Ensemble Gravity:
Coordination Through Contextual Priming vs. Activation-Level Transmission

Jeremy McEntire[1]

March 2026

Abstract

Papers XVI–XVIII established that activation injection at a single layer provides at most 1.2% improvement over text coordination, and that injected geometric structure carries zero reasoning trajectory information when text is absent. This paper tests four coordination *strategies* rather than four transmission *bandwidths*: centroid injection (A), priming selection (B), Socratic scaffold (C), and shared vocabulary induction (D).

Five agents with different contextual priming histories process each probe. The domain-matched primed agent's continuation serves as the prediction target. The central result: **priming selection — choosing the right structured input sequence — closes 48.9% of the gap** between neutral coordination (CE = 1.93) and expert priming (CE = 1.20). Shared vocabulary induction closes 19.5%. Centroid activation injection closes 9.1%. Socratic scaffolding *worsens* performance by 19.9%. Priming selection achieves 86% gap closure for medical and science probes, confirming that the right input sequence is radically more effective than the right activation vector.

Coordination protocols are programs, not channels. The quality of multi-agent coordination depends on the structure of the shared input sequence, not the bandwidth of the activation transmission.

1 Introduction

Papers XVI–XVIII converged on a single finding: text carries function, activations carry geometry, and the two are dissociated. Text summaries preserve 63.7% of sender continuation perplexity (Paper XVIII). Activation injection at layer 10 adds at most 1.2%. Full activation injection without text scaffold achieves perplexity indistinguishable from no context at all, despite preserving 4× higher representational geometry (Paper XVII).

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

[1] Correspondence: jmc@cageandmirror.com

The distinction is fundamental. Transmission injects a state snapshot at one layer and hopes the remaining 17 layers preserve it. Programming runs the receiver's *entire* forward pass through a structured sequence designed to produce convergent processing. One writes to a single register; the other runs the full stack.

Five agents with different contextual priming histories (medical, legal, code, science, neutral) process each domain probe. For each probe, the domain-matched primed agent serves as the "expert" whose 64-token continuation is the prediction target. Four coordination conditions test whether the receiver can predict the expert's specific word choices:

- **Condition A**: Centroid injection — inject the ensemble mean activation at layer 10.
- **Condition B**: Priming selection — select the priming history whose agent is closest to the ensemble centroid. No injection.
- **Condition C**: Socratic scaffold — model-generated analytical framework as text prefix.
- **Condition D**: Shared vocabulary — model-generated structured labels (domain, key concepts, framework, phase) as text prefix.

2 Methods

2.1 Primed agents

Each of five agents processes a ∼150-token multi-turn conversational history before receiving the domain probe. The medical agent's history discusses cardiac enzyme interpretation; the legal agent's discusses unconscionability under UCC § 2-302; the code agent's discusses memory allocator fragmentation; the science agent's discusses spectroscopic anomalies; the neutral agent's discusses meeting note organization. Each priming history is designed to bias the agent's processing context toward a specific domain without overlapping with the probes themselves.

All agents use Qwen 2.5-7B. For each of the 160 domain probes (40 per domain), all five agents generate a greedy 64-token continuation from [priming + probe]. Layer-10 activations are captured during a separate forward pass.

2.2 Coordination conditions

The expert continuation (from the domain-matched primed agent) is the prediction target. The receiver is evaluated under seven conditions:

1. **No coordination**: Neutral priming + probe, no injection. Baseline.

2. **Centroid injection (A)**: Neutral priming + probe, with centroid activation injected at layer 10 ($\alpha = 1.0$).

3. **Expert priming**: Domain-matched priming + probe. Upper bound.

4. **Priming selection (B)**: Auto-selected priming (agent closest to centroid) + probe. No injection.

5. **Socratic scaffold (C)**: Model-generated analytical summary + probe.

6. **Shared vocabulary (D)**: Model-generated structured labels + probe.

7. **No priming**: Probe only, no priming or injection.

2.3 Perplexity computation

Continuation perplexity uses the KV-cache approach from Paper XVIII: process the receiver's context (with optional injection) to build a key-value cache, then teacher-force the expert's continuation tokens through the cached attention mechanism. Per-token cross-entropy is computed and averaged.

