INLP Projection Transmission: Domain-Specific Activation Transfer at the Natural Language Boundary

Jeremy McEntire[1]

March 2026

Abstract

Paper XV measured domain-specific compression loss at 1.4% of activation variance. Paper XV-c revealed this as a measurement artifact: domain-specific information is ∼22% of variance at every layer, encoded in layer-specific directions that rotate through the forward pass. This paper tests whether domain-specific information can be transmitted between model instances by injecting a 36-dimensional INLP projection at layer 10.

The critical finding is that **the text baseline already achieves 95% domain classification accuracy**. The natural language bottleneck preserves domain identity far better than the geometric measurements suggested. INLP injection provides a marginal improvement (+1.9%), outperforming both random injection (0%) and full activation injection (+1.3%). However, the sender and receiver encode domain information in nearly orthogonal directions (INLP cosine = 0.27), and Procrustes alignment fails to recover a consistent rotation (residual > 1).

These results reframe the coordination problem: domain identity is not lost at the NL boundary. What may be lost is finer-grained domain expertise that classification accuracy cannot detect — the geometric encoding of *how* the model processes a domain, not *which* domain it is processing.

1 Introduction

Paper XV established that the natural language interface is a uniform lossy channel, with domain-specific information comprising 1.4% of the compression loss at layer 10. Paper XV-c overturned this measurement: when INLP directions are computed at the measurement layer rather than transferred from the terminal layer, domain-discriminative variance is 22.6% at layer 10 — a 16× correction. The domain encoding *rotates* through the forward pass; the terminal-layer INLP basis is the wrong coordinate system for intermediate layers.

This rotation has a direct implication for multi-agent coordination. If two model instances both encode domain information at 22% variance fraction but in different directions, they

[1]Correspondence: jmc@cageandmirror.com

might be unable to share domain-specific representations even though the information is abundant in both. The question is whether the INLP basis provides a shared coordinate system for domain-specific communication.

This paper tests the simplest possible activation-level coordination protocol: extract a 36-dimensional domain fingerprint (INLP projection at layer 10), transmit it, and inject it into the receiver's hidden state at the same layer. If this 36-float message improves domain alignment beyond what text alone achieves, it validates activation-level coordination as a supplement to natural language.

2 Methods

2.1 Experimental design

All experiments use Qwen 2.5-7B with 28 transformer layers ($d = 3584$). The injection layer is 10 (the selectivity peak from Paper IX) and the terminal layer is 27.

Phase 1: Sender capture. Process each of 160 domain probes (40 per domain $\times$ 4 domains) through the model. Capture last-token activations at layers 10 and 27.

Phase 2: INLP computation. Compute 36 INLP directions (9 per domain) at layer 10 via iterative ridge regression on the sender activations. Orthonormalize via QR decomposition to obtain basis $\mathbf{Q} \in \mathbb{R}^{3584\times 36}$. For each probe i, compute the INLP projection:

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

where $\boldsymbol{\mu}$ is the mean layer-10 activation across all probes. This $\mathbf{v}_i$ is the 36-dimensional domain fingerprint reconstructed in the full 3584-dimensional space.

Phase 3: Summary generation. For each probe, generate a 3-sentence summary using the prompt: "Summarize the following in exactly three sentences, preserving key technical findings."

Phase 4: Receiver text baseline. Feed each summary to the same model with prefix "Based on the following information, continue the work:" and capture layer-10 and layer-27 activations. No injection. This establishes the text-only baseline.

Phase 5: INLP injection. Repeat the receiver forward pass, but at layer 10, modify the hidden state via a forward hook:

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

where $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$. Capture terminal-layer activations and output logits.

Phase 6: Cross-domain matrix. At the optimal α, inject domain-mean INLP projections across all source–target domain pairs. Measure the 4×4 pull matrix: the fraction of target-domain probes classified as the source domain after injection.

Phase 7: Controls.

- **Random injection**: Random direction in $\mathbb{R}^{3584}$, norm-matched to the INLP projection.
- **Full activation injection**: Full centered activation ($\mathbf{h}^{(10)}_i - \boldsymbol{\mu}$) in all 3584 dimensions (upper bound on injectable information).

2.2 Metrics

- **Domain classification**: Ridge classifier trained on sender terminal activations, tested on receiver terminal activations. Measures whether injection shifts the receiver toward the sender's representation.
- **INLP cosine (L10)**: Cosine similarity between sender and receiver INLP encodings ($\mathbf{Q}^\top \mathbf{h}$) at layer 10. Measures geometric alignment of domain encoding.
- **Terminal cosine**: Cosine similarity between sender and receiver full activations at layer 27.
- **Procrustes alignment**: Estimate the optimal orthogonal rotation $\mathbf{R}$ between sender and receiver INLP encodings. The residual measures whether a consistent rotation exists.

3 Results

3.1 Text baseline: the NL bottleneck preserves domain identity

The text-only baseline achieves 95.0% domain classification accuracy at the terminal layer (Table 1). The natural language 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.

Table 1: Text baseline metrics. The NL summary preserves domain identity with high fidelity.

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 — close to orthogonal. 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 $\alpha = 1.0$. The signal is real but small: the 36-float domain fingerprint shifts the receiver's terminal representation marginally closer to the sender's.

The improvement ceiling is 5% (from 95% to 100%). INLP injection at $\alpha = 1.0$ captures 38% of the available improvement (+1.9/5.0).

Table 3: Controls at $\alpha = 1.0$. INLP projection (36 dims) outperforms full activation (3584 dims).

