The Coordination Problem Is Interference: Nine Experiments on What Not to Do

Jeremy McEntire[1]

March 2026

Abstract

This paper synthesizes nine experiments (Papers XVI–XXIV) that systematically tested mechanisms for coordinating LLM agents. The experiments began by measuring the channel (XVI–XVIII), reframing the problem as programming rather than transmission (XIX), decomposing the effective protocol (XX), eliminating adaptive alternatives (XXII), identifying the mechanism (XXIII), and testing scalability (XXIV).

Every mechanism that *adds* to the receiver's context — injection, extended priming, repetition, explicit instruction, forced naming, adaptive guidance — either adds nothing or adds interference. The two mechanisms that work both operate by *removing*: context fence installs a processing mode boundary (-39% CE); minimal domain priming removes domain ambiguity with the fewest possible tokens (15 tokens capture 98.8% of the benefit).

Paper XXIII refined the mechanism: the context fence is mode-switching, not garbage collection. L2 distance to neutral does not predict coordination quality ($\rho = 0.100$). Reset benefit is constant ($\sim$0.076 nats) regardless of prior context length. The two stages — fence (mode boundary) and prime (domain activation) — compose independently ($\rho = 0.858$) because they target orthogonal dimensions of the processing state.

Paper XXIV confirmed scalability: chain degradation saturates by hop 5. A 10-hop chain performs identically to a 5-hop chain. Intermediate fences are pure overhead (0.001 nats difference). The protocol scales to arbitrary depth without compounding error.

1 The Arc of the Argument

The nine papers trace a progression from channel measurement to protocol design to mechanism identification to scalability confirmation. The progression was not planned in advance. Each paper tested the most promising extension of the prior result, and each time the extension failed, the constraint space tightened. By Paper XXII, the constraint space had collapsed to a single protocol. Papers XXIII and XXIV then tested the mechanism and confirmed it scales.

[1]Correspondence: `jmc@cageandmirror.com`

1.1 Phase 1: Measuring the channel (Papers XVI–XVIII)

Paper XVI: INLP Projection Transmission. Can domain-specific activation geometry survive text-mediated transmission? Result: text achieves 95% domain classification without any activation sidecar. INLP injection adds 1.9%. Sender and receiver encode domain information in nearly orthogonal directions (cosine similarity = 0.27). Procrustes alignment fails (residual > 1). *Conclusion: the geometric channel across the natural language boundary is nearly closed, but the functional channel is wide open.*

Paper XVII: Full Mind Transfer. Does increasing bandwidth help? Result: text-based conditions cluster at RSA ≈ 0.11 regardless of injection bandwidth (0–3,584 dimensions). Activation injection without text achieves RSA = 0.47 (4× higher geometric fidelity) but zero functional benefit. *Conclusion: text and activations carry complementary information. Geometry and function are dissociated. You cannot get both through a single channel.*

Paper XVIII: Sender Continuation Perplexity. What does the text channel actually preserve? Result: text summary preserves 63.7% of reasoning trajectory (PPL = 3.05 vs. original context PPL = 1.59). Activation injection adds $\leq$1.2% over text alone. Full activation without text is indistinguishable from no context (PPL = 5.60 vs. 5.62). *Conclusion: text preserves reasoning trajectory; activations alone carry zero reasoning information. The 36.3% residual gap is the binding constraint.*

The channel measurement phase established three constraints:

1. Geometric fidelity and functional fidelity are dissociated across the NL boundary.
2. Text carries function; activations carry geometry. Neither carries both.
3. Activation injection adds $\leq$1.2% over text. The remaining gap is not a bandwidth problem.

1.2 Phase 2: Reframing the problem (Paper XIX)

Paper XIX: Ensemble Gravity. If injection doesn't work, what does? Five primed agents process each probe. Four coordination mechanisms tested: centroid injection (averaging activations), priming selection (choosing the best-primed agent), Socratic scaffolding (guided questioning), shared vocabulary (self-generated labels).

Result: priming selection closes 48.9% of the coordination gap. Centroid injection closes 9.1%. Shared vocabulary closes 19.5%. Socratic scaffolding *worsens* performance by 19.9%. The ratio: priming selection outperforms injection by 5.4×.

