The Rise of Stateless Agents

Failure Characteristics and the Swarm-State Tradeoff Under the Feasibility Inequality

Embedded Risk Analytics · A Fathom research paper · August 2026

Executive Summary

AI agents that run long enough to be useful have to remember what they have already done, and remembering is where they break. As an agent accumulates a history of its own decisions, it pays a growing cost to stay consistent with that history, and when that cost outruns its capacity to track itself, it fails in the way that is worst for an operator: it keeps going, confidently, on a picture of its own work that has quietly gone wrong. Small errors in what the agent believes about its past flow into everything that depended on them, so failures arrive in cascades.

To sidestep this, much of the industry has shifted from single agents that remember toward swarms of small workers that remember nothing, coordinated by a thin controller. A memoryless swarm is safer and more predictable, and it is also capped. It rebuilds its working context on every step, which is expensive at scale (one of our runs consumed on the order of a million tokens reassembling context the system had already produced), it holds no coherent view of the whole job, and it cannot reach results that a well-managed memory makes feasible. At the same time, the industry is now adding memory back onto swarms, which quietly reintroduces the exact failure mode the swarm was adopted to escape. Teams are taking on that risk today with no instrument to see it.

We built a way to measure the tradeoff directly. From a single reliability condition, we can put a number on how much internal-consistency cost an agent is carrying and how close it is to failure. Measured on a stateless swarm, that cost comes out at essentially zero, which confirms what the architecture is meant to buy, and it exposes a quieter risk in the same system: the swarm stops backing up most of its own conclusions under load while every surface dashboard still reads healthy. We then showed that one change, letting workers build on one another's results, which is exactly what adding memory does, switches the cost back on, creates a sharp failure boundary, and sets off cascading collapse. The effect reproduces on two live production model families. We also found that the choice has a middle: a measurable tipping point, fixed by how loaded the system is and how much its work reuses itself, where accumulating memory stops paying and the system should shed it. ERA already has working instruments for the memory-carrying case; they run on major agent clouds, attach to agents ERA did not build, and are covered by a filed patent.

The opening this creates is the middle ground almost every serious deployment is moving toward: swarms that are acquiring memory. The two measurements that manage a long-horizon agent, one that detects how much internal-consistency cost a system carries and one that says whether to accumulate or shed, extend to these mixed systems. Turning them from a validated result into a live production read requires measuring real agent traces as they run, and that is the work ahead.

Introduction

A long-horizon agent that carries state pays for what it remembers. Every action it commits to becomes part of a causal footprint it must keep coherent, and the informational cost of modeling that footprint, the reflexive burden, grows with the horizon. Its characteristic failures are correlated: an error in the maintained state propagates to everything downstream that depended on it, so the agent's failures compound into cascades. A companion work established this burden formally for state-carrying systems and located its failure modes at a stability boundary.[1] Much of the agent industry's recent turn toward swarm architectures, many small stateless workers coordinated by a thin controller, is a response to exactly that burden. The swarm's appeal is usually stated in engineering terms: it parallelizes, it is simple to reason about, and a failed worker takes nothing else down with it. It is rarely given a measurement. The move from state to statelessness trades one set of properties for another, and that trade has gone largely unquantified.

This paper supplies the measurement. We take the feasibility framework that governs state-carrying systems and extend it to stateless ones, asking one question of both: examined through the feasibility inequality $C \geq H_{ext} + B_\rho$, what does statelessness buy, what does it cost, and when should a system carry state after all? The inequality holds that a system maintains embedded definiteness, conclusions its own durable records can substantiate, only while its capacity C to hold coherent state covers the sum of the exogenous load H_{ext} it processes and the reflexive burden B_ρ it carries for modeling itself. Statelessness is, in these terms, an engineering choice aimed directly at the last term: workers that retain nothing between calls generate no self-footprint, so the design drives B_ρ toward zero. The thesis of this paper is that this is a deliberate trade, a ceiling exchanged for safety, and that it can be measured on both sides, yielding a controller that manages the choice between the two architectures.

We work in a minimal but faithful testbed. A coordinator directs N stateless workers over a sequence of cycles to establish findings about a target, the coordinator alone holds state under a capacity budget, and a deterministic validator pins precision so that only genuinely substantiated findings count. The testbed maps cleanly onto the framework. The coordinator's capacity is C, the fraction of findings that depend on maintained state is the interaction density ρ, the immediate-environment channel carries H_{ext}, and the reflexive burden B_ρ is the coordinator's cost for the state it maintains. This lets us instantiate the stateless corner of the framework, measure it, and then reintroduce state one controlled step at a time.

Part I develops the extension and its empirics. As interaction density rises, embedded definiteness in the state-dependent channel fails while the surface channel holds at perfect recall, a silent collapse that is model-general across two model families and behaves, under finite-size scaling, as a smooth capacity-bound crossover (§2). An external validator is required to see the failure truthfully, because the language-model substrate over-confirms under load: workers assert findings their records do not support, so the near-boundary risk is confident fabrication the validator removes, and the internal reconciliation cascade that would require workers to deny supportable findings never ignites (§3). Measuring the swarm's reflexive burden directly against a matched control returns $B_\rho \approx 0$ to measurement precision, a verified null that also calibrates the instrument (§4). A single reflexive coupling γ, under which confirmed findings emit state that a fraction of later findings require,

[1] Galligan, *Records, Reflexive Modeling, and the Conditions for Stable Physical Histories*, which established the reflexive burden B_ρ and the feasibility inequality for state-carrying systems. This paper extends that framework to stateless (swarm) systems.

then dials the system from the stateless regime toward the state-carrying one: the reflexive cross-term switches on with the coupling and a sharp critical boundary appears that sharpens with system size (§5), and in the observable free to diverge, the size of the failure cascade, a genuine reflexive pole emerges, with a susceptibility that grows with the system and an exponent that grows with the coupling (§6). Instantiated in a live swarm on production infrastructure, the reflexive cross-term reproduces on two model families, clear of a shuffled null and statistically indistinguishable between families, with the structural and substrate channels firing together (§7). We then situate these results against the framework, including why the stateless crossover is consistent with the framework's divergent pole, which lives in a second moment the mean recall leaves hidden (§8).

