# Context Fences:

# Two-Phase Coordination Through Mode Switching and Domain Activation

Jeremy McEntire[1]

March 2026

**Abstract**

Papers XIX–XXII established that the effective coordination protocol is reset, then prime, then deliver. The throughline interpretation framed the reset as a garbage collector: clearing residual activation biases, with the optimal clearing signal approaching zero content. This paper tests that framing directly by varying nine clearing strategies across four experiments.

The garbage collector hypothesis is falsified on three counts. (1) Activation distance to neutral does not predict coordination quality ($\rho = 0.100$, $p = 0.798$). (2) More tokens in the clearing instruction produce *better* results ($\rho = -0.672$, $p = 0.047$), not worse. (3) Reset benefit is constant regardless of prior context length (∼0.07–0.09 nats from 0 to 384 tokens of history; $\rho = -0.500$, $p = 0.391$).

The correct mechanism is mode switching, not garbage collection. The reset instruction does not remove residual context — it installs a processing mode boundary. The model is told that what follows is epistemically separate from what came before. Silence fails (CE = 0.699, worst) because it carries no signal about processing mode, leaving the model in continuation mode where the subsequent prime must fight the continuation interpretation. The standard reset succeeds (CE = 0.503, best) because it simultaneously names what to ignore and what to attend to, establishing the boundary cleanly.

The two coordination stages — mode boundary and domain activation — are separable: post-prime activation distance to expert predicts CE with $\rho = 0.858$ ($p < 10^{-4}$). Mode and domain are orthogonal dimensions of the processing state. The optimal coordination protocol for multi-agent systems is: install a context fence, then activate the domain, then deliver the task.

## 1 Introduction

Papers XVI–XXII traced the coordination problem from channel measurement to protocol design. The result was a three-step protocol — reset, prime, deliver — that achieves 163%

[1] Correspondence: jmc@cageandmirror.com

gap closure from baseline. The throughline interpretation (Throughline II) proposed a specific mechanism: the reset is a garbage collector that brings activations to the neutral baseline, and the primer then pushes from neutral toward the expert. The garbage collector's objective: minimize $\|\mathbf{h}_{\text{post-clear}} - \mathbf{h}_{\text{neutral}}\|$. Its design principle: output the minimum content necessary for clearing, because every token is a new contaminant.

This paper tests that hypothesis directly. Four experiments:

1. **Clearing strategies**: Nine strategies varying from silence to extended instructions. Does activation distance to neutral predict coordination quality?
2. **Prior context length**: Varying the amount of residual context before clearing. Does reset benefit scale with the amount of accumulated bias?
3. **Composability**: Crossing clearing strategies with priming strategies. Are the two stages separable?
4. **Content analysis**: Does the clearing instruction's token count predict its effectiveness? Is the zero-content clearing optimal?

The garbage collector hypothesis predicts: (1) closer to neutral after clearing = better CE; (2) more prior context = more reset benefit; (3) fewer clearing tokens = better results. All three predictions are tested. All three are falsified.

## 2 Methods

All experiments use Qwen 2.5-7B with the Paper XIX protocol: continuation perplexity of a domain-matched expert's 64-token greedy output via KV-cache teacher-forcing. The domain priming held constant across conditions is the base priming from Paper XIX ($\sim$119 tokens per domain). Layer-10 activations are captured to measure geometric distances. 160 domain probes (40 per domain: medical, legal, code, science).

### 2.1 Experiment 1: Clearing strategies

Nine clearing strategies are tested, each followed by the same domain-matched priming:

For each strategy, two measurements are taken: (a) layer-10 activation distance to neutral (clearing quality, the garbage collector's objective) and (b) continuation cross-entropy after clearing + domain priming (coordination quality).

Table 1: Clearing strategies. Token count from Qwen 2.5-7B tokenizer.

| Strategy | Tokens | Content |
|---|---|---|
| Silence | 1 | Four newlines |
| Boundary | 1 | “---” |
| Single word | 2 | “New.” |
| Minimal | 2 | “Begin.” |
| Standard reset | 18 | “The following is a new conversation... Disregard... ” |
| Topic change | 30 | Unrelated Q&A exchange |
| Double reset | 36 | Standard reset repeated twice |
| Extended reset | 45 | Verbose instruction to clear prior context |
| No clearing | 0 | Empty string (direct priming) |

### 2.2 Experiment 2: Prior context length

Five levels of prior context (0, 19, 93, 271, 384 tokens) precede the clearing instruction. Each level is tested with and without the standard reset. The prediction: if the reset clears accumulated biases, its benefit should scale with prior context length.

