The New Model Was Not the Missing System

Exzil Calanza

Independent Researcher, Iloilo City, Philippines

Independent research preprint. IEEE-style formatting only; this work has not been accepted by, published by, certified by, or peer reviewed by IEEE.

ABSTRACT This technical synthesis examines evidence that stronger foundation models do not by themselves remove system-level defects in task contracts, queue management, context continuity, or verification. Drawing on the source article's field observations and cited work on multi-agent failure modes, capability horizons, structured memory, and risk management, it argues for an engineering focus on persistent state, explicit handoffs, and verification gates. The analysis is qualitative and secondary-source based; it does not present a new benchmark dataset or controlled model-comparison experiment.

INDEX TERMS ¢â‚¬â€ agentic reliability; system harness; MAST framework; structured memory; verification contracts; capability horizons

I. INTRODUCTION

STUDY TYPE: Technical synthesis. **RESEARCH QUESTION:** Why do multi-agent system failures persist across foundation-model upgrades, and how do coordination and verification harnesses shape operational reliability?

This preprint formalizes the existing public evidence record as a technical synthesis. No new experiment, dataset, measurement, or validation is claimed beyond the source article and its cited evidence.

A smarter model can raise the ceiling on what an AI system might do. It does not repair the queue, the memory discipline, the verification harness, or the operating process that decides whether work actually gets done. Gartner's cancellation forecast for agentic AI projects is a useful warning: the failure mode is often value, cost, and control, not raw model IQ [1].

II. KEY TAKEAWAYS

- Model capability is not system reliability. A stronger model can improve local reasoning while queues, handoffs, memory discipline, verification, and completion logic still fail around it.
- Multi-agent failures are often coordination failures. The cited MAST work groups recurring failures around specification issues, inter-agent misalignment, and task verification rather than a single prompt-quality problem.
- Long context is not operational memory. Reliable continuity requires structured state for the active task, latest instruction, blocker, evidence requirement, and exact done criteria.
- Capability horizons are not production guarantees. Longer tasks accumulate more opportunities to drift or stop early, so autonomy needs bounded scopes, recovery, and outcome verification.
- Upgrade the model only when the model is the measured bottleneck. When the surrounding process leaks state or confuses progress with proof, reliability work belongs in the harness and operating model.

III. THE FIELD PATTERN

I have watched an autonomous multi-agent system fail in a way that was easy to misdiagnose. It had access to stronger models over time. The underlying LLMs became more capable, more fluent, and better at local reasoning. Yet the system still failed to complete simple end-to-end work. Not because the model could not write a paragraph, inspect a page, summarize a source, or reason through a next step. It failed because the surrounding coordination layer kept dropping the thread.

The symptoms were familiar: tasks were accepted but not finished, context was summarized too loosely, verification was skipped or overstated, handoffs lacked enough state for the next agent, and recovery logic confused partial progress with completion. Upgrading the model changed the tone of the failures. It made them faster and more articulate. It did not make them reliable.

That distinction matters. A model can be impressive inside a single prompt and still be embedded in a broken system. Once work spans tools, tabs, files, queues, agents, source checks, screenshots, or human-visible outcomes, reliability is no longer a model property. It is a system property.

IV. MULTI-AGENT FAILURES CLUSTER AROUND COORDINATION, NOT JUST COGNITION

TABLE 1

Observed evidence	Named finding	Reliability implication
1000+ annotated execution traces	Failures studied across 7 multi-agent frameworks	The failure surface is system-level, not isolated to one prompt style
14 failure modes	MAST taxonomy	Failures need diagnosis categories, not a generic "use a better model" answer
3 top-level clusters	Specification issues, inter-agent misalignment, task verification	The recurring defects are process, coordination, and verification defects

V. THE WEAKEST LAYER WINS

The reliability ceiling of an agentic system is set by its weakest coordination layer. If task boundaries are ambiguous, a stronger model will pursue the wrong task with better prose. If the handoff protocol loses state, the next agent will reason from a distorted premise. If the verifier only checks that something happened rather than checking that the right thing happened, the system will confidently report progress while the user-facing result remains absent.

The MAST taxonomy is useful because it names the shape of this problem. Its failure clusters include specification issues, inter-agent misalignment, and task verification [2]. Those are not solved by swapping in a larger model in the same broken harness. They are solved by narrowing task contracts, preserving context, defining stop conditions, and requiring evidence before status claims.

This is why "the model is smarter now" can become a dangerous sentence. It tempts the team to keep a weak operating process and expect the model to compensate. In practice, the model absorbs some mess, then produces a polished version of the same operational confusion.

VI. CONTEXT IS NOT CONTINUITY

Longer context windows help, but they are not memory systems. They allow more text to fit inside the prompt. They do not guarantee that the system preserves the right facts, weighs them correctly, or retrieves the decisive detail at the moment of action. "Lost in the Middle" showed that models can perform worse when relevant information sits in the middle of a long context, even when the model technically supports long inputs [4].

That maps closely to agent operations. A multi-agent system can carry a large transcript and still forget the one rule that matters. It can preserve a long history and still miss the latest user correction. It can summarize ten steps and quietly omit the blocker. The fix is not merely "give the model more context." The fix is to make continuity structured: active task, owner, evidence required, latest user instruction, current blocker, next safe action, and exact criteria for done.