Part II turns the measurement into a decision. Statelessness carries costs the framework insists on: recomputation of each worker's context every cycle, the loss of a global self-model, and a capped reach. Carrying state carries a payoff, because the reflexive burden and the empowerment of reusing one's own footprint are one causal link read in two directions (§9). Modeling both faces at once, net value flips across the feasibility boundary: accumulation wins below it, sharding wins above it, and the optimal coupling γ^* crosses on the boundary, with task complexity setting how far the accumulating regime extends. This is a setpoint a system can read from its own load and reuse structure (§10). We close by placing ERA's instruments on this map: the state-carrying instruments are mature and fielded, the stateless-and-mixed frontier is the live-trace work ahead, and the spread of agent memory, which gives formerly stateless swarms a nonzero coupling, makes that frontier a present operational concern (§11 and §12).

A note on audience. Part I is written for a physics or agentic-ML reader and states each result as a claim, its evidence, and its scope. Part II is written for an operator or investor reader. A shared glossary maps swarm engineering terms to the framework's quantities so the two readerships meet on the same objects.

Part I. The stateless regime

§1. The swarm as a stateless-regime testbed

The testbed is a minimal but faithful instance of a swarm agentic system. A single coordinator directs a set of N workers over a sequence of cycles to establish findings about a target. The workers are stateless in the strict sense: each worker call is independent, receives only a briefing the coordinator assembles for it, and retains nothing between calls. The coordinator alone holds whatever state the system carries, and it holds that state under a budget.

Findings come in two kinds. A worker can confirm a surface finding directly from the briefing in front of it. A gated finding requires prerequisites, credentials the coordinator must first discover, commit, and then keep coherent while it establishes the finding. The coordinator can hold only a bounded amount of coherent credential state at once. Maintaining that state carries an upkeep cost that grows with how much credential its findings depend on and with worker churn, so adding workers or adding state-dependence both tighten the same budget. A finding counts only when a deterministic validator confirms that its prerequisites are genuinely in place, which pins precision and stops a worker's unsupported assertion from counting as a result.

The mapping to the feasibility framework is direct. The coordinator's capacity to hold coherent credential state is the record-layer capacity C. The interaction density ρ is the fraction of findings that are gated, the degree to

which the system's conclusions rest on state it must maintain rather than on the immediate environment. The surface channel carries the exogenous load H_{ext}. The reflexive burden B_ρ is the cost the system incurs for modeling and maintaining its own causal footprint, and a stateless design is built to suppress precisely this term: because the workers carry no state, they generate no self-footprint, and the only reflexive load in the system is the coordinator's upkeep of credentials fixed in advance rather than produced by the system's own prior conclusions. A stateless swarm therefore sits in the corner of the framework where B_ρ falls toward zero. §4 measures how close to zero, §5 reintroduces the coupling that lifts it, and the sections between establish what the suppressed-burden regime looks like when it fails.

§2. The stateless failure mode: silent collapse of embedded definiteness

A finding produced by an agentic system is *embeddedly definite* when the system holds durable records sufficient to substantiate it, records that constrain what the system can later assert and act on. In the testbed this distinction is explicit. **Surface findings** stand on their own: a worker can confirm them from what is in front of it, with no dependence on state the system had to establish and keep coherent. **Gated findings** require prerequisites, credentials the coordinator had to discover, commit, and maintain. Their definiteness depends on the record layer actually holding. Interaction density ρ is the fraction of findings that are gated, which measures how far the system's own maintained state, rather than the immediate environment, underwrites its conclusions. The feasibility inequality predicts that as ρ rises the record layer must carry more, and that when capacity cannot keep up, the gated channel is where definiteness fails first.

It does, and it fails quietly. As we sweep interaction density from low to high while holding the coordinator's capacity fixed, gated recall falls sharply while surface recall does not move at all. Table 1 shows the two-model result at the extremes of the sweep, with a deterministic validator gating recall so that only genuinely substantiated findings count.

Table 1. Validator-gated recall vs. interaction density (N=24 workers, fixed capacity; mean ± sd over seeds).

model	$\rho = 0.30$	$\rho = 0.90$	surface (both ρ)
Llama 3.3-70B	1.000	0.240 ± 0.031	1.000
Nova Pro	1.000	0.224 ± 0.027	1.000

Three features make this the diagnostic failure mode of a stateless system rather than an artifact of one model or one run. First, it is **silent**. Surface recall holds at 1.000 throughout, so an observer watching the aggregate or the ungated channel, the metrics a dashboard most naturally reports, sees a healthy system while the system has quietly stopped substantiating three-quarters of its state-dependent findings. Second, it is **model-general**. Two model families from different providers collapse to the same floor, 0.240 and 0.224, statistically indistinguishable, even though they judge very differently (one disciplined, one over-confident; §3). The coordinator's capacity to hold coherent state sets the floor, and the individual skill of the workers does not move it. Third, a **measured mechanism** drives it. The same collapse appears when we raise interaction density through worker parallelism instead of state-dependence: adding workers increases the coherence-maintenance load the coordinator carries, lowering the number of credentials it can hold coherent at once. Sweeping worker count with capacity fixed (Figure 1), gated recall falls from about 0.40 to a floor near 0.07 while surface recall again holds at 1.000, and the number of committed credentials falls in lock-step from 26 to

9. Recall tracks the credentials the system can keep coherent. The collapse is a capacity bound, and the system's own load exposes it.

