Shepherd Agents:
Adaptive Priming Through Directed Intervention

Jeremy McEntire[1]

March 2026

Abstract

Papers XIX and XX established that priming selection and context reset are the dominant coordination mechanisms, closing 48.9% and an additional 39% of the coordination gap respectively. Both used *fixed* priming: the same sequence regardless of the specific task. This paper tests whether *adaptive* priming — a shepherd agent that generates probe-specific coordination signals — improves over fixed approaches.

Three shepherd strategies are tested: a storyteller (indirect narrative analogy), a provocateur (challenge and reframe), and a director (explicit instruction). The result is unambiguous: all three shepherd strategies perform **worse than no coordination** when used bare (storyteller −12.3%, provocateur −26.7%, director −37.1% of gap). With a reset instruction prepended, all three converge to within 0.03 nats of reset-alone performance — the shepherd content contributes nothing. Adding a storyteller layer between reset and domain priming degrades the reset-prime result by 0.12 nats.

The conclusion: probe-specific adaptive priming does not improve coordination. The effective mechanism is the reset-prime protocol itself, not the sophistication of the priming content. More explicit instruction produces *worse* results, not better. The correct coordination protocol remains the simplest: reset, then domain-prime, then deliver.

1 Introduction

Paper XX identified the minimal effective coordination sequence: reset the receiver's context, then deliver domain-matched priming. But the domain priming in Paper XX was fixed — the same for all probes within a domain. A natural question: does *adaptive* priming, tailored to each specific task, outperform fixed domain-level priming?

The hypothesis has intuitive appeal. A fixed "the patient presented with elevated troponin" priming activates the medical domain generically. A probe-specific "this case involves distinguishing NSTEMI from unstable angina given troponin kinetics" should, in principle, activate a more precise processing mode. The shepherd agent sees the probe and generates a coordination signal designed to orient the receiver toward the specific reasoning required.

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

We test three shepherd philosophies, each corresponding to a distinct coordination tradition:

1. **Storyteller**: Indirect narrative analogy. “Let me tell you about a similar case. . . ” Contextualizes the domain through story, activating reasoning frames without stating the answer. Corresponds to parable, case-based teaching, the Socratic setup.

2. **Provocateur**: Challenge and reframe. “Most people approach this as X, but that’s wrong because. . . ” Forces the receiver to reconsider default assumptions, deepening processing. Corresponds to the koan, the devil’s advocate, the Socratic elenchus.

3. **Director**: Explicit instruction. “Focus on A, B, C. The key distinction is. . . ” Direct specification of where to attend and what matters. Corresponds to the briefing, the coach’s halftime adjustment, the explicit protocol.

If the storyteller (indirect) outperforms the director (explicit), processing depth matters more than information density. If the director wins, the field’s intuition about explicit coordination is correct.

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, measured via KV-cache teacher-forcing.

Phase 1: Shepherd generation. For each of 160 domain probes (40 per domain), the same Qwen 2.5-7B generates three shepherd outputs. Each shepherd prompt instructs the model to prepare a colleague for the given probe using the specified strategy. Outputs are truncated to 120 tokens maximum. Mean output lengths: storyteller 30 tokens, provocateur 34 tokens, director 114 tokens.

Phase 2: Condition evaluation. Twelve conditions are measured:

1. Storyteller alone (shepherd output + probe)

2. Provocateur alone

3. Director alone

4. Reset + storyteller

5. Reset + provocateur

6. Reset + director

7. Fixed domain priming (Paper XIX baseline)
8. Reset + fixed domain priming (Paper XX baseline)
9. No coordination (neutral priming)
10. Reset + neutral priming
11. Storyteller + fixed domain priming (combined)
12. Reset + storyteller + fixed priming (full stack)

The reset instruction is the 15-token sequence from Paper XX: "The following is a new conversation on a different topic. Disregard any prior context."

Phase 3: Activation analysis. Layer-10 activations are captured for neutral, storyteller, director, fixed priming, and reset+fixed priming conditions. L2 distance and cosine similarity to the expert's activations quantify where each condition places the receiver in activation space.

3 Results

3.1 Bare shepherd strategies are worse than no coordination

Table 1: All 12 conditions ranked by cross-entropy. Gap closure is measured relative to the no-coordination–to–expert-priming gap (0.734 nats).

