the domain-matched primed agent serves as the "expert" whose 64-token continuation is the prediction target.

Four coordination strategies are tested:

- **Centroid injection (A)**: Inject the ensemble mean activation at layer 10. The activation-level approach.
- **Priming selection (B)**: Select the priming history whose agent's activation is closest to the ensemble centroid. No injection—the coordination signal determines input, not internal state.
- **Socratic scaffold (C)**: A model-generated analytical framework ("This text belongs to the medical domain and applies a classification framework...") as a text prefix.
- **Shared vocabulary (D)**: Model-generated structured labels (domain name, key concepts, analytical framework, phase) as a text prefix.

Priming selection uses activation centroids to select which input the receiver processes. It is activation-*guided* text, not pure text. The activations inform the selection; the input itself does the work.

6.2 Primary result: priming selection closes 48.9% of the gap

Priming selection closes 48.9% of the expert–baseline gap, outperforming centroid injection by 5.4× (48.9% vs. 9.1%; Table 10). Shared vocabulary closes 19.5%, outperforming injection by 2.1×. Socratic scaffolding *worsens* performance by 19.9%—the analytical summary displaces conversational priming that would have been more effective.

The ranking is decisive: selecting the right input sequence outperforms injecting the right activation vector at every comparison point.

Table 10: Continuation perplexity across coordination strategies. Gap measures fraction of the CE interval between no-coordination baseline and expert priming.

Condition	CE	PPL_{geo}	Gap Closed	RSA_{expert}
Expert priming	**1.197**	**3.31**	100%	1.00
Priming selection	1.572	4.82	48.9%	0.79
Shared vocabulary	1.788	5.98	19.5%	0.63
Centroid injection	1.864	6.45	9.1%	0.79
No coordination	1.931	6.90	0%	0.79
Socratic scaffold	2.077	7.98	−19.9%	0.60
No priming	3.935	51.2	—	0.75

6.3 Per-domain efficacy

Table 11: Per-domain gap closure for priming selection. Selection distribution: science 77%, medical 11%, legal 8%, code 4%.

Domain	CE_{base}	CE_{select}	CE_{expert}	Gap	Closed
Medical	1.862	1.241	1.140	0.722	86.0%
Science	1.721	1.322	1.257	0.464	86.0%
Legal	1.916	1.738	1.263	0.653	27.3%
Code	2.226	1.987	1.130	1.096	21.8%

Priming selection closes 86% of the gap for medical and science probes but only 22–27% for legal and code (Table 11). The explanation: science priming was selected for 77% of all probes (123/160). Science priming produces the most "average" processing state—closest to the ensemble centroid regardless of probe domain. This may simply mean that science is the best default prime for this model and probe set, rather than reflecting a deeper property of domain processing.

The domain match rate is only 15.6% (25/160)—the auto-selected priming rarely matches the probe's actual domain. The mechanism is processing style alignment: the right priming puts the receiver in a compatible reasoning mode, even when the domain is wrong.

6.4 Why Socratic scaffolding fails

Socratic scaffolding worsens performance because it provides propositional content (domain identity, framework name) rather than processing content (the multi-turn reasoning pattern that shapes how the model engages with the probe). Domain identification alone is near-useless—Section 3 showed text achieves 95% classification without any intervention. What

matters is the reasoning trajectory, and an analytical summary strips this down to its propositional skeleton.

6.5 Inter-agent agreement

Table 12: Domain coherence: mean PPL when non-expert agents predict the expert's continuation.

Domain	Non-expert PPL	Interpretation
Medical	8.0	Higher coherence
Science	9.0	Higher coherence
Legal	27.9	Lower coherence
Code	27.4	Lower coherence

Medical and science expert continuations are relatively predictable by non-expert agents (PPL $\approx$ 8–9), while legal and code continuations are highly distinctive (PPL $\approx$ 28; Table 12). Legal and code reasoning creates the most domain-specific processing trajectories—consistent with the lower gap closure for priming selection in those domains.

7 Discussion

7.1 The dissociation

The four experiments converge on a single finding: geometric similarity between representations does not imply functional equivalence in behavior. Table 13 summarizes the dissociation across all metrics.

Table 13: Summary of the geometry–function dissociation across all experiments.

Metric	Text	Activation Injection
Domain classification	95.0%	+1.9% (INLP)
RSA (overall)	0.107	0.466 (BOS+Full)
KL divergence	7.35 nats	6.97 nats (BOS+Full)
Continuation PPL	3.05	5.60 (BOS+Full)
Gap closure	63.7%	1.2% (injection) 48.9% (selection)

Activations carry structural information—which probes are similar to which, which domains cluster together—as measured by RSA and domain classification. Text carries functional information—what the model will say next, what reasoning trajectory it will follow—as measured by continuation perplexity and KL divergence. The two types of information are

dissociated: 4× higher geometric fidelity (RSA) translates to zero improvement in functional alignment (PPL).

7.2 Why injection fails: the 17-layer overwrite

The mechanism is straightforward. Injection at layer 10 writes a state snapshot to the residual stream. The subsequent 17 transformer layers (11–27), each performing attention over the text tokens in the context, reshape the residual stream to match the text content. By the time the activation reaches the output head at layer 27, the injected signal has been overwritten.

This is not a claim that injection has zero effect—it shifts classification accuracy by 1.9% and RSA by $< 1\%$. But the effect is small because the text-conditioned attention layers dominate the representation. The injected signal competes with 150 tokens of text processed by the full forward pass; the text wins.