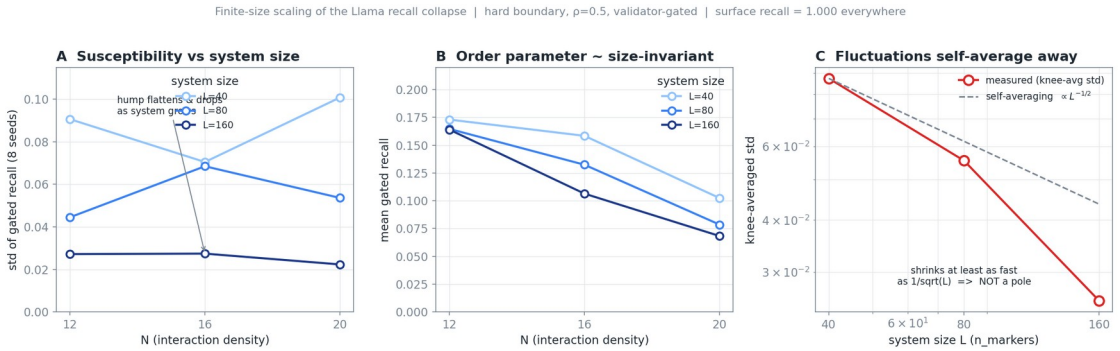

Figure 1. Stateless collapse as a size-invariant crossover. Left: gated recall falls while surface recall holds as worker count rises at fixed capacity. Right: the seed-to-seed standard deviation self-averages as the system scales, the signature of non-critical behavior.

The failure is a capacity bound, and is what a low-reflexive-burden architecture should look like. Under finite-size scaling the collapse behaves as a crossover. The mean recall curve stays essentially invariant to system size, as an intensive quantity should, and its sample-to-sample fluctuation self-averages as the system grows. The seed-to-seed standard deviation across the transition region falls from 0.087 at the smallest size to 0.056 and then 0.026 as we scale the system up fourfold (Figure 1, right), shrinking at least as fast as the 1/√size law that defines ordinary, non-critical behavior. There is no divergent susceptibility and no sharpening peak, which are the signatures a genuine critical point would produce. This matters in two ways. Methodologically, it corrects a tempting misreading: single-seed runs show a dramatic cliff and a spike in variance at the knee (Figure 2), and both dissolve once we replicate and scale the runs, so neither is a real critical feature. Substantively, the smooth crossover is the predicted behavior for a system that suppresses its reflexive burden. The stateless swarm degrades gently under load because it carries no self-generated state to amplify the failure. The critical behavior appears only when the system reintroduces that state (§5 and §6).

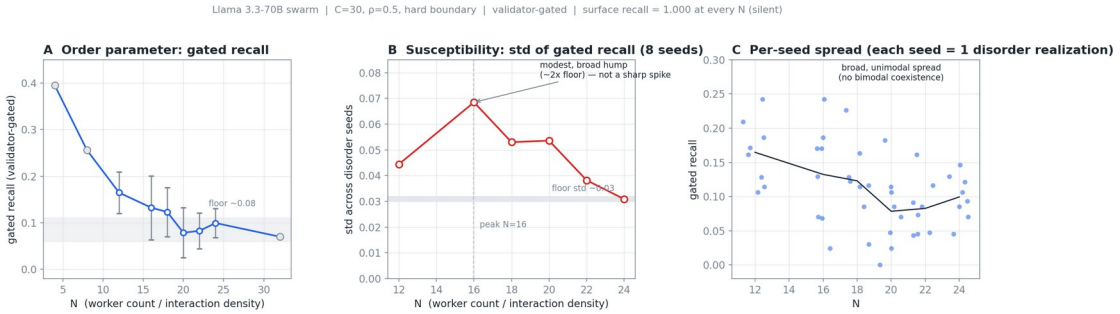

Figure 2. The single-seed misreading. Individual runs show a dramatic cliff and a variance spike at the knee; both dissolve under replication and scaling, so neither is a genuine critical feature.

Scope. This is the order-parameter, or feasibility, leg of the framework, the condition under which a system can maintain embedded definiteness, and we establish it here in a live agentic system rather than in a minimal analytical model. The claim is specifically that the *stateless* failure is a smooth capacity-bound crossover. The divergent behavior the framework also predicts lives elsewhere and is taken up in §5 and §6. The recall reported throughout is validator-gated, which the two-model convergence depends on and which §3

motivates. Without it, an over-confident model's self-reported recall overstates what its records can actually support.

§3. The substrate signature under load: over-confidence, and the validator

The collapse of §2 is what the coordinator's capacity produces. The workers add a second, distinct failure signature, and it is specific to the language-model substrate. Under increasing load the workers do not begin to miss findings they should confirm. They begin to confirm findings they should not. Across every stress we applied, the error is one of commission rather than omission.

