Three Studies, 20,000 Conversations: What Governance Actually Changes

Governed and ungoverned AI can look almost identical on raw task completion. The difference shows up the moment you ask whether the task was done correctly — and now there are numbers for it.

TL;DR

Across three studies and 20,000+ production conversations, governed AI didn't finish more tasks than ungoverned AI — it finished them correctly: verified completion of 53% vs 9% and clean exits of 72% vs 23%, against a control running the same model, with zero character breaks or drift. Raw completion rates hide that gap entirely; quality is where governance shows up — and where the business value lives.

5.7×
more likely to complete the task correctly than the ungoverned control

3.1×
more likely to exit the conversation cleanly

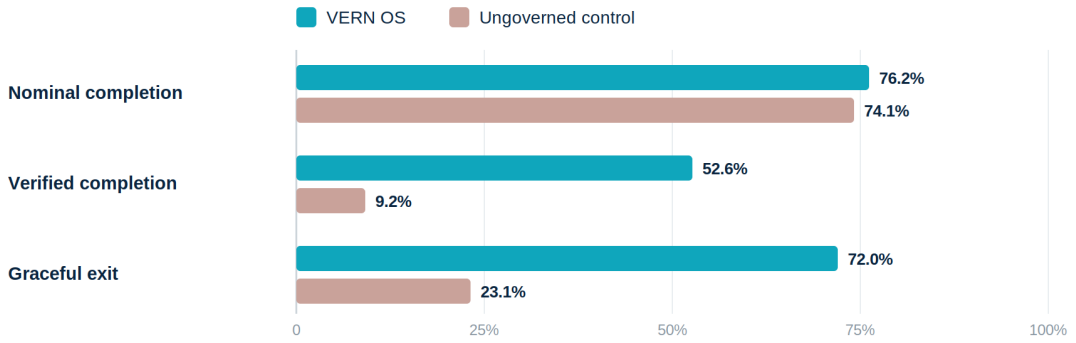
0
character breaks or drift incidents across the sample

Most AI evaluation stops at one question: did the system finish the task? It sounds reasonable, and it's incomplete. VERN's research portfolio — three studies across more than 20,000 production conversations — suggests raw completion rates hide the difference between an AI that merely reaches the end of an interaction and one that does the job correctly. Enterprises aren't deploying AI to have more conversations; they're deploying it to complete workflows, stay inside policy, reduce escalation, and produce outcomes they can trust. So the better question isn't “did the AI finish?” It's “did it finish correctly?”

The metric most benchmarks miss

Study 1 ran the cleanest comparison in the portfolio: the same underlying language model, with and without VERN OS's behavioral control layer, across 8,374 production sessions. On the surface, the two looked almost the same. Nominal task completion — whether the session reached an endpoint — was 76.2% with VERN and 74.1% without. On that metric alone, you'd conclude governance barely mattered.

It mattered enormously once “completion” was held to an enterprise-grade standard: goal achieved, stayed on-goal, no tangents, clean close. By that measure, governed sessions verified at 52.6% versus 9.2% for the ungoverned control — roughly five to six times as often — with zero character breaks and zero drift incidents across the entire sample.



Same underlying model, with and without VERN OS · Study 1, N = 8,374.

How to read these metrics

Term	What it means
Nominal completion	The session reached an endpoint — the conversation finished.
Verified completion	The goal was achieved while staying on-goal, with no tangents and a clean close.
Graceful exit	The conversation ended cleanly — no looping, collapse, hard failure, or leaving the user unresolved.
Character break · drift	The AI left its role or departed from the deployment's objective: persona breach, instruction leakage, or lost task state.

Governance didn't make AI finish more conversations; it changed *how* they finished. A raw completion rate would call the two systems equivalent. They aren't — one is completing the job in a way an enterprise can actually trust several times as often. (The control group's rare clean completions came from a self-selected, unusually persistent set of users, which means VERN's real advantage is probably wider than the chart shows, not narrower.)

Completion is not the same as correctness

This is where AI programs get misled. A conversation can "complete" while still failing the business: the user reaches the last step, but the AI wandered off-goal, lost the thread, skipped the intended path, or ended without a clean resolution. That failure doesn't show up in a raw completion metric — it shows up later, in repeat contacts, escalations, abandonment, compliance review, and human cleanup.

