How We're Wrong on Purpose

A diligence page for a market where every governance claim is self-attested

Every vendor in the AI-governance market makes accuracy and governance claims. Nearly all of them are self-attested: the company grading its own homework. You cannot diligence a claim like "our agents are reliable" directly.

What you *can* diligence is a company's track record of publishing findings **against itself** — because that record is expensive to fake and impossible to backfill. Here is ours, from the last ninety days. Each item is either public or available for inspection in a diligence session.

1. We retitled a published essay because our own thesis failed verification — and said so in the essay. The working frame for essay #49 ("AI raises the mean and explodes the variance") was our founder's own public claim. Two independent research passes found the most-cited field experiments say the opposite for bounded work. The published essay opens by owning the correction ("So the frame was wrong"), rebuilds on the evidence that survived, and the original title was retired because it named the wrong problem. *Public: shyanming.substack.com/p/the-selection-problem.*

2. Our research program pre-registers hypotheses and publishes its own nulls. Recent internal R&D on our verification tooling recorded, verbatim, in the permanent record: "reliability advantage did NOT replicate," "the manifest was NOT the active ingredient," "the 'cliff' finding was wrong," and "null result, and the design flaw was mine." Four refuted hypotheses in one series, kept, not buried. *Available for inspection; portions scheduled for open-source release.*

3. When our own system fails silently, the incident becomes a named control. Three self-caught production incidents in one week: a config store that grew past a database hard limit because a sync error was silently swallowed, and two build-pipeline failures that bricked deploys. Each fix commit names the incident, and the silent-failure path was converted into a logged, bounded, monitored one. No customer reported any of these; our own operation of the system did. *Commit-level receipts available in diligence.*

4. Our ungoverned baseline caught an error in our governed layer — and we wrote it up as a hazard class. While benchmarking our governed interface, the raw-model control arm flagged a value our curated layer had wrong. We verified against the binding source, corrected it, and documented the general lesson: a curated layer whose values are not verifiably derived from cited sources is itself a new hallucination surface. The finding names our own product's failure mode. *Write-up available in diligence.*

Why this matters to you: the claim is *not* that we make fewer errors. It is that errors in our operation get caught by the operation itself, recorded where they cannot be quietly deleted, and turned into controls — before a customer or an auditor finds them. In a self-attested market, a paper trail of self-corrections is the only governance claim that arrives with its own evidence.

JieGou · jiegou.ai · Current as of 29 July 2026