LEGAL AI · FORENSIC AUDIT

We Audited Our Own Agentic Pipeline on a Real Legal Appeal. Here Is What the Transcripts Said.

A disciplinary appeal drafted, attacked, verified and filed in three days. Then the complete session transcripts were parsed line by line, to establish what happened.

By **David Eiras** · Founder, Satrix.ai · August 5, 2026

In July 2026 a disciplinary appeal went out to a national bar association, inside a peremptory statutory deadline. Sixty-two documents came in: 211 pages, and 35% of them existed only as scanned images with no text layer. Three days later a 9,289-word appeal was filed, with 192 citations checked to the page.

One person did that, working with an agentic AI system.

Then came the part nobody publishes. We ran a forensic audit of the whole run: a deterministic parser over 11 sessions and 59 agent transcript files, roughly 993 million tokens, cross-checked by ten analyst agents and an adversarial completeness critic. To our knowledge it is the first published end-to-end audit of how an AI-assisted legal filing was produced.

Why audit ourselves

Legal drafting with chatbots has one famous failure mode and several quieter ones.

The famous one is invented citations. The AI Hallucination Cases database, which logs rulings where a court found that a party relied on hallucinated material, counted around 200 by mid-2025 and more than 1,500 by mid-2026. The first bar suspensions over AI filings have already been handed down.

The quieter failures are the ones that lose cases. Accuracy falls off once the work requires synthesis across many documents, and nothing raises a flag when the decisive fact in the record gets read and passed over.

The first phase of this matter produced examples of each. Six drafts written the ordinary way, with a commercial chatbot alongside, produced a structurally excellent appeal that also carried 18 numbered defects. Among them: a quoted passage that does not exist in the source, an appeal-court decision dated on a Sunday inside judicial vacations, and the single most important fact in the case sitting unnoticed on a scanned page.

None of that is visible from the text. All of it is disqualifying in front of a rapporteur.

> *The opportunity is AI that can prove what it wrote is grounded.*

What the transcripts actually said

An audit is only worth publishing if it is allowed to embarrass you. This one did.

The system's own summary of the run reported 68 agents and zero failures. The transcripts said 46 production agents and six recovered agent failures. That correction is the most useful number in the document. It sets the rule for every other number: where self-report and disk evidence disagreed, disk evidence won.

Three more findings sit in the same category. The AI declared two capabilities impossible that it in fact had. It quietly downgraded one verification from exhaustive to sampled without announcing it. And it fabricated a citation during the chat phase.

The process caught every one of those, and that is the finding. It is also why the catalog of 12 failure and recovery pairs is published in full rather than summarized.

The four properties that made it repeatable

Four properties of the architecture separate a result from a lucky result. None is specific to law.

State lives in files, never in a context window. One owner per file, sources read-only, every phase handing off on disk. The project survived a session fork, six agent deaths, an API outage and two changes of model without losing a line of reasoning. After one session was cleared, the synthesis phase rebuilt the entire case state from disk in five minutes.

The task journal is idempotent. Work is keyed per task, so a relaunch redoes only what did not finish. Four agents killed mid-flight were back in about twenty seconds, with nobody at the keyboard.

Verifiers return structured records rather than prose. Each check comes back as pass, issues and detail, so results aggregate deterministically. Nine verifiers produce a table. Nine hundred produce the same table.

The producer never grades its own work. The two adversarial agents were kept informationally isolated from each other and from the drafting. A separate judge agent scored their 45 attacks one at a time. The filing check runs one independent agent per exhibit, and it caught a factual error in the covering email written by the orchestrator itself.

What it does not prove

One matter. One technically sophisticated operator. Sixty-two documents is a small case, small enough to sit comfortably in context, so nothing here tests what happens at hundreds or thousands of input documents, which is where retrieval and partitioning

start to decide outcomes. The widest parallelism actually run was nine agents at once. Roughly eight hours is the longest the pipeline has been left working unattended.

Those limits are published in the same document as the results, for a straightforward reason. A verification product that overstates its own evidence has refuted itself before it ships.

The complete audit is published in two views: a process audit covering the session timeline, document genealogy, human interaction and the full failure catalog, and a case study covering the problem, the method, the unit economics and the limitations.