Condition	Accuracy	Δ vs Text
Text baseline (no injection)	0.950	—
INLP injection (36 dims)	**0.969**	+0.019
Full activation (3584 dims)	0.963	+0.013
Random injection (matched norm)	0.950	+0.000

3.3 Controls: INLP outperforms full activation

The 36-dimensional INLP projection outperforms the full 3584-dimensional activation injection (+1.9% vs +1.3%). This demonstrates that the INLP projection acts as a *denoising filter*: it strips the non-domain dimensions that introduce interference when injected. The full activation carries domain-agnostic structure (positional, syntactic, token-level) that conflicts with the receiver's existing representation at those dimensions.

Random injection has exactly zero effect, confirming that the improvement is direction-specific, not a norm artifact.

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. Text baseline confusion shown for comparison.

	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

The pull matrix has diagonal dominance of 35.5. Science injection is the most effective: it raises code-probe misclassification-as-science from 5.0% to 12.5% and pushes science self-recognition to 100%. Medical, legal, and code injections largely reproduce the text baseline.

The asymmetry is notable: science injection pulls probes from other domains, but medical injection does not. This may reflect the hierarchical structure of domain overlap — science shares vocabulary with medical and code, while medical is more lexically isolated.

3.5 Rotation analysis: the encoding is content-dependent

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**

The raw INLP cosine between sender and receiver is 0.266 — the domain encodings are nearly orthogonal. Procrustes alignment improves this to only 0.359, with a residual of 1.18 (exceeding the signal norm). 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.

Code shows the smallest Procrustes improvement ($0.375 \rightarrow 0.377$), indicating that code's INLP encoding is already as aligned as it can get — the remaining misalignment is intrinsic, not rotational. Legal shows the largest improvement ($0.313 \rightarrow 0.437$), suggesting legal domain encoding has a more consistent rotational component.

4 Discussion

4.1 Why text works and geometry doesn't

The central surprise of Paper XVI is that the 1.4% DS_{frac} from Paper XV, and even the corrected 22.6% layer-specific VF from Paper XV-c, are both misleading about functional domain preservation. The text summary preserves domain identity with 95% accuracy not because it preserves the geometric encoding (it doesn't — cosine 0.27) but because it preserves the *semantic content*: the words, the terminology, the structure. The receiver re-encodes domain identity from semantic features, arriving at the same classification through a completely different geometric path.

This dissociation between geometric similarity and functional equivalence is a central finding. Two representations can be nearly orthogonal in INLP space ($\cos = 0.27$) and yet functionally equivalent for domain classification (95% accuracy). The INLP subspace is one

of many coordinate systems the model can use to encode domain identity; the receiver simply uses a different one.

4.2 The 36-float protocol as denoising filter

The INLP injection's advantage over full activation injection (36 dims $>$ 3584 dims) demonstrates a principle for activation-level coordination: transmit the projection, not the activation. The INLP projection acts as a band-pass filter that retains domain-discriminative signal and strips domain-agnostic structure that would interfere with the receiver's internal representation.

This result has implications for activation-sharing protocols. Q-KVComm and similar approaches that transmit raw KV-cache entries may benefit from projecting onto task-relevant subspaces before transmission. The optimal transmission is not the most information — it is the most *relevant* information.

4.3 What classification accuracy cannot measure

The 95% text baseline sets a high bar, but classification is a coarse measure. It asks "which domain?" but not "what specific expertise?" A summary that says "this was a medical analysis" preserves domain identity (classification) without preserving the specific diagnostic reasoning, the differential diagnosis structure, or the treatment protocol logic that was active in the sender's representation.

The remaining 5% error and the 0.27 INLP cosine may reflect exactly this: the loss of fine-grained domain-specific processing that classification accuracy cannot detect. Papers XVII–XIX will require more sensitive metrics — representational similarity analysis, task-specific performance, and information-theoretic measures — to probe what is lost beyond domain identity.

4.4 Implications for multi-agent coordination

1. **Domain identity survives the NL bottleneck.** Text summaries preserve which domain the sender was operating in. The coordination failure reported in multi-agent systems is not about domain confusion.

2. **Domain expertise may not survive.** The geometric encoding of domain-specific processing is nearly orthogonal between sender and receiver (cosine 0.27). The fine-grained activation structure that encodes *how* the model processes within a domain is lost.

3. **INLP projection as sidecar protocol.** A 36-float sidecar message alongside the text summary captures 38% of the remaining classification improvement. At 144 bytes, this is negligible bandwidth.

4. **Consistent rotation does not exist.** A fixed correction matrix cannot align sender and receiver encodings. Any alignment protocol must be adaptive — conditioned on the specific content being communicated.

5 Conclusion

The INLP projection transmission experiment reveals a dissociation between geometric and functional domain preservation. The natural language bottleneck destroys geometric alignment (INLP cosine = 0.27) while preserving functional domain identity (95% classification accuracy). INLP injection provides marginal improvement (+1.9%) by adding geometric alignment atop the already-strong text signal, outperforming both random injection (0%) and full activation injection (+1.3%) through its denoising effect.

The absence of a consistent rotation between sender and receiver encodings (Procrustes residual > 1) means that domain-specific representations are not merely rotated by the NL bottleneck — they are *re-encoded* from semantic content. The coordination problem for multi-agent LLM systems is not domain identity (which text preserves) but domain expertise: the fine-grained processing structure that classification accuracy is too coarse to detect.

Data Availability

All results, including per-probe metrics, alpha sweep data, cross-domain matrices, and rotation analysis, are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper16/`.

Series: Activation Geometry of Domain-Selective Noise Injection, Paper XVI.