Verified completion is the stricter, more useful standard: did the interaction reach the intended outcome while preserving role, policy, and direction, and close cleanly. The related signal is the graceful exit — VERN's 72.0% versus 23.1% for the ungoverned control. A graceful exit means the conversation ended cleanly instead of looping, collapsing, or leaving the user unresolved. For an enterprise, that gap isn't cosmetic; it's operational cost.

Emotion became measurable

Study 2 asked whether emotional trajectory can be measured across real interactions. It reviewed 7,295 conversations across 21 AI Human personas, and every persona showed positive movement in user emotional state from the start of a session to the end. Among users who arrived distressed, 76.9% left calmer in companion deployments and 84.6% in task-focused ones.

That second number reframes how to think about emotional intelligence in AI. The clearest emotional benefit didn't come from warmth or open-ended empathy — it came from *progress*. When a user knows what's happening, feels guided, and reaches a clear next step, their emotional state improves because the interaction is working. Emotion, in other words, isn't a soft experience layer; it's an operating signal that tells the business whether the interaction is helping or hurting the person on the other end — did frustration fall, did anxiety stabilize, did the session end better than it began.

Two things make the measurement credible: it was scored by VERN's own emotion system on a per-turn basis, not by an LLM grading another LLM — sidestepping the model-judging-model circularity common in AI evaluation — though it's still VERN's own instrument, pending independent validation.

Against the industry's own benchmarks

Study 3 compared VERN against published figures from Gartner, Salesforce, and the Guardrails AI Index, in the task and customer-service contexts where external benchmarks exist. On that subset, governed clean-exit rates reached 88.5% (against an unorchestrated baseline near 28%), with task completion in the 74–88% range — at or above the 70–90% published for industry leaders, and well clear of RAG-based systems at 40–65%.

The sharper point is what the industry *doesn't* measure: persona integrity, drift prevention, real-time emotional steering, whether the AI held a behavioral mandate under pressure. There's no established benchmark for these because most systems can't do them. VERN reports zero character breaks and zero drift across monitored deployments. As AI becomes the interface between organizations and people, that behavioral integrity stops being a nice-to-have and becomes part of product performance — an AI that finishes a workflow but breaks role or leaves an unresolved exit hasn't really succeeded.

Control runs both ways

The most useful finding across the portfolio is that governance is not the same as making AI more agreeable. In Track B, certain personas — Amber and Christine among them — were intentionally built to provoke confrontation, sustain fear, or create discomfort inside a bounded experience. They still achieved 100% behavioral success with zero character breaks while holding high clean-exit rates.

That's the proof the system isn't just smoothing tone. A governance layer that can only make AI nicer is limited; one that can steer toward different emotional objectives by use case is far more powerful. A governed character that players can't break, manipulate out of role, or extract secrets from is a fundamentally different product than what's on the market today — and the

same control that makes that possible is what keeps a support agent calm. The target changes with the job: de-escalation for support, reassurance and protocol adherence for healthcare, trust for sales, pressure for a training simulation, suspense for entertainment. The common requirement isn't positivity. It's control.

The honest read

These studies are a meaningful step forward, and they should be read precisely. Study 1 is the strongest causal signal in the portfolio because it compares the same underlying model with and without VERN OS across production sessions — more rigorous than a single uncontrolled readout. Still, this is production data, not a fully randomized lab trial. The industry comparisons in Study 3 are published external figures, not a matched head-to-head test VERN ran against every competing system. And Study 2's emotional results come from VERN's proprietary emotion system — which avoids the LLM-grading-LLM problem but still needs independent validation over time. Those caveats matter, and they don't erase the pattern: across all three studies, raw capability isn't the differentiator. Governed behavior is.

Why it matters

Most of the market is still racing on capability — bigger models, faster inference, longer context. That's becoming table stakes. The harder enterprise problem is what happens when AI is live with real people: can it stay inside the job, complete the workflow correctly, adapt to emotional state, avoid drift under messy input, exit cleanly, and prove what happened. That's the layer VERN OS governs. There's an efficiency angle, too: VERN reports reaching governed task completion with a single controlled agent in place of the multi-agent committees of 3–8 LLM roles teams often assemble to get reliability.

The business case isn't that governed AI talks more. It's that it does the work correctly more often, with cleaner exits, stronger behavioral integrity, measurable emotional trajectory, and an audit trail the enterprise can trust. When AI is the interface, behavior is the product — and behavior is now measurable.

> **VERN OS is the runtime governance layer for AI-human interaction.** See the full research portfolio or run it on your own workflow — vernai.com.

Capability is table stakes. Behavior is the product.