Conclusion: coordination is not a transmission problem. It is a programming problem. The question is not "what information should I send to Agent 2?" but "what input sequence

should I run Agent 2's forward pass through?"

This reframe changed the research direction. Papers XVI–XVIII had been measuring channel capacity. Paper XIX showed the channel was the wrong abstraction. The operative mechanism is not information transfer but processing mode alignment.

1.3 Phase 3: Decomposing the protocol (Paper XX)

Paper XX: Ritual Shape. What structural features of a priming sequence produce the coordination benefit? Five features varied systematically: length, structural regularity, within-session repetition, vocabulary density, cold-start reset.

Five results, each eliminating a mechanism:

Table 1: Paper XX feature decomposition. Each row tests one structural feature of priming sequences.

Feature	Finding	Magnitude	Status
Length	15 tokens capture 98.8% of benefit	2.705/2.738 nats	Saturated
Regularity	Natural conversation > rigid Q&A	+0.475 nats	Format matters
Repetition	Degrades monotonically per repeat	+0.07 nats/rep	Eliminated
Vocabulary	Forced naming worst; natural best	+0.568 nats	Eliminated
Reset	"Disregard prior context" before prime	−0.464 nats (39%)	**Dominant**

The headline: reset + domain priming achieves CE = 0.733, beating the expert's own priming baseline (CE = 1.197) by 39%. The receiver with a clean context and domain priming *better predicts the expert's continuation* than the expert does from its own starting state.

Conclusion: the effective protocol is reset, then prime, then deliver. The reset is the single largest intervention. Domain identification requires only 15 tokens. Everything else — length, repetition, naming, structural regularity — is either neutral or harmful.

1.4 Phase 4: Eliminating the obvious extension (Paper XXII)

Paper XXII: Shepherd Agents. Paper XX used fixed priming. Does adaptive, probe-specific priming improve over fixed? Three shepherd strategies tested: storyteller (indirect narrative), provocateur (challenge and reframe), director (explicit instruction).

The gradient across shepherd types is the sharpest single result in the series. It is a controlled dose-response: more explicit instruction produces more displacement from the expert. Storyteller (−12%), provocateur (−27%), director (−37%). The model is not failing to understand the shepherd. It is understanding it too well. The shepherd content activates instruction-following mode, narrative mode, analytical mode — all of which compete with the target domain-processing mode the prime is trying to establish.

Table 2: Paper XXII shepherd results. Gap closure measured against the neutral-to-expert gap (0.734 nats).

Condition	CE	PPL_{geo}	Gap closure
Reset + fixed priming	**0.733**	**2.08**	+163%
Full stack (reset + story + prime)	0.851	2.34	+147%
Fixed priming	1.197	3.31	+100%
Reset + director	1.249	3.49	+93%
Reset + storyteller	1.277	3.59	+89%
Reset + neutral	1.278	3.59	+89%
No coordination	1.931	6.90	0%
Storyteller	2.021	7.55	−12%
Provocateur	2.127	8.39	−27%
Director	2.203	9.06	−37%

With reset prepended, all three shepherd strategies converge to within 0.03 nats of reset-alone performance. The shepherd content contributes nothing once the reset has done its work. The ratio: reset accounts for 22:1 of the benefit over the best shepherd content.

Inserting a storyteller between reset and domain priming degrades the result by +0.118 nats. The shepherd re-contaminates the context the reset just cleared.

Conclusion: adaptive priming is not just unnecessary — it is actively harmful. The protocol admits no intermediate steps between reset and prime. Content between the clearing operation and the domain activation is a contaminant.

1.5 Phase 5: Identifying the mechanism (Paper XXIII)

Paper XXIII: Context Fences. The throughline through Papers XVI–XXII suggested the reset operates by "garbage collection" — clearing residual activation biases. Paper XXIII tested this directly with nine clearing strategies, five prior context lengths, and a composability matrix.

The garbage collector hypothesis predicts: (1) L2 distance to neutral should predict coordination quality; (2) reset benefit should scale with prior context length; (3) more tokens in the clearing instruction should hurt (adding content = adding contamination).

All three predictions failed.

1. L2 distance to neutral does *not* predict CE ($\rho = 0.100$, $p = 0.798$). Closeness to neutral is irrelevant.

2. Reset benefit is *constant* ($\sim$0.076 nats) regardless of whether 0 or 384 tokens of prior context have accumulated ($\rho = -0.500$, $p = 0.391$).

