The Inter-Instance Compression Barrier: Domain-Specific Information Loss at the Natural Language Interface

Jeremy McEntire[1]

March 2026

Abstract

When LLM instances coordinate through natural language summaries, how much domain-specific information survives the compression? Papers IV and IX–XIV established that domain-specific information occupies a geometrically small subspace of the activation space (36 INLP directions in $\mathbb{R}^{3584}$). This paper measures whether the natural language interface disproportionately drops domain-specific information relative to domain-agnostic content.

The primary hypothesis (H1) predicts that the domain-specific fraction of compression loss should decrease with increasing compression — that the natural language bottleneck selectively drops domain-discriminative information. **H1 is not supported.** The domain-specific fraction of activation distance is stable at ∼1.4% (layer 10) and ∼4.4% (layer 27) across all three compression levels (full reproduction, 3-sentence summary, 1-sentence abstract). The domain-specific to domain-agnostic ratio is 0.07 at layer 10 and 0.13 at layer 27, invariant to compression level.

This null result is the stronger finding. The NL interface is a *uniform* lossy channel that does not discriminate between domain-specific and domain-agnostic information. But because domain-specific information is already only 1.4% of the activation variance, any lossy channel destroys it — even a fair one. The concentration barrier (Paper XI) is the sole constraint; the communication interface is innocent.

1 Introduction

Multi-agent LLM systems coordinate through natural language: one instance processes a task, summarizes its work, and passes the summary to the next instance. This natural language interface is a compression bottleneck. The question for multi-agent coordination is whether this bottleneck selectively drops domain-specific information — the portion of the internal representation that encodes what domain the model is operating in.

If the bottleneck is selective, then coordination failures in multi-agent systems would partly originate at the communication interface: domain expertise gets lost in translation. If

[1]Correspondence: jmc@cageandmirror.com

the bottleneck is uniform, then coordination failures originate upstream, in the geometric structure of the representation itself.

Papers IV and IX–XIV established the relevant geometry:

- 36 INLP directions span the domain-discriminative subspace in $\mathbb{R}^{3584}$ (Paper IV)
- Domain selectivity peaks at layer 10 (Paper IX)
- The forward-pass Jacobian is isotropic, treating domain and non-domain directions equally (Paper X)
- Effective dimensionality $d_{\text{eff}} \approx 20$ bounds the achievable domain fraction at $k/d_{\text{eff}} \leq 1.8$ (Paper XI)
- Classification-optimized (INLP) directions outperform causally-derived alternatives (Paper XIV)

This paper directly measures what happens to domain-specific information when it passes through the NL compression interface.

2 Methods

2.1 Experimental design

The experiment captures a single model's internal representations before and after NL compression, using the same model as both sender and receiver to isolate the compression effect from model heterogeneity.

Step 1: Original capture. Present each of 160 domain probes (40 per domain) to Qwen 2.5-7B. Capture last-token activations at layers 10 and 27. Generate a 64-token response for use in compression.

Step 2: Compress. For each probe, generate NL summaries at three levels:

- **Full**: "Repeat the following text exactly, preserving all technical details."
- **Summary**: "Summarize in exactly three sentences. Preserve key technical findings."
- **Abstract**: "Describe what the following accomplishes in a single sentence."

Step 3: Re-capture. Feed each summary to the same model with a receiving prompt ("Based on the following information, continue the work:") and capture activations at layers 10 and 27.

Step 4: Decompose. For each probe, decompose the L2 activation distance between original and compressed representations into three components:

$$\|\boldsymbol{\delta}\|^2 = \|\Pi_{\text{DS}}\boldsymbol{\delta}\|^2 + \|\Pi_{\text{DA}}\boldsymbol{\delta}\|^2 + \|\boldsymbol{\delta}_{\text{res}}\|^2 \tag{1}$$

where $\boldsymbol{\delta} = \mathbf{h}_{\text{orig}} - \mathbf{h}_{\text{comp}}$, Π_{DS} projects onto the 36-dimensional INLP subspace (domain-specific), and Π_{DA} projects onto the top-20 PCA directions (domain-agnostic). The residual is computed via QR orthogonalization of the combined basis.