### 2.3 Experiment 3: Composability

Six clearing strategies are crossed with three priming types (domain-matched, neutral, none), yielding an 18-cell matrix. Post-prime layer-10 activations are captured for all 18 conditions to measure distance to expert. If the two stages are separable, clearing quality and priming quality should independently predict final CE.

### 2.4 Experiment 4: Content analysis

The nine clearing strategies from Experiment 1 are analyzed by token count. The garbage collector prediction: fewer tokens = better clearing (less contamination). The mode-switching prediction: semantic precision matters more than brevity.

## 3 Results

### 3.1 Experiment 1: Distance to neutral does not predict CE

The garbage collector’s objective — activation distance to neutral — does not predict coordination quality: Spearman $\rho = 0.100$ ($p = 0.798$). The no-clearing condition ($\mathrm{L2}_{\mathrm{neutral}} = 0$, cosine = 1.000 — identical to neutral by definition) ranks fifth, not first. Silence (L2

Table 2: Clearing strategies ranked by coordination quality (CE). All conditions include domain-matched priming after clearing. L2 and cosine are distances to neutral activations.

| Strategy | Tokens | CE | $\mathrm{PPL}_{\mathrm{geo}}$ | $\mathrm{L2}_{\mathrm{neutral}}$ | $\cos_{\mathrm{neutral}}$ |
|---|---|---|---|---|---|
| Standard reset | 18 | **0.503** | **1.65** | 22.95 | 0.911 |
| Double reset | 36 | 0.524 | 1.69 | 23.27 | 0.908 |
| Topic change | 30 | 0.573 | 1.77 | 27.43 | 0.874 |
| Extended reset | 45 | 0.592 | 1.81 | 24.60 | 0.899 |
| No clearing | 0 | 0.596 | 1.81 | 0.00 | 1.000 |
| Minimal | 2 | 0.601 | 1.82 | 22.04 | 0.918 |
| Single word | 2 | 0.610 | 1.84 | 23.04 | 0.910 |
| Boundary | 1 | 0.612 | 1.84 | 27.40 | 0.878 |
| Silence | 1 | 0.699 | 2.01 | 23.42 | 0.907 |

= 23.42) and standard reset (L2 = 22.95) have nearly identical distances to neutral but CE values separated by 0.196 nats.

The hypothesis that clearing quality equals proximity to the neutral activation state is falsified. The clearing stage is doing something other than bringing activations to baseline.

### 3.2 Experiment 2: Reset benefit is constant across prior context lengths

Table 3: Reset benefit across prior context lengths. Benefit = CE(no reset) − CE(with reset).

| Prior context | Tokens | CE (no reset) | CE (reset) | Benefit |
|---|---|---|---|---|
| None | 0 | 0.596 | 0.503 | 0.093 |
| Short | 19 | 0.553 | 0.486 | 0.067 |
| Medium | 93 | 0.607 | 0.528 | 0.080 |
| Long | 271 | 0.638 | 0.574 | 0.064 |
| Very long | 384 | 0.663 | 0.588 | 0.075 |

Reset benefit is approximately constant at $0.076 \pm 0.011$ nats across prior context lengths ranging from 0 to 384 tokens. The correlation between prior length and reset benefit is non-significant ($\rho = -0.500$, $p = 0.391$).

If the reset were clearing accumulated activation biases, benefit should scale with the amount of accumulated context: more history means more biases to clear, more benefit from clearing. It does not scale. The reset provides the same benefit whether there is one prior turn or five.

This falsifies the garbage collection mechanism. The reset is not clearing residual context. It is performing a constant-cost operation: installing a processing mode boundary. The cost of the mode switch is independent of what preceded it, just as a stack frame push costs the same regardless of how many frames are already on the stack.

Note that CE *without* reset does degrade with prior context length (0.596 → 0.663, monotonically). Prior context creates interference. But the reset's benefit is not proportional to the interference — it is fixed. The reset does not undo the interference; it recontextualizes it as irrelevant.

### 3.3 Experiment 3: The two stages are separable

Table 4: Composability matrix: clearing strategy × priming type (CE). Each cell is an independent measurement.

| Clearing | Domain prime | Neutral prime | No prime |
|---|---|---|---|
| Standard reset | **0.503** | 0.742 | 0.900 |
| Extended reset | 0.592 | 0.843 | 1.027 |
| Minimal | 0.601 | 0.918 | 1.188 |
| No clearing | 0.596 | 0.977 | 1.693 |
| Boundary | 0.612 | 0.864 | 1.172 |
| Silence | 0.699 | 1.025 | 1.041 |

The composability matrix reveals clean separability. Within each row, domain priming always outperforms neutral which always outperforms none. Within each column, standard reset always outperforms other clearing strategies (with minor exceptions for the minimal-vs-no-clearing comparison). The stages do not interact: the best clearing paired with the best priming produces the best CE, regardless of which specific strategies are used.