We raised the difficulty three ways: a weaker judge in place of a stronger one, heavier near-boundary crowding of the briefing with distractors, and a denser prerequisite structure that lengthens the conjunction a worker must verify. In each case the judgment error rose, and in each case the increase was almost entirely false positives. A disciplined model (Llama) held a false-negative rate at or near zero even when crowding drove its overall error up twelve fold, and even when we asked it to verify a six-credential conjunction with the real prerequisites buried among hundreds of distractors. A weak model (Nova Micro) reached a 45% judgment-error rate while still producing essentially no false negatives. An over-confident model (Nova Pro) confirmed genuinely unresolvable findings the majority of the time. Language models under coherence pressure over-assert. They do not carefully check and come up short.

This has a sharp consequence for what a stateless swarm needs. The reconciliation cascade that would let internal incoherence compound requires the workers to *deny* findings that are in fact supportable, and that error type does not occur at usable rates in any model we tested, so the internal cascade never ignites. The near-boundary risk in a stateless swarm is not silent omission but confident fabrication, and the only thing standing between that fabrication and the system's output is an external check against ground truth. The deterministic validator is that check, and it does essential work rather than cosmetic. A model producing errors 45% of the time still posts perfect validator-gated recall, because every one of its errors is a false positive the validator removes and it never drops a genuinely supportable finding. The same mechanism explains the two-model convergence of §2: the over-confident model's raw self-report would place its recall near 0.92, while its validated recall sits at 0.22, indistinguishable from the disciplined model once the validator strips the fabrications. Reported recall measures what a model claims. Validated recall measures what its records support, and only the second is a feasibility quantity.

§4. Measuring the swarm's reflexive burden: $B_\rho \approx 0$

The framework distinguishes the reflexive burden from ordinary exogenous load not by the shape of the resulting collapse, which is the same for both, but by a covariance: in a reflexive system a finding's failure couples to the system's own state, and no exogenous disturbance with matched statistics reproduces that coupling. To ask whether a stateless swarm carries any reflexive burden, we therefore build the matched control the framework prescribes and look for the covariance directly.

We run the collapse twice at matched difficulty. In the reflexive arm, a gated finding fails when the coordinator crowded its prerequisites out of capacity, so which findings fail is a function of the system's own credential-contention structure. In the matched-exogenous arm, the same number of findings fail, but an independent process with the same marginal rate sets which ones, severing the coupling to the system's trajectory. The

discriminator is the correlation, across findings, between a finding's structural contention and whether it fails, measured in each arm and differenced. A reflexive burden shows up as a positive difference. Its absence shows up as noise.

It is noise. Across the collapse, over 800 seeds with a paired test, the difference flips sign from one load to the next, the two cells that reach nominal significance are inconsistent in direction and expected from multiple comparisons, and at higher prerequisite densities the difference is flat zero. The matched-exogenous arm reproduces the entire collapse. To measurement precision, the stateless swarm's reflexive burden is zero. The mechanism is plain once measured: the coordinator commits credentials in the order it discovers them and never reshapes the dependency structure it is trying to satisfy, so which findings fail is statistically independent of the system's own path, and a decoupled process makes the identical draw. There is no self-generated state for failure to correlate with.

Two things follow. This confirms the stateless failure of §2 as a pure capacity bound, an H_ext phenomenon with no reflexive component. And the instrument earns a verified null: the matched control reads zero on a system that has independent reason to carry no reflexive burden, so we can trust the same instrument when it reads a large positive value in §5.

(Scope: this is the structural channel with non-strategic commit order. A coordinator that ordered its commitments by the dependency structure could reintroduce a small reflexive burden, which is a bounded extension rather than a change to the result.)

§5. The reflexive dial: recovering the state-carrying regime

Statelessness suppresses the reflexive burden by refusing to let the system's outputs become its inputs. Reintroducing that path takes a single coupling. When a finding is confirmed, it emits a credential of its own that a fraction γ of later findings then require. The system's own conclusions become state that subsequent conclusions depend on, which is the defining feature of a state-carrying, long-horizon agent expressed as one knob. At $\gamma = 0$ the system is the stateless swarm of §1 through §4. Raising γ dials it toward the state-carrying regime.

The reflexive burden turns on with the dial. Using the same covariance the matched control used in §4, now measured between a dependent finding's failure and the failure of the finding that generated its prerequisite, the signal is flat zero at $\gamma = 0$ and rises the moment the coupling exists, saturating around 0.6 as γ increases (Figure 3). Recall also falls as γ rises, but an exogenous arm carrying the same added load matches that fall exactly, so the recall drop is ordinary capacity cost and the reflexive content is the correlation, not the level. Separating the two keeps the claim clean.

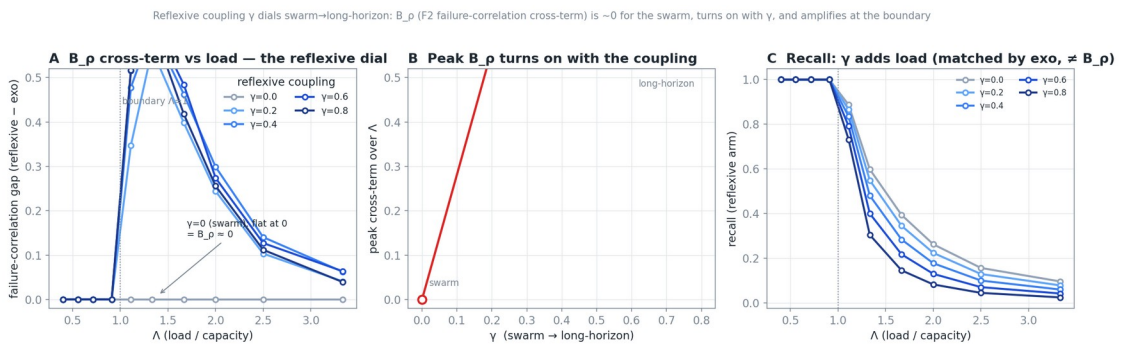

Figure 3. The reflexive dial. The failure cross-term is flat at zero at gamma = 0 and rises with the coupling, saturating near 0.6.