2.2 Controls

No-context upper bound. Capture activations for a generic prompt with no task content. The distance between original domain probes and this baseline represents the maximum possible divergence.

Metrics.

- $\text{DS}_{\text{frac}} = \|\Pi_{\text{DS}}\boldsymbol{\delta}\|^2/\|\boldsymbol{\delta}\|^2$: fraction of compression loss in domain-specific directions
- $\text{DA}_{\text{frac}} = \|\Pi_{\text{DA}}\boldsymbol{\delta}\|^2/\|\boldsymbol{\delta}\|^2$: fraction in domain-agnostic directions
- $\text{DS/DA} = \text{DS}_{\text{frac}}/\text{DA}_{\text{frac}}$: the primary test statistic

H1 prediction. If the NL interface selectively drops domain-specific information, DS/DA should *increase* with compression level (from full to abstract). Specifically, DS_{frac} should grow as a proportion of total loss as more information is lost.

3 Results

3.1 Primary result: DS/DA ratio is stable

The DS/DA ratio shows no systematic increase with compression. At layer 10, the ratio fluctuates narrowly between 0.068 and 0.085. At layer 27, between 0.128 and 0.136. **H1 is not supported**: the NL interface does not selectively drop domain-specific information.

Table 1: Compression decomposition at two layers across three compression levels.

Level	Layer	DS_{frac}	DA_{frac}	DS/DA	Total L2
Full	10	0.0137	0.195	0.070	58.3
Summary	10	0.0144	0.171	0.085	62.2
Abstract	10	0.0141	0.206	0.068	56.9
Full	27	0.0434	0.340	0.128	542.9
Summary	27	0.0435	0.322	0.135	581.2
Abstract	27	0.0443	0.325	0.136	523.9

3.2 Absolute magnitudes

DS_{frac} is $\sim$1.4% at layer 10 across all compression levels. This means that of the total activation distance introduced by NL compression, only 1.4% lies in the domain-discriminative subspace. The remaining 98.6% is in domain-agnostic directions (19.5% in top-20 PCA) or in the residual subspace ($\sim$79%).

At layer 27, DS_{frac} rises to $\sim$4.4%. This is consistent with Papers IX–X: the terminal layer has higher INLP alignment but lower selectivity, because more of the activation energy is domain-correlated at the output stage.

3.3 Per-domain breakdown

Table 2: Per-domain DS_{frac} and DA_{frac} at layer 10, "full" compression.

Domain	DS_{frac}	DA_{frac}	Total L2
Medical	0.0143	0.230	59.4
Legal	0.0133	0.189	56.1
Code	0.0134	0.179	58.1
Science	0.0138	0.182	59.7

DS_{frac} is remarkably uniform across domains (0.0133–0.0143). Domain-agnostic fraction varies more (0.179–0.230), with medical showing the highest DA fraction. The uniformity of DS_{frac} across domains reinforces the conclusion that the compression loss is not domain-sensitive.

3.4 No-context upper bound

The no-context upper bound has *lower* total L2 than the compression conditions at layer 10 (42.4 vs. 56–62). This is unexpected: a completely uninformed prompt is "closer" to the original domain probes than the NL summaries. The explanation is that domain probes

Table 3: No-context baseline comparison. Maximum possible divergence.

Layer	DS_{frac}	DA_{frac}	DS/DA	Total L2
10	0.0174	0.418	0.042	42.4
27	0.0953	0.445	0.214	432.6

have high token-level variance (long, detailed prompts), while both generic prompts and NL summaries are short, so the activation distance is dominated by length/style differences rather than domain content.