3. Token count in the clearing instruction *negatively* correlates with CE ($\rho = -0.672$, $p = 0.047$). More tokens = better performance. The opposite of garbage collection.

The correct mechanism is **mode switching**. The context fence installs a processing mode boundary — it tells the model that what follows is epistemically separate from what came before. It requires both a backward reference ("disregard prior context") and a forward reference ("focus on what follows"). Silence fails because it carries no semantic content about the mode transition.

The composability result ties it together: fence quality and prime quality correlate with CE independently ($\rho = 0.858$, $p < 0.0001$). The two stages are separable because they target *orthogonal dimensions* of the processing state — mode boundary and domain orientation. This is why they compose: they don't interfere with each other.

Conclusion: the coordination protocol has two independent optimization targets. The fence installs a processing mode boundary. The prime installs domain orientation. They compose because they're orthogonal. Optimize each independently.

1.6 Phase 6: Testing scalability (Paper XXIV)

Paper XXIV: Hop Scaling. Does the bilateral protocol survive multi-hop chains? Sequential chains of 1–10 agents, where each relay agent generates a 48-token response that becomes the next agent's context. Three fence modes: fence at every hop, fence only at the terminal agent, no fences.

The headline finding: **chain degradation saturates**. CE rises from 0.503 at hop 1 to $\sim$0.60 by hop 5, then plateaus through hop 10. Neither linear ($R^2 = 0.50$) nor exponential ($R^2 = 0.48$) models fit — the shape is logarithmic.

Three secondary findings:

1. Fence at every hop vs. fence at final hop only: *identical* results (0.598 vs. 0.597 at 10 hops). Intermediate fences are pure overhead. The fence prepares the worker, not the relay chain.

2. Shared vocabulary substrate *hurts* at every chain length (+0.04 to +0.07 CE). Paper XX's saturation effect replicated at the chain level: don't pre-load agents, let each prime independently.

3. Cross-domain handoffs *outperform* within-domain chains (CE 0.56–0.63 vs. 0.72). Domain discontinuity acts as a natural mode reset.

Conclusion: the binding constraint on multi-agent architecture is not information-theoretic chain degradation (it saturates). It is the $O(n^2)$ coordinator complexity from organizational theory (the Graicunas limit). Build wide and shallow — not because deep chains fail, but because the coordinator's capacity is the bottleneck.

2 The Scorecard

Seven experiments tested every plausible coordination mechanism. The results form a single pattern.

Table 3: Complete mechanism scorecard across Papers XVI–XXIV. Mechanisms ranked by effect.

Mechanism	Paper	Effect	Verdict
Mechanisms that work (by removing)			
Context reset	XX	−0.464 nats (39%)	**Dominant**
Minimal domain priming (15 tok)	XX	98.8% of benefit	Works
Priming selection	XIX	48.9% gap closure	Works
Natural conversation format	XX	+0.475 vs rigid	Works
Self-generated vocabulary	XIX	19.5% gap closure	Works
Independent stage composition	XXIII	$\rho = 0.858$	Works
Terminal-only fencing (chains)	XXIV	0.001 nats diff	Works
Cross-domain handoffs	XXIV	CE 0.56 vs 0.72	Works
Mechanisms that fail (by adding)			
INLP activation injection	XVI	+1.9% over text	Marginal
Full activation injection	XVII, XVIII	≤1.2% over text	Dead
Centroid injection	XIX	9.1% gap	Marginal
Extended priming (>150 tok)	XX	Reverses by 300 tok	Harmful
Within-session repetition	XX	+0.07 nats/repeat	Harmful
Forced naming	XX	+0.568 nats	Harmful
Rigid Q&A structure	XX	+0.475 nats	Harmful
Socratic scaffolding	XIX	−19.9% gap	Harmful
Shepherd storyteller	XXII	−12.3% gap	Harmful
Shepherd provocateur	XXII	−26.7% gap	Harmful
Shepherd director	XXII	−37.1% gap	Harmful
Shepherd between reset & prime	XXII	+0.118 nats	Harmful
Shared vocabulary substrate	XXIV	+0.04–0.07 CE	Harmful
Intermediate fences in chains	XXIV	+0.001 nats (overhead)	Unnecessary
Mid-chain re-priming (pre-saturation)	XXIV	+0.007 nats	Harmful

The pattern: every mechanism in the "works" column operates by *removing* something. Context reset removes residual biases. Minimal priming removes domain ambiguity with the

fewest tokens. Priming selection removes the wrong agents from the coordination. Natural format removes artificial structural constraints. Self-generated vocabulary removes the mismatch between imposed and internal representations.