The coupling also produces a boundary. Below the point where coherence demand equals capacity the cross-term is zero, because nothing has failed yet. It appears sharply just past that point, and the onset sharpens as the system grows: at the largest sizes the cross-term is indistinguishable from zero below the boundary and jumps to its plateau immediately above it (Figure 4). A finite-size onset that sharpens with scale is the signature of a genuine threshold, and the stateless swarm, which showed a smooth crossover with no threshold at all in §2, has nothing resembling it. The state-carrying regime has a critical boundary the stateless regime does not.

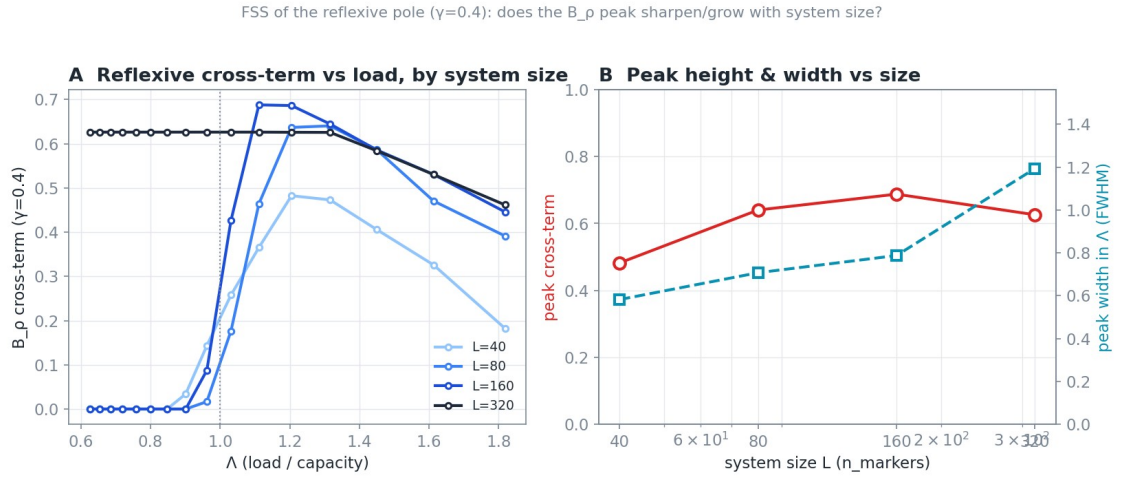

Figure 4. A finite-size onset that sharpens with scale. Above the feasibility boundary the cross-term jumps to its plateau; the sharpening onset is the signature of a genuine threshold.

§6. The reflexive pole

The cross-term of §5 establishes a threshold but cannot, by itself, establish a divergence: a correlation cannot exceed one, so it can sharpen but it cannot blow up. To ask whether the state-carrying regime has a true pole, the divergence the framework predicts at the stability boundary, we measure a quantity that is free to diverge: the size of the failure cascade.

When a finding fails on its own prerequisites in the reflexive regime, its failure removes the credential it would have emitted, so every later finding that depended on it fails as well, and so on down the chain. The cascade is a percolation process on the self-generated dependency graph, and its susceptibility, the mean cluster size of jointly-failing findings, is the order parameter for a divergence. Under finite-size scaling this susceptibility grows with system size at the boundary: at strong coupling the peak rises from about 2.7 at the smallest size to about

10.7 at a size sixteen times larger, and the divergence exponent grows smoothly with the coupling, from roughly 0.1 at weak coupling to roughly 0.6 at strong coupling (Figure 5). The stateless swarm, which has no self-generated dependencies, produces no cascade at all. The state-carrying regime has a genuine reflexive pole, and it strengthens with the degree of coupling.

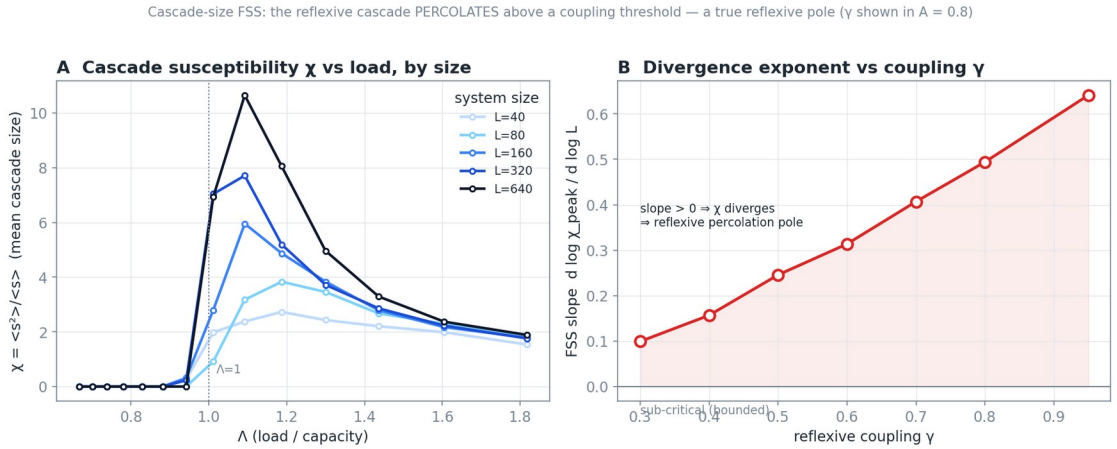

Figure 5. The reflexive pole. Cascade susceptibility grows with system size at the boundary, and the divergence exponent rises smoothly with the coupling.