At layer 27, the no-context DS_{frac} is 9.5% — substantially higher than the compression conditions (4.3–4.4%). This is because the terminal layer's representation more strongly reflects domain content, and removing all domain context creates a larger domain-specific displacement.

3.5 PCA structure

Top-20 PCA directions explain 68.9% of activation variance at layer 10 and 66.8% at layer 27. The PCA spectrum at both layers shows a gradual decay (no sharp elbow), consistent with the high effective dimensionality ($d_{\text{eff}} \approx 20$) reported in Paper XI.

4 Discussion

4.1 The stronger null result

H1 predicted selective compression loss. The reality is simpler and more constraining: the NL interface is a uniform lossy channel. It drops approximately 80% of activation information at layer 10, and this loss is distributed proportionally across domain-specific and domain-agnostic subspaces.

This is the stronger result because it eliminates a degree of freedom in the coordination failure explanation. If H1 had been supported, one could improve coordination by designing a better communication protocol — one that preserves domain-specific information. Since H1 fails, the communication protocol is not the bottleneck. The bottleneck is that domain-specific information is only 1.4% of the activation variance to begin with.

4.2 The concentration barrier as the sole constraint

Paper XI established that INLP variance fraction is bounded by k/d_{eff}, where k is the number of domain-specific directions and d_{eff} is the effective dimensionality. With $k = 9$ per domain

and $d_{\text{eff}} \approx 20$, the bound is $\sim$45% per domain or $\sim$1.8*overall*.

The measured 1.4% DS_{frac} at layer 10 is well below this bound, which reflects the fact that DS_{frac} measures a different quantity than the INLP variance fraction from Paper XI: it measures the *compression loss* in domain directions, not the *total variance* in those directions. The compression loss depends on what the NL summary preserves, and the summary preserves the high-variance (domain-agnostic) directions better than the low-variance (domain-specific) ones — not because it selects against them, but because high-variance directions carry more signal by construction.

4.3 Implications for multi-agent coordination

The 1.4% figure has direct implications for multi-agent LLM systems that coordinate through NL:

1. **Domain expertise is not communicated.** When an agent summarizes its domain-specific work, the receiving agent recovers <2% of the domain-specific activation structure. The summary preserves task-level content (what was done) but not the computational substrate (which internal representations were active).

2. **Improving the summary protocol cannot fix this.** Since the loss is proportional (not selective), a "better" summary format would reduce total loss but not change the domain-specific fraction. The only way to increase the domain-specific fraction is to change the underlying representation — which requires changing the model, not the communication protocol.

3. **Activation-level communication bypasses the barrier.** Passing internal activations directly (e.g., via KV-cache transfer) would preserve the domain-specific subspace. This motivates approaches like Q-KVComm and Communicating Activations, which share activation-level representations instead of NL summaries.

4.4 Layer comparison

The DS_{frac} increase from layer 10 (1.4%) to layer 27 (4.4%) is consistent with the terminal layer's higher INLP alignment (Paper X). The terminal layer encodes more domain-discriminative information because it is closer to the output head, where domain identity matters for next-token prediction. However, this higher DS fraction comes with lower selectivity (Paper IX): the domain information is more present but less isolable.

The DS/DA ratio at layer 27 (0.13) is nearly double that at layer 10 (0.07), suggesting that the domain-specific subspace at the terminal layer is relatively better represented in the

compression loss. This is because the terminal layer's INLP directions are more aligned with high-variance PCA directions — domain classification at the output stage shares variance with general output preparation.

5 Conclusion

The natural language compression interface is a uniform lossy channel. It does not selectively drop domain-specific information — it drops everything proportionally. But because domain-specific information occupies only 1.4% of the activation variance at the selectivity-peak layer, uniform compression is sufficient to destroy the domain signal. The concentration barrier from Paper XI, not the communication protocol, is the binding constraint on domain-specific information transfer in multi-agent LLM systems.

Data Availability

All results, including per-probe decompositions, PCA spectra, and summary texts, are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper15/`.

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