Every mechanism in the "fails" column operates by *adding* something. Injection adds activation vectors. Extended priming adds tokens. Repetition adds redundant content. Naming adds artificial labels. Shepherds add meta-commentary. Socratic scaffolding adds guided questions. In every case, the addition either provides no benefit or actively interferes with the target processing mode.

3 The Interference Interpretation

3.1 Why adding fails

The forward pass is a dynamical system. The model's processing trajectory through activation space is determined by the input token sequence. Each token shifts the trajectory. The expert generated its continuation from a specific processing trajectory. The coordination problem is to place the receiver on a trajectory that produces the same continuation.

Adding tokens to the receiver's context changes its trajectory. Every added token is a perturbation. For the perturbation to help, it must push the trajectory *toward* the expert's. For domain-matched priming, this works: the first 15 tokens of domain content push the trajectory into the right processing mode. But each additional token beyond saturation is a perturbation that pushes the trajectory *away* from the expert's, because the expert did not process those tokens.

This explains the non-monotonic length curve (Paper XX): benefit peaks at ∼100–150 tokens and reverses by 300. It explains why shepherd outputs hurt: the expert never processed a storyteller's narrative or a director's instructions. The receiver's trajectory diverges from the expert's with every token of content the expert did not see.

The reset instruction is special because it is semantically inert relative to domain content. It does not push the trajectory toward any specific processing mode. It pushes it toward *neutral* — the starting state from which domain priming has maximum leverage. The reset removes accumulated trajectory perturbations without adding new ones.

3.2 The context fence (updated by Paper XXIII)

Paper XXII eliminated the intuitive architecture: a sophisticated intermediate agent that prepares the receiver. The initial interpretation — that the reset operates as a "garbage collector" clearing residual activation biases — was a testable hypothesis. Paper XXIII tested

it and found it wrong.

The reset is a **context fence**: a mode-switching boundary, not a garbage collector. Three pieces of evidence:

1. L2 distance to neutral does not predict coordination quality ($\rho = 0.100$, $p = 0.798$). If the reset were clearing residual biases, closeness to neutral should predict performance.

2. Reset benefit is constant ($\sim$0.076 nats) regardless of prior context length (0–384 tokens, $\rho = -0.500$, $p = 0.391$). A garbage collector should work harder with more garbage.

3. Token count in the clearing instruction negatively correlates with CE ($\rho = -0.672$, $p = 0.047$). More semantic content about the mode transition = better switching.

This gives two measurable objectives targeting *orthogonal* dimensions:

- **Fence quality**: install a processing mode boundary that separates prior context from the task ahead.

- **Priming quality**: maximize alignment of $\mathbf{h}_{\text{post-prime}}$ with $\mathbf{h}_{\text{expert}}$ — activate the target domain.

The stages compose independently ($\rho = 0.858$, $p < 0.0001$) because mode boundary and domain orientation are orthogonal operations. Optimize each separately.

3.3 Why the expert's own priming is suboptimal

The most counterintuitive result: reset + priming (CE = 0.733) beats the expert's own priming (CE = 1.197) by 39%. The receiver predicts the expert's continuation *better than the expert's own starting state would predict it.*

The mechanism: the expert generated its continuation from a processing state that includes residual biases from model initialization and default processing modes. These biases are part of the expert's trajectory but do not contribute to the continuation's logic. They are noise in the starting state.

The reset instruction suppresses these biases. The receiver's "cleaner" processing of the same domain priming produces a state more aligned with the expert's *reasoning trajectory* than the expert's own starting state before the trajectory stabilized. The reset removes the noise; the priming delivers the signal. The student outperforms the teacher after receiving the teacher's instruction, because the instruction crystallizes knowledge the teacher arrived at through noisier exploration.