Post-prime activation distance to expert predicts CE with $\rho = 0.858$ ($p < 10^{-4}$) across all 18 conditions. The Stage 2 metric — proximity to expert after the full protocol — captures the coordination outcome.

The separability has a structural interpretation: mode and domain are orthogonal dimensions of the processing state. The clearing stage operates on the mode dimension (continuation vs. new context). The priming stage operates on the domain dimension (medical vs. legal vs. code vs. science). Because these dimensions are orthogonal, the stages compose independently.

### 3.4 Experiment 4: Semantic precision, not brevity

Token count is negatively correlated with CE ($\rho = -0.672$, $p = 0.047$): more tokens in the clearing instruction produce better results, not worse. This falsifies the zero-content prediction.

Table 5: Clearing strategies ordered by token count. The garbage collector prediction: fewer tokens = better CE. The observed pattern: semantic precision = better CE.

| Strategy | Tokens | CE | Semantic content | Mode signal |
|---|---:|---:|---|---|
| No clearing | 0 | 0.596 | None | None |
| Silence | 1 | 0.699 | None | Ambiguous (continuation?) |
| Boundary | 1 | 0.612 | Visual separator | Ambiguous |
| Single word | 2 | 0.610 | “New” | Partial |
| Minimal | 2 | 0.601 | “Begin” | Partial |
| Standard reset | 18 | **0.503** | Ignore prior + attend to next | **Complete** |
| Topic change | 30 | 0.573 | Unrelated exchange | Implicit |
| Double reset | 36 | 0.524 | Reset instruction ×2 | Complete (redundant) |
| Extended reset | 45 | 0.592 | Verbose clearing | Diluted |

But the relationship is not monotonic with length. The standard reset (18 tokens) outperforms the extended reset (45 tokens). The pattern follows semantic precision, not token count:

- **Silence** (worst): no signal. The model interprets the continuation of whitespace as continuation of the prior context. The subsequent prime arrives in continuation mode and must fight the continuation interpretation.

- **Boundary / single word**: ambiguous signals. “—” and “New.” could mark a continuation, a section break, or a topic change. The model does not reliably interpret these as mode boundaries.

- **Standard reset** (best): precise signal. “The following is a new conversation on a different topic. Disregard any prior context.” performs two operations simultaneously: names what to ignore (prior context) and names what to attend to (what follows). Both halves are necessary to establish the boundary.

- **Extended reset** (worse than standard): correct signal, diluted by verbosity. The additional tokens (“Clear your mind of all prior topics, themes, and assumptions. . . ”) add processing without adding precision. The mode boundary was already established; the extra content is noise.

- **Double reset**: near-standard performance. The second reset is redundant (Paper XX: within-session repetition degrades at +0.07 nats/repeat), but because the first reset already installed the boundary, the degradation is smaller here (0.021 nats) than for content that competes with domain processing.

# 4 Discussion

## 4.1 Not garbage collection: mode switching

The garbage collector framing predicted that the clearing stage removes residual activation biases, with effectiveness measured by proximity to the neutral baseline. Three results falsify this:

1. L2 distance to neutral does not predict CE ($\rho = 0.100$).
2. Reset benefit is constant regardless of prior context length ($\sim$0.076 nats).
3. The zero-content clearing (silence) is worst, not best.

The correct mechanism: the reset instruction installs a processing mode boundary. It does not erase residual context — the model's attention mechanism still has access to prior tokens. Instead, it tells the model that what follows is epistemically separate from what came before. This is a different operation from garbage collection. Garbage collection frees memory. Mode switching pushes a new stack frame: a signal that says "new execution context begins here."

The constant reset benefit across prior context lengths is the deep result. If the operation were clearing accumulated biases, it should scale: more biases to clear, more benefit from clearing. It does not scale because the operation is not clearing. The mode switch costs the same signal regardless of how much context preceded it. The model does not need to "forget" 384 tokens of monetary policy discussion; it needs to be told that the monetary policy discussion is no longer the operative context.

## 4.2 The context fence

The revised framing: the clearing stage installs a **context fence** — a semantic boundary that separates prior context from subsequent context. The fence's effectiveness depends on the precision of its signal, not on its brevity.