The centroid for each probe is the arithmetic mean of the five agents' layer-10 activations (3584 dimensions). Centered centroids (mean-subtracted across probes) are used for injection. Priming selection identifies the agent whose layer-10 activation has minimum ℓ_2 distance to the centroid.

3 Results

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

The gap between no coordination (CE = 1.93) and expert priming (CE = 1.20) is 0.734 nats. Priming selection closes 48.9% of this gap (CE = 1.57), shared vocabulary closes 19.5% (CE = 1.79), and centroid injection closes 9.1% (CE = 1.86). Socratic scaffolding *increases* cross-entropy by 0.15 nats, performing worse than the neutral baseline.

The ranking is decisive: selecting the right input sequence outperforms injecting the right activation vector by 5.4× (48.9% vs. 9.1%). Structured labeling outperforms injection by 2.1× (19.5% vs. 9.1%). Both text-level strategies beat activation-level transmission.

3.2 RSA: coordination through text reduces geometric alignment

RSA vs. expert is high for all priming conditions (≈ 0.79) and lower for text-prefix conditions (0.60–0.63). Adding analytical text (scaffolds, annotations) before the probe pushes the

Table 1: Continuation perplexity across coordination conditions. CE is mean cross-entropy (nats). $\mathrm{PPL}_{\mathrm{geo}}$ is geometric mean perplexity = exp(CE). Gap measures fraction of the CE interval between baseline and expert.

Condition	CE	$\mathrm{PPL}_{\mathrm{geo}}$	Gap Closed	$\mathrm{RSA}_{\mathrm{expert}}$
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

terminal-layer geometry *away* from the expert's geometry. Yet priming selection achieves better perplexity (lower CE) despite identical RSA to the no-coordination baseline. This extends the Paper XVII finding: functional alignment and geometric alignment remain dissociated even within text-based coordination strategies.

3.3 Per-domain breakdown: priming selection shows domain-specific efficacy

Table 2: Per-domain CE and gap closure for priming selection. Priming selection distribution: science 77%, medical 11%, legal 8%, code 4%, neutral 1%.

Domain	$\mathrm{CE}_{\mathrm{base}}$	$\mathrm{CE}_{\mathrm{select}}$	$\mathrm{CE}_{\mathrm{expert}}$	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. The explanation lies in the selection distribution: 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 centroid-proximity makes science priming an effective substitute for medical and science probes (where the analytical reasoning style is similar) but a poor substitute for legal and code probes (where domain-specific vocabulary and reasoning patterns diverge sharply).

The domain match rate is 15.6% (25/160) — the auto-selected priming rarely matches the probe's domain. Yet it closes 48.9% of the gap overall. The mechanism is not domain matching but *processing style alignment*: the right priming puts the receiver in a compatible reasoning mode, even when the domain is wrong.

3.4 No-priming condition: expert continuations are radically probe-specific

The no-priming condition (probe text only, no conversational history) achieves CE = 3.94, corresponding to geometric mean PPL = 51.2. This is far worse than Paper XVIII's text-summary baseline (PPL = 3.05) because the prediction target is different: Paper XVIII predicted the *unprimed* model's continuation; this paper predicts the *domain-primed* model's continuation. The 150-token priming history fundamentally changes what the model generates. Without access to that history, the receiver cannot predict the expert's specific word choices.

3.5 Inter-agent agreement and domain coherence

Table 3: Domain coherence: mean PPL when non-expert primed agents predict the expert's continuation.

Domain	Non-expert PPL	Interpretation
Medical	8.0	High coherence (predictable continuations)
Science	9.0	High coherence
Legal	27.9	Low coherence (distinctive continuations)
Code	27.4	Low coherence

Medical and science expert continuations are relatively predictable by non-expert agents (PPL ≈ 8–9), while legal and code continuations are highly distinctive (PPL ≈ 28). This mirrors the priming selection results: science-primed agents can partially predict medical and science expert continuations but not legal and code ones. Legal and code reasoning creates the most domain-specific processing trajectories.

3.6 Ensemble characterization

The Mahalanobis distance between expert activations and ensemble centroids in the INLP subspace (36 dimensions) averages 6.46, with low cross-domain variance (science: 6.04, code: 6.81). The ensemble spread in the INLP subspace (mean standard deviation across agents) is 0.58. These values indicate that contextual priming creates genuine activation variation in the domain-discriminative subspace, confirming that priming histories function as activation-shaping inputs rather than superficial text changes.