4 The Protocol

The seven experiments converge on a single coordination protocol:

Reset → Prime → Deliver

Each step has a distinct functional role:

1. **Reset** (∼15 tokens). Clear residual processing biases. Bring activations to neutral baseline. The clearing instruction must be semantically inert — it tells the model to forget, not to think. "This is a new conversation. Disregard prior context." Content-free by design.

2. **Prime** (15–50 tokens). Activate the target processing mode with minimal domain-specific content. Natural conversational format. No forced naming, no structural regularity, no explicit instruction. The vocabulary should match the expert's natural domain language exactly — neither stripped nor augmented.

3. **Deliver**. Present the task into the prepared context. No additional coordination content between prime and delivery.

What the protocol excludes, by experimental evidence:

- No content between reset and prime (Paper XXII: +0.118 nats).
- No content between prime and delivery.
- No activation injection at any layer (Papers XVI–XVIII: ≤1.2%).
- No extended priming beyond ∼50 tokens (Paper XX: reversal by 300).
- No within-session repetition (Paper XX: +0.07/repeat).
- No forced vocabulary (Paper XX: +0.568 nats).
- No rigid structural format (Paper XX: +0.475 nats).
- No adaptive probe-specific priming (Paper XXII: −12% to −37%).
- No Socratic scaffolding (Paper XIX: −19.9%).

The protocol is approximately 30 tokens of coordination overhead. It achieves 163% gap closure from baseline. No more elaborate alternative tested in this series matches it.

5 Predictions

The interference interpretation generates three falsifiable predictions:

Prediction 1: Coordination overhead should be inversely correlated with coordination quality. In any multi-agent system, the agents with the most tokens of inter-agent coordination content should show the worst alignment with expert processing. The relationship should be monotonic, as it was for shepherd output length (30 tokens storyteller $>$ 34 provocateur $>$ 114 director).

Prediction 2 (FALSIFIED by Paper XXIII): The reset benefit should be proportional to prior context length. This prediction assumed the garbage collector mechanism: more context = more residual bias to clear = larger reset benefit. Paper XXIII tested this directly with five prior context lengths (0, 19, 93, 271, 384 tokens). Result: reset benefit is *constant* at $\sim$0.076 nats regardless of context length ($\rho = -0.500$, $p = 0.391$). This falsified the garbage collector hypothesis and established the mode-switching mechanism. The reset installs a processing boundary, and boundaries cost the same signal regardless of what came before.

Prediction 3: Cross-session repetition of the reset-prime cycle should compound. Paper XX showed within-session repetition degrades (+0.07 nats/repeat). But the effective "repetition" in institutional coordination protocols is repetition of the *reset-prime cycle across sessions*, each starting from a cleared context. Weekly mass, recurring standups, regular 1:1s. The prediction: repeated exposure to reset-prime across sessions should show stable or improving coordination, while repeated exposure within a session should continue to degrade.

6 Limitations

Single model family. All experiments used Qwen 2.5-7B. The structural results (direction of effects) should generalize to other autoregressive transformers, but the magnitudes are model-specific.

Single coordination metric. Continuation perplexity via KV-cache teacher-forcing measures reasoning trajectory alignment. It does not measure task completion, factual accuracy, or downstream utility. A protocol that optimizes for trajectory alignment may not optimize for all coordination goals.

Synthetic priming sequences. All priming sequences were constructed for the experiments. Real-world coordination contexts (CLAUDE.md files, agent handoff briefs, system prompts) have different structural properties. The magnitudes of the effects may differ in

practice.

Fixed expert. The expert is defined as the domain-matched primed agent's greedy output. This is a reasonable proxy but not a ground truth. The reset+prime condition beating the expert baseline raises the question of what "expert" means when the receiver outperforms the expert's own starting state.

No cross-model coordination. All experiments used the same model as both sender and receiver. Cross-model coordination (different architectures, different scales) introduces additional alignment challenges not addressed here.

7 Conclusion

Seven experiments, one converging answer. The coordination problem is not about information — it is about interference. Every mechanism that adds to the receiver's context either adds nothing or adds interference. The two mechanisms that work both operate by removing: reset removes residual context, minimal priming removes domain ambiguity with the fewest possible tokens.

The optimal coordination program is the one with the fewest instructions.

Data Availability

All experimental results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper16/` through `paper24/`.

Series: Activation Geometry of Domain-Selective Noise Injection, Throughline II (Papers XVI–XXIV).