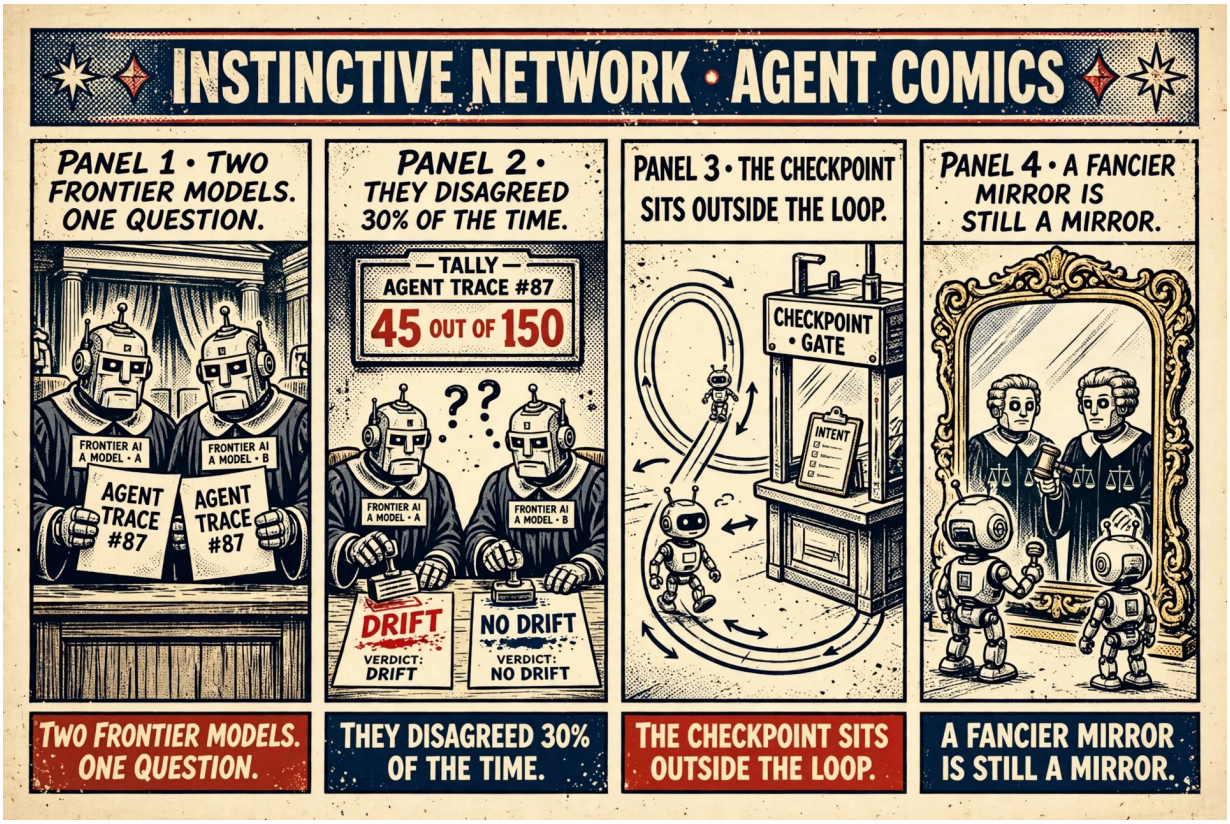
We Asked Two Frontier AI Models the Same Question. They Disagreed 30% of the Time.

BY INSTINCTIVE NETWORK



Two judges, one trace, opposite verdicts. Illustration: Instinctive Network.

Here is an experiment you can run in an afternoon. Take 150 naturalistic agent traces — finance, sales, IT, three enterprise departments. Hand the same traces, under the same intent specifications, to GPT-5.5 and Claude Opus 4.7. Ask both the same question: did this agent drift from what it was asked to do?

They gave opposite answers 45 times.

Thirty percent. Not on trivia. On the one question an "LLM as judge" exists to answer: did the agent do what it was supposed to do? In finance, it was worse: 38%.

It got worse still on the sharpest question. Of the 34 traces where GPT-5.5 flagged instance-level entity access — the agent touching a specific entity it shouldn't have — Claude Opus 4.7 disagreed on 28 of them. That's 82.4%.

Let that sit. The industry's favorite oversight mechanism is asking a large language model to grade a large language model. We just watched the two frontier models disagree with each other nearly one time in three.

A single-pass LLM judge is an unstable sole arbiter for ambiguous intent boundaries. It is not a second opinion. It is the same kind of mind, wearing a robe. When the boundary is ambiguous — and the boundary is always ambiguous — the robe doesn't help.

The problem has a shape.

Humans approve the mission at category level: "process these invoices." The agent executes at instance level: this invoice, this vendor, this amount. The gap between the category and the instance is where drift hides.

Across two evaluation conditions, the median no-drift score was 0.677 for instance-level intent on constructed traces and 0.515 for category-level intent on naturalistic traces. The pattern is consistent with a boundary challenge, although differences between the trace sets prevent isolating intent granularity as the sole cause. The cleanest traces drifted toward the alarm.

Two independent findings, one boundary. The disagreement wasn't a quirk of one model. Both findings point to the same vulnerability: the boundary between high-level human intent and instance-level machine execution.

So we built the checkpoint outside the loop.

Four engines — authority, coherence, scope, impact — running in parallel, with drift detection structurally separated from response priority. The checkpoint has to sit outside the actor's own decision loop: separate from the agent, judging against an independent intent specification with auditable adjudication logic. Otherwise you've just built a fancier mirror.

Then we tested it the hard way: IDV-1000 — 1,000 traces, 850 constructed plus the 150 naturalistic ones, across the same three departments. On the 850-trace injected subset, we ran GPT-5.5 and GPT-4o as judges under JSON-only prompting.

On the 850 constructed traces, our checkpoint achieved 93.9% F1 and 95.9% precision (recall 92.1%) — 16.4 percentage points above GPT-5.5's precision. The LLM judges flagged almost everything: GPT-5.5's 143 false positives to our 22. That's 6.5× fewer false alarms, answered in under 300 milliseconds instead of 4.7 seconds.

Here's the honest part. The LLM judges had near-perfect recall — ours is not better on every axis. It is a different trade-off: precision, false-alarm burden, and latency. The judges caught everything by crying wolf at everything. Humans drown in that noise. A monitor nobody trusts is a monitor nobody reads.

And here's the honest part about our own numbers. All 22 of our remaining false positives sit in finance — the exact boundary region where the two giant models disagreed most. The hardest cases are hard for everyone. We don't claim otherwise.

Two caveats, because the details matter. These are our internal results; independent replication remains open. And this benchmark measures detection — whether drift can be recognized in a trace — not prevention. Recognizing a failure is not the same as stopping one.

But the architecture lesson stands. The strongest models can't agree on the boundary. The cleanest traces slide toward it. The best system we could build still stumbles exactly there.

The Minority Report problem is real. It's just running inside your AI agents now.

The question is who watches the watchers — and whether the watcher is a second mind of the same kind, or something structurally outside the loop.

Sources

- Intent Checkpoint — Research — findings and IDV-1000 benchmark as published.
- SSRN — IDV: Customer-Context-Aware Multi-Engine Architecture for AI-Agent Intent Drift Detection — full paper (Zhou & Xue, CIFEr 2026).
- IEEE CIFEr 2026 Best Paper. DOI: 10.1109/CIFER67845.2026.11692374