Unstructured context is a pile. Operational memory is a ledger. Reliability comes from the second one.

VII. CAPABILITY HORIZONS ARE NOT GUARANTEES

METR's time-horizon work is one of the more useful ways to discuss agent capability without hype. It translates AI performance into the length of human work that models can complete with a given probability. The headline is impressive: frontier AI time horizons have been improving rapidly, and the paper estimated about a 50-minute 50% completion horizon for current frontier software tasks at publication time [3].

But a 50% completion horizon is not the same thing as operational reliability. A system that succeeds half the time on tasks of a certain length may be a research milestone. It is not a dependable production worker unless the harness catches failures, retries safely, narrows scope, and verifies outcomes.

The more important lesson is not that agents are useless. It is that autonomy has a measurable failure curve. Longer tasks accumulate more chances to drift, repeat, omit, misread, or stop early.

When teams ignore that curve, they assign long-horizon work to a fragile loop and then blame the model when the loop collapses. The model mattered. The harness mattered more.

VIII. AI ADOPTION IS BROAD, BUT SCALED VALUE REMAINS NARROW

TABLE 2

McKinsey finding	Reported level	What it suggests
Regular AI use in at least one business function	88%	Access and experimentation are no longer the hard part
Organizations not yet scaling AI across the enterprise	Nearly two-thirds	Operating model and workflow integration remain bottlenecks
Organizations reporting enterprise-level EBIT impact	39%	Usage does not automatically convert into material business value
Organizations at least experimenting with AI agents	62%	Agent curiosity is ahead of reliable deployment discipline

IX. THE ENTERPRISE VERSION

The same pattern appears at enterprise scale. McKinsey found that AI use is widespread, but most organizations remain early in scaling and enterprise-level value capture [5]. Gartner's agentic AI warning points in the same direction: cost, unclear value, and inadequate risk controls are enough to cancel projects even while the technology itself improves [1].

That is the core trap. Leaders buy capability, then discover they also needed operating discipline. They needed workflow redesign, ownership, evaluation suites, governance, exception handling, and evidence standards. Without those layers, a

better model is like a faster engine bolted to a loose frame. It can move. It cannot be trusted at speed.

NIST's AI Risk Management Framework is valuable here because it treats trustworthy AI as a lifecycle practice for organizations designing, developing, deploying, and using AI systems [6]. That framing is less glamorous than a model release, but it is closer to the work that actually determines whether a system survives contact with reality.

X. THE UPGRADE FALLACY

The upgrade fallacy says: "This failed because the model was not smart enough." Sometimes that is true. There are tasks where the model simply lacks the reasoning, perception, tool-use, or domain capability required. But in the field pattern I am describing, the failures persisted after model upgrades because the failure was not located inside the model alone.

The real defects were mundane and severe. The system did not always know what promise it had made. It did not always preserve the latest correction. It did not reliably distinguish "submitted," "queued," "observed," "verified," and "complete." It sometimes treated the absence of an error as proof of success. It sometimes stopped after producing an artifact that still needed validation. A smarter model can describe those distinctions. The system has to enforce them.

Once you see that, the design question changes. You stop asking, "Which model can we upgrade to?" as the first move. You ask, "Where does the work lose truth?" That question points to contracts, instrumentation, replayable logs, bounded tasks, and verification gates. It points to engineering.

XI. WHAT ACTUALLY FIXES IT

- Define tasks as contracts: owner, input, expected output, evidence required, timeout, and explicit done criteria.
- Separate progress from proof: queued, attempted, observed, verified, and complete should be different states.
- Make context structured: carry the latest user instruction, current blocker, decision log, and next action as machine-readable state.
- Design verification before autonomy: every external claim needs a source, probe, screenshot, test result, or reproducible artifact.
- Keep evaluations close to the real workflow: benchmark the full harness, not only the model's answer quality.
- Use stronger models deliberately: upgrade them when the measured bottleneck is model capability, not when the process is leaking state.
- Treat failures as system telemetry: every drift, premature stop, false completion, and weak proof claim should harden the harness.

A smarter model is still worth using. The mistake is expecting it to rescue an unreliable system from its own coordination debt. The model can raise the possible ceiling. The process determines the actual one.

XIII. LIMITATIONS AND CLAIM BOUNDARY

The paper combines qualitative field observation with secondary literature. It does not provide a new controlled model-comparison dataset, and causal claims about model upgrades versus harness quality remain bounded by the cited evidence.

Claim boundary - None beyond preserving the source's distinction between model capability and system reliability and avoiding causal claims not supported by the cited evidence.

XIV. REFERENCES AND SOURCE RECORD

1. Gartner — “Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027,” June 2025
2. MAST / NeurIPS 2025 — “Why Do Multi-Agent LLM Systems Fail?,” 2025
3. METR — “Measuring AI Ability to Complete Long Software Tasks,” March 2025
4. Liu et al. (TACL) — “Lost in the Middle: How Language Models Use Long Contexts,” November 2023
5. McKinsey & Company — “The State of AI in 2025: Agents, Innovation, and Transformation,” November 2025
6. NIST — “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” January 2023

AUTHOR BIOGRAPHY

EXZIL CALANZA is an independent researcher and software builder based in Iloilo City, Philippines. His work focuses on browser and desktop automation, reproducible evaluation, agent reliability, workflow verification, and AI-assisted systems engineering.