4 Discussion

4.1 Coordination as programming, not transmission

The central result — priming selection outperforms centroid injection by 5.4× — reframes multi-agent coordination from a transmission problem to a programming problem. Injection attempts to write a state snapshot to one layer and relies on the remaining 17 layers to preserve it. Priming selection runs the receiver's *entire* forward pass through a structured input designed to produce the target processing state. One modifies a single register; the other executes a program.

This explains why injection is consistently weak across Papers XVI–XIX despite preserving representational geometry (Paper XVII). The geometry is correct but the processing state is wrong: 17 subsequent transformer layers, each performing attention over the text tokens, reshape the injected activation to match the text content. The injection is not preserved — it is dominated.

Priming 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.

4.2 Why Socratic scaffolding fails

Socratic scaffolding (Condition C) worsens performance by 19.9%. The analytical summary ("This text belongs to the medical domain and applies a classification framework...") displaces the conversational priming that would have been more effective. The scaffold provides *propositional* content (domain identity, framework name) but not *processing* content (the multi-turn reasoning pattern that shapes how the model engages with the probe).

This distinction is critical: domain identification alone is near-useless (Paper XVI showed text achieves 95% classification without any injection). What matters is the specific *reasoning trajectory* the expert took, and an analytical summary strips this trajectory down to its propositional skeleton.

4.3 The vocabulary mechanism

Shared vocabulary induction (Condition D, 19.5% gap closure) outperforms centroid injection (9.1%) by 2.1×. Structured labels — "Domain: Lymphoma. Key concepts: Non-Hodgkin lymphoma, cell of origin, clinical behavior. Framework: Classification" — provide semantic anchors that orient the receiver's processing without prescribing the reasoning trajectory.

The labels function as compressed priming: they activate domain-specific representations through the model's embedding and attention layers, without the overhead of a full conversational history. They lose information relative to full priming (19.5% vs. 48.9%) because the compression discards the multi-turn reasoning structure that shapes processing dynamics.

4.4 Implications for coordination protocol design

1. **Run the forward pass, don't bypass it.** Any coordination signal that must survive the forward pass is weaker than a signal that *is* the forward pass input. Design input sequences, not injection vectors.

2. **Process-level priming beats propositional framing.** Conversational priming ("Here is how we analyzed a similar problem...") is 2.5× more effective than analytical framing ("The domain is X and the framework is Y").

3. **Naming provides portable coordination.** Structured labels survive outside the priming context and can be deployed in arbitrary future conversations. They are less effective than full priming (19.5% vs. 48.9%) but infinitely more portable.

4. **Centroid injection is weakly positive, not negative.** Properly measured (cross-entropy, not arithmetic mean PPL), centroid injection closes 9.1% of the gap. Small but non-zero. Activation-level communication is real but secondary to text-level coordination.

5. **Domain-specific efficacy varies radically.** Priming selection closes 86% of the gap for medical and science but only 22% for code and legal. Coordination protocol design must account for domain-specific processing divergence.

5 Conclusion

Priming selection — choosing the right structured input sequence based on activation-space proximity — closes 48.9% of the gap between neutral and expert coordination, outperforming activation injection by 5.4×. Shared vocabulary induction closes 19.5%. Socratic scaffolding worsens performance. The ranking is unambiguous: the quality of multi-agent coordination depends on the structure of the shared input, not the bandwidth or fidelity of activation transmission.

The result reframes the activation geometry program. Papers XVI–XVIII asked: how much internal state can be transmitted between agents? The answer is: almost none, functionally. Paper XIX asks the complementary question: how effectively can shared input produce convergent processing? The answer is: substantially, when the input sequence is structured to match the target processing mode.

The 36.3% perplexity gap identified in Paper XVIII remains the binding constraint for text-based coordination. This paper demonstrates that the gap is partially addressable through input design (priming selection) rather than state transmission (activation injection). The remaining gap — 51.1% of the expert-to-baseline interval — represents the information carried exclusively by the expert's specific priming history: the multi-turn reasoning structure, the implicit processing biases, and the contextual associations that cannot be reconstructed from domain labels, structured annotations, or ensemble statistics alone.

Data Availability

All results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper19/`.

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