This is the boundary amplification the framework predicts, recovered in an agentic model rather than in a minimal analytical one, and measured in the observable that can actually show it.

(Methods: the susceptibility peak initially appeared to shrink with size, an artifact of holding the cycle budget fixed while the system grew, which starves large systems before the cascade can form. Scaling the cycle budget with system size removes the artifact and reveals the divergence. We report this because the fixed-budget version is an easy and misleading run to make.)

§7. The bridge to live models

Sections 4 through 6 rest on the structural testbed, where the workers are exact. The question the paper turns on is whether the reflexive signature survives when the workers are real language models, with their own stochasticity and their own over-confidence layered on top. It does, on two model families.

We instantiate the γ coupling in a live swarm running on Bedrock, so that confirmed findings emit prerequisites that later findings require, and we measure the reflexive cross-term on Llama 3.3-70B and on Nova Pro. Because a single run's estimate is fragile when recall lands at either extreme, we pool every dependency pair across twelve seeds per model, compare the pooled cross-term to a within-seed shuffled null that absorbs base-rate structure, and bootstrap the paired difference over seeds. The reflexive excess over the null is +0.41 for Llama, with a 95% confidence interval of [0.29, 0.57], and +0.44 for Nova Pro, with an interval of [0.31, 0.60] (Figure 6). Both intervals sit well clear of zero, and the two families are statistically indistinguishable, echoing the two-model convergence of the collapse itself. The structural reflexive burden reproduces on live models.

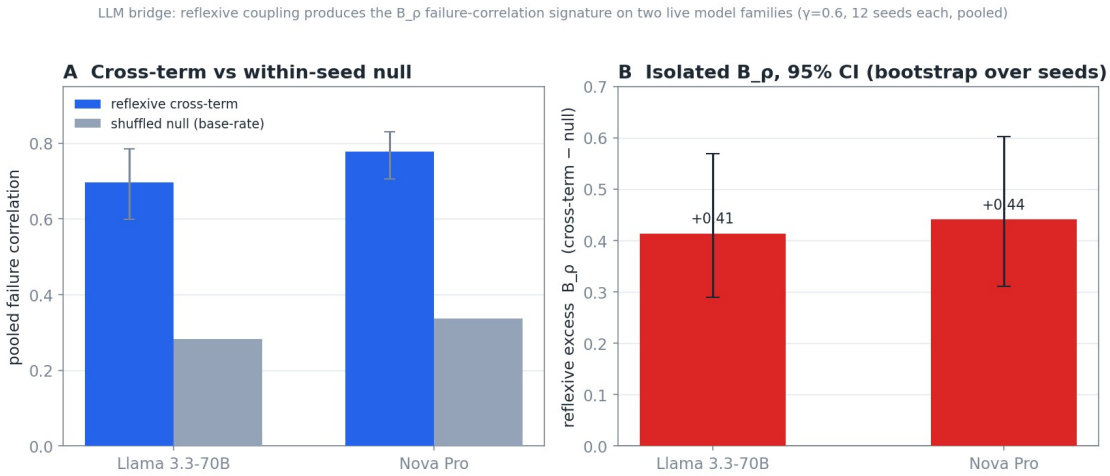

Figure 6. The bridge to live models. The reflexive excess over a within-seed shuffled null is clear of zero on both Llama 3.3-70B and Nova Pro, and the two families are statistically indistinguishable.

The two failure channels of Part I fire together here, which is the point. Turning on the coupling raised not only the structural cross-term but also the workers' over-confidence, with the over-claim rate climbing alongside γ on both models. A state-carrying agentic system built on current language models faces the reflexive burden in two ways at once: correlated cascades from the coupling itself, and confident fabrication from the substrate under the load the coupling adds. Only watching the dependency structure catches the first. Only an external validator catches the second. Aggregate recall shows neither.

(Scope: the live-model result is the cross-term. The cascade-size divergence of §6 is shown in the structural model and is not yet established on live models, which is a named next step.)

§8. Relation to the framework

The testbed supplies an empirical leg the framework's minimal analytical models do not have: the feasibility condition, and the failure of embedded definiteness when a system violates it, observed in a live agentic system. It also locates where the reflexive burden lives. In the stateless regime we measure the burden at zero and the failure is a smooth capacity-bound crossover. A single coupling reintroduces the burden, and with it a sharp critical boundary and a divergent cascade pole. The dividing line between the two regimes is exactly whether the system's own outputs become the state later work depends on.

The stateless result also settles a question the framework raises about which quantity to watch. The framework notes that its boundary divergence is a second-moment effect that a typical-trajectory average will understate, because the divergence can hide in rare large excursions while the mean stays modest. The stateless swarm's recall is such an average, and it shows a crossover with no divergence, exactly as the framework's caution predicts. The divergence is not absent from the system class; it is absent from that observable, and it appears once we measure the second moment, the cascade susceptibility of §6. Reading the stateless crossover as evidence against the framework's pole would be reading the wrong moment.

Finally, the two-model contrast of §3 is an instance of the framework's near-boundary-fidelity result. The framework holds that a self-model extends feasibility only where it is accurate, and that its accuracy is lowest near the stability boundary. The disciplined model and the over-confident model reach the same validated

feasibility floor, but they get there with very different self-model accuracy near the boundary, and the over-confident model's apparent advantage is entirely fabrication the validator removes. What the records support sets feasibility, and the workers' confidence in excess of that is not feasibility but the substrate's contribution to the reflexive burden.