Priming selection avoids this problem entirely. The priming history is processed by *all* layers, from embedding through the output head. Every attention layer attends to the priming tokens. The priming shapes the entire forward pass, not just one layer's residual stream. One approach writes to a single register; the other runs the full stack.

7.3 Connection to entanglement

The parent entanglement paper [McEntire, 2026a] established that concept representations in transformers are geometrically entangled: every direction in the informative subspace carries every concept. The present results extend this: entangled representations resist not just extraction but *transfer*. You cannot isolate the "medical reasoning" component of a hidden state and transplant it, because medical reasoning is distributed across all dimensions, interleaved with every other concept the model encodes. The INLP projection captures the maximally discriminative 36 dimensions, but these carry only 47% of the activation norm—the other 53% is entangled structure that INLP cannot isolate.

7.4 Practical recommendation

For multi-agent coordination where agents need to follow each other's reasoning—in this model, at this scale, with single-layer injection at layer 10—the recommendation is direct: design the input sequence rather than injecting activations. Whether this finding generalizes to multi-layer injection, terminal-layer injection, or cross-architecture transfer remains open (Section 8).

Specifically:

1. **Text is the primary channel.** A three-sentence summary preserves 63.7% of reasoning trajectory alignment.

2. **Structured priming outperforms injection by 5.4×.** Selecting the right conversational context closes 48.9% of the expert gap; centroid injection closes 9.1%.

3. **Activations are useful for selection, not injection.** The best approach uses activation centroids to *choose* which input to provide, not to inject state.

4. **Propositional framing is counterproductive.** Analytical scaffolds ("the domain is X") are worse than no coordination. Process-level priming (multi-turn reasoning examples) is 2.5× more effective than propositional framing.

5. **The effective bandwidth ceiling is ∼100 dimensions.** Beyond 100 PCA components, additional injection bandwidth is wasted.

8 Limitations

Same-model transfer. All experiments use the same model (Qwen 2.5-7B) as both sender and receiver, processing different inputs. This is the best case for activation-level transfer—the sender and receiver share identical weights, so the only alignment challenge is content-dependent rotation, not architectural or training-dependent misalignment. Cross-model transfer, which faces additional representational alignment challenges, was not tested. The negative results reported here establish an upper bound: if activation injection fails in same-model transfer, it faces steeper obstacles in cross-model scenarios.

Single injection layer. All injection occurs at layer 10, chosen because earlier work in this series identified it as the domain selectivity peak—the layer where domain-discriminative signal is strongest. This choice is empirically motivated but not exhaustive. Multi-layer injection, injection at later layers (closer to the output head), or injection via architectural mechanisms (prefix tuning, adapter layers) might maintain the injected signal more effectively through the forward pass. The 17-layer overwrite mechanism is specific to single-layer injection at an intermediate layer; the claim is that this specific approach is insufficient, not that all forms of activation-level communication are impossible.

Continuation perplexity as proxy. Perplexity measures word-level prediction, not task completion. A receiver with PPL 3.05 might still complete the sender's task correctly despite predicting different specific tokens. Perplexity is a proxy for reasoning trajectory alignment, not a direct measure of coordination success on downstream tasks.

Resource asymmetry. The text channel transmits ∼150 tokens processed by all 28 layers; activation injection transmits 36–3,584 floats at one layer. This comparison is deliberately asymmetric—we asked whether activation injection can improve on text, not whether equal-bandwidth channels perform equally. A fairer comparison might constrain text to the same information content as the injection, but this is not the relevant engineering question. In practice, text is available and cheap; the question is whether activation injection adds value on top of it.

Four domains, one model. Results are established across four domains (medical, legal, code, science) on one model architecture at one scale. The domain reversal finding (Section 4) and the processing style alignment mechanism (Section 6) may be specific to this model and probe set.

9 Conclusion

Four experiments on activation-level state transfer converge on a single finding: geometric similarity does not imply functional equivalence. Text summaries preserve functional information (63.7% of reasoning trajectory) through completely different geometric means than the sender's representation. Activation injection preserves geometric structure (4× higher RSA) but carries zero reasoning trajectory information when text is absent.

The strongest coordination strategy is not transmission but selection: using activation-space proximity to choose the right input sequence closes 48.9% of the expert–baseline gap, outperforming activation injection by 5.4×. The forward pass is the coordination mechanism. The best way to produce a target processing state in a receiver is to run the right input through all 28 layers, not to inject a state snapshot at one.

The practical implication for multi-agent AI systems is direct: invest in input design, not activation-level communication protocols. The 36.3% perplexity gap between text summary and original context remains the binding constraint on coordination fidelity. Closing that gap requires richer input sequences—not richer activation channels.

Data Availability

All results, including per-probe metrics, alpha sweeps, RSA matrices, perplexity distributions, and priming selection logs, are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper16/`, `paper17/`, `paper18/`, and `paper19/`.

This paper consolidates four previously published papers: INLP Projection Transmission (DOI: 10.5281/zenodo.18843836), Bandwidth vs Fidelity in Activation-Level Coordination (DOI:

10.5281/zenodo.18843840), Sender Continuation Perplexity (DOI: 10.5281/zenodo.18843842), and Ensemble Gravity (DOI: 10.5281/zenodo.18843844).

References

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3519–3529, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

McEntire, J. Universal entanglement in transformer activation space. Zenodo, DOI: 10.5281/zenodo.19409951, 2026.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020.

Schönemann, P. H. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carber, Y., Kornblith, S., and Schmidt, L. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 23965–23998, 2022.

Frankle, J. and Carlin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Bansal, Y., Nakkiran, P., and Barak, B. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. E. Convergent learning: Do different neural networks learn the same representations? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.