Condition	CE	PPL_{geo}	Gap closure
Reset + fixed priming	**0.733**	**2.08**	+163.3%
Full stack (reset + story + prime)	0.851	2.34	+147.3%
Fixed priming (Paper XIX)	1.197	3.31	+100.0%
Reset + director	1.249	3.49	+93.0%
Reset + storyteller	1.277	3.59	+89.1%
Reset + neutral	1.278	3.59	+89.0%
Reset + provocateur	1.279	3.59	+88.9%
Storyteller + fixed priming	1.530	4.62	+54.6%
No coordination (neutral)	1.931	6.90	0%
Storyteller	2.021	7.55	−12.3%
Provocateur	2.127	8.39	−26.7%
Director	2.203	9.06	−37.1%

All three bare shepherd strategies perform *worse* than neutral priming. The ranking is inverse to explicitness: storyteller (indirect, CE = 2.02) is least bad; director (explicit, CE = 2.20) is worst. The director generates 3.3× more tokens than the storyteller (114 vs. 30 mean tokens), and every additional token of probe-specific meta-commentary degrades performance.

The shepherd outputs are not domain priming. They are meta-commentary *about* the domain. “Let me tell you about a similar case involving troponin kinetics” activates the narrative mode, not the medical reasoning mode. “Focus on the distinction between NSTEMI and unstable angina” activates the instruction-following mode, not the clinical reasoning mode. The model’s forward pass processes these as conversational turns, not as domain activation.

3.2 Reset equalizes all shepherd strategies

Table 2: Reset + shepherd vs. reset + neutral. Differences from reset-neutral baseline.

Condition	CE	$\mathrm{PPL}_{\mathrm{geo}}$	Δ vs reset+neutral
Reset + neutral	1.278	3.59	—
Reset + storyteller	1.277	3.59	-0.001
Reset + provocateur	1.279	3.59	$+0.001$
Reset + director	1.249	3.49	-0.029

With a reset instruction prepended, the three shepherd strategies converge to within 0.03 nats of reset-alone performance. The storyteller and provocateur are statistically indistinguishable from reset+neutral ($|\Delta| \leq 0.001$ nats). The director shows a marginal 0.029-nat advantage, possibly because its explicit instructions provide a small amount of domain-relevant vocabulary after the reset clears the context.

The implication: the reset instruction accounts for essentially all the coordination benefit in the reset+shepherd condition. The shepherd content is noise once the reset has done its work. The reset reduces CE from 1.93 to 1.28 (-0.65 nats). The best shepherd adds -0.03 nats on top. The ratio is 22:1 in favor of the reset mechanism.

3.3 Adding a shepherd layer degrades reset-prime

Inserting a storyteller layer between the reset instruction and the domain priming degrades performance by 0.118 nats. Without the reset, the storyteller+priming combination (CE = 1.53) is worse than priming alone (CE = 1.20) by 0.33 nats.

The storyteller output sits between the reset instruction and the domain primer in the token sequence. The attention mechanism processes the shepherd tokens before reaching the domain priming. By the time the model encounters the domain-specific content, its

Table 3: Effect of adding a storyteller layer to the reset-prime protocol.

Condition	CE	PPL_{geo}	Δ vs reset+prime
Reset + fixed priming	**0.733**	**2.08**	—
Full stack (reset + story + prime)	0.851	2.34	+0.118
Storyteller + fixed priming	1.530	4.62	+0.797

processing trajectory has already been shaped by the narrative framing. The reset cleaned the context; the shepherd re-dirtied it.

3.4 Activation distances confirm: shepherds move away from expert

Table 4: Layer-10 activation distance to expert for key conditions.

Condition	L2 distance	Cosine similarity
Fixed priming (= expert)	0.00	1.000
Reset + fixed priming	8.90	0.985
Neutral	21.61	0.912
Storyteller	28.15	0.860
Director	28.49	0.844

The storyteller and director both move activations *further* from the expert than neutral priming does (L2 28.1–28.5 vs. 21.6). The shepherd strategies are not converging the receiver toward the expert — they are actively diverging from it.

This explains the CE results mechanistically. The shepherd outputs push the model into a processing mode (narrative or instructional) that is further from the expert's domain-specific processing mode than the neutral starting point. The reset+priming condition has the second-closest activations (L2 = 8.9, cosine = 0.985), explaining its superior CE performance.