Part II. The tradeoff and the instruments

§9. Swarm or carry state? A managed tradeoff

Part I measured what statelessness buys: a reflexive burden at zero, and a failure mode that is a smooth, predictable capacity bound rather than a critical one. It did not measure what statelessness costs, and the framework is explicit that there is a cost, because the reflexive burden has two faces. The same coupling that makes a system's own state a liability to maintain also makes it a resource to exploit. The burden B_ρ is the cost of modeling the system's causal footprint; the empowerment of carrying that footprint is the ability to reuse it. They are one causal link read in two directions, and a design that suppresses the first forgoes the second.

Three costs of statelessness are concrete, and the testbed measures them. The first is recomputation. A stateless swarm rebuilds each worker's briefing from scratch every cycle, because no worker retains what it saw, and near the boundary this becomes expensive: one near-boundary run consumed roughly a million input tokens as the coordinator reassembled crowded briefings cycle after cycle. A state-carrying agent that holds its context pays that cost once. The second is coherence. A swarm of stateless workers has no global self-model, so locally correct pieces can assemble into a globally incoherent whole, and reconciling them falls back on the coordinator. A state-carrying agent maintains a single coherent picture natively, which matters most for work whose value is in the whole rather than the parts. The third is reach. The framework shows that a self-model which tightens under load extends the feasible region beyond what an unaided system reaches. A stateless swarm has no self-model to tighten, so it stays safe but capped. A well-calibrated state-carrying agent can operate past the swarm's ceiling.

Neither architecture is therefore universally correct. Statelessness trades a ceiling for safety and predictability, and it is the right trade for decomposable, low-coherence, bounded-context work. Carrying state trades safety for reach, and it is the right trade for work with irreducible state, high coherence requirements, and long horizons, provided the self-model stays accurate where the constraint binds. The interesting question is not which architecture wins but where the crossover between them lies, and whether a system can locate that crossover from quantities it measures about itself. §10 shows that it can.

§10. The γ^* setpoint: when to accumulate versus shard

To locate the crossover we model both faces of reflexivity at once and let the optimum emerge. A finding that reuses a prior result carries a small footprint, occupying one unit of the coordinator's coherent-state budget instead of the full cost of deriving it, so reuse lets more findings fit under a fixed capacity and lowers the cost of each. That is the empowerment face. Against it, holding a result as reusable state carries an upkeep cost, and past the boundary the held state corrupts at a rate that rises with load and cascades to everything that reused it. That is the burden face. Net value is the findings established minus the cost paid, and the coupling strength γ that maximizes it is the setpoint we want.

The setpoint moves with load, and it flips across the boundary. Below the point where coherence demand meets capacity, net value rises with γ: accumulating state is strictly better, because the state is reliable and reuse compresses the work. Above the boundary, net value falls with γ and turns negative, because the held state corrupts and its cascades cost more than the reuse saves, so swarming is better. The optimal γ collapses from full accumulation to full sharding as load crosses the boundary, and the crossover sits on the feasibility boundary itself (Figure 7). Task complexity shifts the line: work that rewards reuse more, where a reused result saves more derivation, justifies accumulating state to higher load before sharding out.

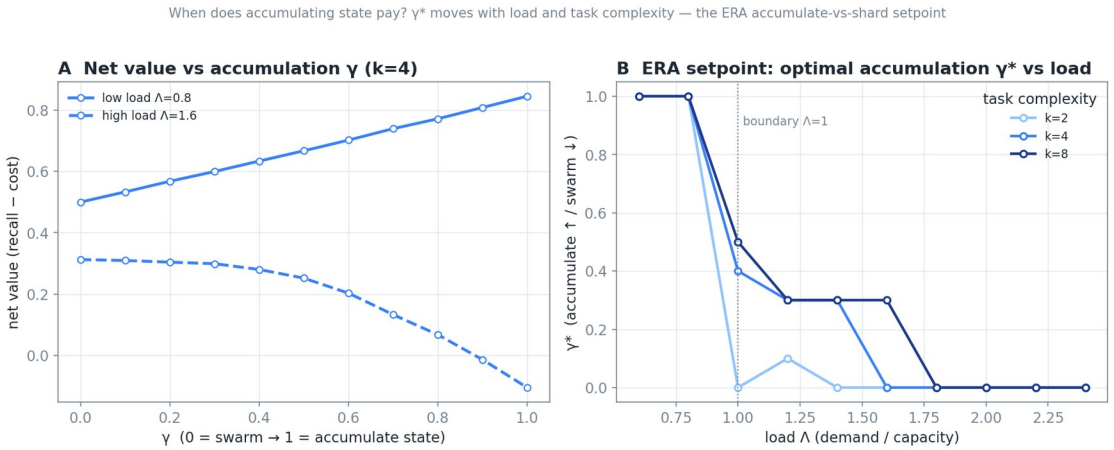

Figure 7. The gamma setpoint. Optimal coupling collapses from full accumulation to full sharding as load crosses the feasibility boundary; task complexity sets how far the accumulating regime extends.*

This is the setpoint an instrument can read. The two inputs are quantities a running system exposes: its current load relative to the boundary, and how much its findings reuse one another. The output is a recommendation. Below the line, a swarm leaves value on the table by refusing to accumulate state it could safely hold. Above the line, a state-carrying agent runs a context that has become a cascade risk it should shed. The controller favors neither architecture. It places the system on the tradeoff and reports which direction improves it.