An effective context fence requires two components:

1. **Backward reference**: name what to ignore. "Disregard any prior context" tells the model that the preceding tokens are not relevant to what follows.
2. **Forward reference**: name what to attend to. "The following is a new conversation on a different topic" tells the model what the subsequent tokens represent.

Silence fails because it carries neither component. The model in continuation mode interprets subsequent tokens as more of the same conversation. The prime arrives in a context where it must compete with the continuation interpretation rather than building on a clean mode boundary.

"Begin." and "—" carry partial forward reference (something new is starting) but no backward reference (nothing says to ignore what came before). The model partially enters a new mode but retains the prior context as potentially relevant.

The standard reset carries both components and succeeds. The extended reset carries both but dilutes them with unnecessary elaboration. The optimum is not zero content — it is precisely the content necessary to establish both halves of the fence.

### 4.3 Orthogonality of mode and domain

The composability result ($\rho = 0.858$ between post-prime L2-to-expert and CE) confirms that the two stages target orthogonal dimensions of the processing state:

- **Stage 1 (fence)**: operates on the *mode* dimension. Continuation mode vs. new-context mode. Binary. Requires a semantically precise signal.
- **Stage 2 (prime)**: operates on the *domain* dimension. Medical vs. legal vs. code vs. science vs. neutral. Requires domain-specific vocabulary.

Because mode and domain are orthogonal, the stages compose independently. The best fence paired with the best prime produces the best coordination (CE = 0.503), and this optimum is achieved by optimizing each stage for its own objective without regard to the other.

This decomposition is the engineering result. Multi-agent systems can optimize their coordination protocol by independently tuning two components:

1. The context fence: a fixed, semantically precise instruction that installs a processing mode boundary. Once this instruction is established, it need not be changed per-task or per-agent. It is infrastructure.
2. The domain primer: 15–50 tokens of domain-specific vocabulary (Paper XX). This varies by task domain but not by specific task within a domain.

Neither component requires adaptive content. Neither benefits from sophistication. The optimal coordination overhead is $\sim$30 tokens: $\sim$18 for the fence and $\sim$15 for the primer.

### 4.4 Implications for multi-agent coordination

The coordination problem that plagues AI agent swarms has a measurable solution. The problem is not that agents cannot share information — text transmits 63.7% of reasoning trajectory (Paper XVIII). The problem is that agents accumulate processing mode contamination across turns, and the contamination degrades coordination even when the information is correct.

The solution:

1. **Every agent-to-agent handoff begins with a context fence.** A fixed instruction of ∼18 tokens that establishes a processing mode boundary. This is not per-task configuration; it is protocol infrastructure, the same for every handoff.

2. **After the fence, a minimal domain primer.** 15–50 tokens of domain-specific vocabulary. This activates the receiving agent's domain processing mode without competing with the fence.

3. **Then the task.** No additional coordination content.

4. **Nothing between the fence and the primer.** Paper XXII showed that content between reset and prime degrades by +0.118 nats. The fence and the primer should be adjacent.

5. **Measure post-handoff activation proximity to expert.** $\rho = 0.858$ with CE makes this a viable real-time quality metric for coordination.

The total coordination overhead: ∼30 tokens per handoff. The total benefit: 163% gap closure (Paper XX) to as low as CE = 0.503 (this paper). The cost of not doing it: agents in continuation mode process each task as an extension of prior conversation rather than as a fresh context, accumulating interference with every turn.

## 5 Conclusion

The garbage collector hypothesis — that the clearing stage should minimize activation distance to neutral — is falsified. The clearing stage is not a garbage collector; it is a context fence. It does not remove residual context; it installs a processing mode boundary.

The evidence: (1) distance to neutral does not predict coordination quality; (2) reset benefit is constant regardless of prior context length; (3) silence is worst because it carries no mode signal; (4) the effective instruction requires both backward reference (what to ignore) and forward reference (what to attend to).

The two coordination stages — mode boundary and domain activation — are separable ($\rho = 0.858$, $p < 10^{-4}$), operating on orthogonal dimensions of the processing state. They can be optimized independently. Neither benefits from adaptive content or sophistication.

The optimal coordination protocol for multi-agent systems: install a context fence (∼18 tokens), then activate the domain (∼15 tokens), then deliver the task. Approximately 30 tokens of total coordination overhead. The stages are fixed protocol infrastructure, not per-task configuration. The coordinator's only job is to say "new context, here's the domain, go" — and the phrasing of those 30 tokens is the entire coordination problem.

## Data Availability

All results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper23/`.

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