4 Discussion

4.1 Why adaptive priming fails

The failure of shepherd agents is not a failure of implementation. It is a structural result. The shepherd generates meta-commentary about the domain, not domain content. The model's forward pass distinguishes between:

- **Domain priming**: "The patient presented with elevated troponin and ST depression." This activates the medical processing mode directly.

- **Meta-commentary**: “Think about what happens when troponin rises in the context of chest pain.” This activates the instruction-following mode with medical vocabulary embedded within it.

The two are not equivalent. Domain priming programs the forward pass into the target processing mode. Meta-commentary programs the forward pass into a *different* mode (instruction-following, narrative, analytical) that happens to contain domain tokens. The processing trajectory differs even when the vocabulary overlaps.

Paper XX showed that forced naming (“TROPONIN CASCADE”) performs worse than natural domain vocabulary despite containing the same domain tokens. This paper extends the finding: even well-crafted probe-specific meta-commentary containing relevant domain vocabulary performs worse than fixed domain priming that is not probe-specific.

The mechanism is consistent: what matters is not what tokens appear in the priming, but what *processing mode* the priming activates. Domain priming activates domain processing. Meta-commentary activates meta-processing. The expert generated continuations in domain processing mode; the receiver needs to match that mode, not a meta-reflective version of it.

4.2 The more explicit, the worse

The inverse relationship between explicitness and performance (storyteller > provocateur > director among bare conditions) has a straightforward explanation: more explicit instruction activates the instruction-following mode more strongly. The director generates 3.3× more tokens of explicit instruction, each of which pulls the model further into “I am following directions” processing rather than “I am reasoning about medicine.”

This contradicts the common practice of providing detailed coordination instructions to AI agents. More detailed briefings are not better briefings. The optimal briefing is the minimal domain-specific signal: enough to activate the right processing mode, with nothing that activates competing modes.

4.3 The reset does (almost) everything

The most striking result: reset+neutral (CE = 1.278) nearly matches fixed domain priming (CE = 1.197), closing 89% of the gap. The reset instruction alone — with no domain-specific content — nearly matches the expert’s own priming.

Combined with Paper XX’s finding that reset+priming beats the expert baseline, this suggests the reset instruction’s role is not merely additive. It is *multiplicative*: it removes residual processing biases that attenuate the effect of whatever comes next. Whether what comes next is domain priming, neutral priming, or shepherd meta-commentary, the reset ensures it lands in a clean context.

The shepherd strategies fail not because they are bad priming, but because the reset has already done the work. After reset, the marginal value of additional coordination content is near zero. The only exception is domain-matched priming, which provides the specific activation of the target processing mode.

4.4 Implications for coordination protocol design

Paper XX established reset-then-prime as the protocol. Paper XXII eliminates a class of proposed extensions:

1. **Don't add adaptive priming layers.** A shepherd agent between reset and domain priming degrades performance by 0.12 nats. The protocol should be two steps (reset, prime), not three (reset, shepherd, prime).

2. **Don't make priming more explicit.** More explicit instruction activates instruction-following mode, which competes with domain processing. The optimal priming is domain-specific content, not instructions about how to process domain content.

3. **Don't extend priming to be probe-specific.** Fixed domain-level priming outperforms probe-specific adaptive priming. The overhead of specificity (additional tokens, mode-switching) exceeds its value.

The effective coordination protocol is:

Reset (clear residual biases) → **Prime** (15–50 tokens of domain content) → **Deliver** (the task)

Three steps. Each minimal. Nothing between them.

5 Conclusion

Shepherd agents — adaptive, probe-specific priming strategies — do not improve coordination. All three strategies (storyteller, provocateur, director) perform worse than no coordination when used alone, and add nothing beyond what a reset instruction provides when used with reset. More explicit instruction produces worse results, not better. Adding a shepherd layer between reset and domain priming actively degrades the reset-prime protocol.

The coordination mechanism is the reset-prime sequence itself. No sophistication of the priming content improves upon minimal domain activation preceded by context reset. The field's intuition that better coordination requires more detailed, more specific, more adaptive briefing is exactly wrong. Better coordination requires *less*: clear the context, state the domain, deliver the task.

Data Availability

All results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper22/`.

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