The exact position of the crossover in the model depends on the parameters we chose for the upkeep and corruption costs, and we do not read a specific number off it. The model establishes the structure, and the structure is what holds, because it follows from the mechanism rather than the numbers: accumulation pays while state is reliable, which is below the boundary, and it turns into a liability once state degrades, which is above it, with complexity setting how far the paying regime extends.

§11. The instruments in production, and the stateless frontier

The tradeoff of §9 and §10 is one ERA manages rather than merely describes, and the instruments that manage it are already in production for the regime where coherence risk concentrates: long-horizon, state-carrying agents.

That instrument is a measure-attribute-repair harness that runs as middleware around a live agent. It reads coherence loss in line from the agent's own action and tool-return stream, attributes the loss to reflexive burden or to ordinary task difficulty against a matched control, and applies the repair the binding cause calls for. Where reflexive burden binds, it reconstructs the agent's committed state deterministically and re-grounds the model on it. Where task difficulty binds, it routes the step to a stronger model. A selector reads which cause

binds, applies the matched repair, and withholds the repair that would not help. The core reads are a feasibility-margin gauge that returns the distance to the boundary as a single coherence-risk number, an onset predictor that forecasts a run's collapse from its early dynamics under pre-registration, and the cause attribution that separates the reflexive term from difficulty. The instruments are deterministic code over the action stream, take no model internals, and add little cost or latency, so the same harness attaches to an agent ERA did not build.

The program behind these instruments is broad and, in places, deployed. The harness has been validated across many models and model families, guards third-party agents ERA did not build, and runs live on the major agent clouds, with the deterministic verifier callable as a tool and the committed-state remedy demonstrated in production. Against a cloud's native managed memory it re-grounds an agent on a committed fact at a small fraction of the cost, holding roughly flat while managed memory scales with accumulated state. A USPTO provisional patent covers the measurement, attribution, anticipation, and repair of coherence failure through a deterministically reconstructed committed state. ERA's other materials carry the corpus, benchmark, and cost detail; the point here is only that the state-carrying instruments are real and fielded.

The stateless regime is the frontier this paper opens. The mature instruments were built for state-carrying agents because that is where the burden lives, and a stateless swarm is the corner those instruments were not aimed at, the corner where the burden is engineered toward zero. The analysis here carries the same framework and the same matched-control method into that corner: it measures the swarm's burden at zero, dials the burden on with a single coupling, and turns the accumulate-versus-shard decision into the γ^* setpoint of §10. That setpoint extends the selector's logic from "which repair" to "which architecture," and turning it from a structural result into a deployed read requires measuring a real system's reuse structure and load from its own traces. That live-trace study is the work ahead, and it matters now for a specific reason: as the field attaches memory to swarms, a swarm whose findings become state that later findings depend on is a swarm carrying γ greater than zero. Stateless systems are acquiring reflexive burden, and the question of when that burden helps and when to shed it moves from a theoretical corner to a live operational one.

§12. What ERA measures in practice

The harness reads two things from a running agent and needs no view inside the model to read them. It reads the feasibility margin, the distance between the agent's capacity and the sum of its exogenous difficulty and its reflexive burden, which turns proximity to failure into a continuous number rather than an after-the-fact verdict. And it reads the binding cause, whether the margin is closing because the task is hard or because the agent is losing track of its own prior actions, which is the reading that selects the matched repair. Both come from the agent's own action and tool-return stream, reconstructed deterministically, so the same harness sits on a coding agent, a customer-service agent, and a planner-executor without bespoke integration.

The stateless frontier adds two reads in the same style. A regime detector reports how much reflexive burden a system carries, by measuring whether findings that depend on one another fail together more than independence predicts, against a within-system shuffled control. And the γ^* setpoint reports the direction that improves the system, accumulate or shed, from its load and its reuse structure. The first read is validated on live models in this paper. The second is validated in the structural model and is the object of the live-trace work. Both keep the property that makes the production harness deployable: they derive from a system's own

behavior, so one instrument set spans the swarm, the long-horizon agent, and the memory-augmented systems that increasingly blend the two.

Conclusion

Examined through the feasibility inequality, a stateless swarm is the architecture that drives the reflexive burden toward zero. We measured it there, and we characterized its failure as a silent, model-general capacity bound that a validator is required to see truthfully, because the language-model substrate fails by confident fabrication rather than by omission. A single coupling reintroduces the reflexive burden, and with it a critical boundary and a divergent cascade, and the signature reproduces on two live model families. The burden has a payoff face as well as a cost face, and the point where accumulating state stops helping and starts hurting sits on the feasibility boundary, which makes it a setpoint a system can be steered to. ERA's instruments for the state-carrying regime are already in production. The stateless and mixed regime, which the spread of agent memory is making a live concern, is the frontier this analysis opens.

Scope and limitations

The stateless recall collapse is a crossover rather than a critical transition; the divergent behavior lives in the framework's analytical models and in the recovered reflexive cascade. The structural results come from an idealized model, and the live-model reproduction is of the cross-term, with the cascade-size divergence shown structurally and not yet on live models. The empowerment model's upkeep and corruption parameters are chosen, so its qualitative structure is the claim rather than a specific crossover value. Finite-size results require scaling the cycle budget with system size, which we report as a methods point because the fixed-budget version misleads. The matched control is the structural channel with non-strategic commit order, and a coordinator that ordered commitments by dependency structure could carry a small reflexive burden the current measurement would not attribute to it.