Can AI Value a Legal Claim? What Two Failed Models Taught Us

two preregistered experiments, two rejected candidates, and the valuation framing that survived them

Argentis Labs Research · July 26, 2026 · 8 min read

Can a claim change in value on the days when nothing is filed?

A motion denied in March may continue to reprice a case in April without a single new page hitting the docket. Eight months of silence may not be the absence of information. Relative to what was expected, it may reveal something about court congestion, the parties' leverage, or a settlement window that is opening or closing.

Yet many models of what a legal claim is worth read the case as a stack of documents and stop there. The gap between how claims reprice in practice and how we model them is the subject of our research program. This is what we found, including the part that did not work.

Calendar time tells us how long a case has existed—not how far it has moved

Financial modeling has long recognized a related problem. On a busy trading day, prices may move more in an hour than they do in a quiet week. The market's effective clock ticks with activity, not merely with the calendar. This insight motivated models that index risk by market activity rather than treating every unit of calendar time as equivalent.

Litigation keeps time in much the same way. A lawsuit may accelerate in the weeks after a denied motion to dismiss, nearly stand still through a discovery stall, and yet never fully stop. Six months may contain several dispositive events in one case and almost no procedural movement in another. Calendar time remains necessary for measuring the interval, but it is not a reliable index of the case's institutional progress. Silence becomes informative only relative to an expectation: eight months is merely duration; eight months where six weeks was expected is news.

That is why a claim cannot be valued as a document or a snapshot. A claim is a moving position inside an institution. Its value today depends on where it sits in that motion: what is likely to happen next, how soon, and what that uncertainty is worth in present dollars.

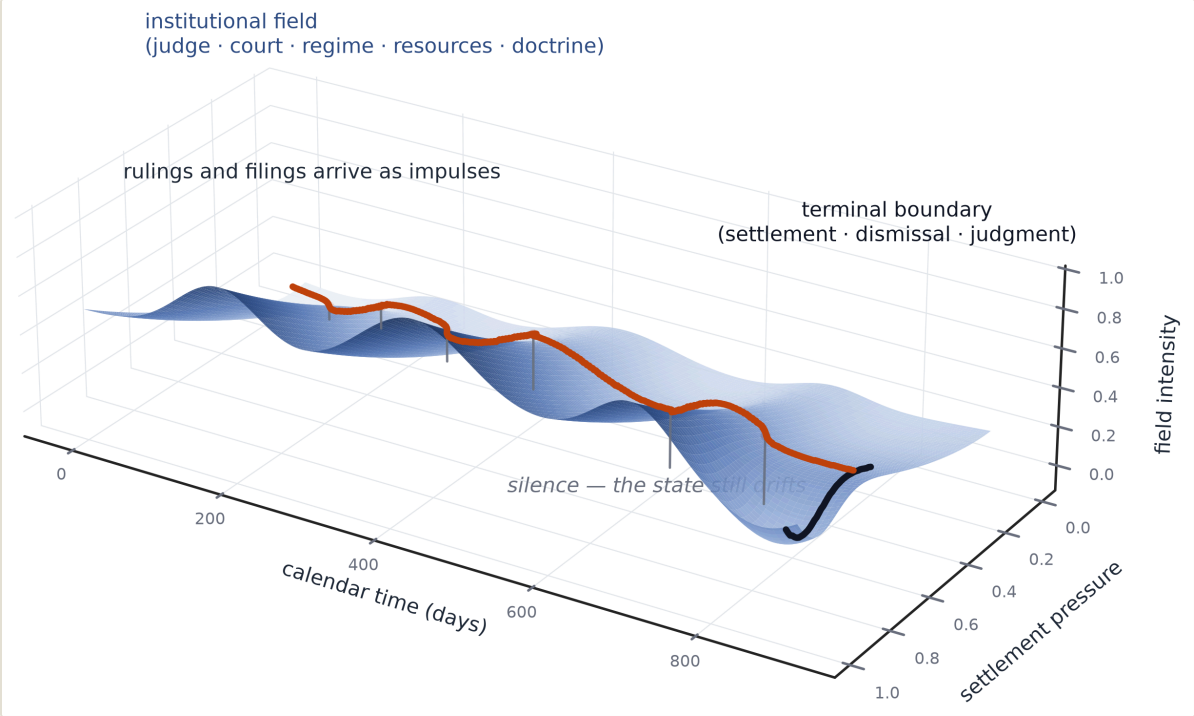

Figure 1. A stylized claim trajectory through an institutional field. Filings and rulings arrive as impulses; the trajectory drifts during silence; some regions of the field act as settlement basins. The value question concerns the whole path, not any single point along it.

What we tested, and what failed. Twice.

Our thesis favored the newer AI tools. Because legal events arrive irregularly—three filings in a week, then silence, then a settlement—models built to handle continuous, uneven time should, in principle, capture something that conventional models miss. That was the bet.

We built synthetic litigation worlds: controlled environments in which the true mechanics are known, allowing a model's behavior to be evaluated against ground truth. We then froze the pass/fail rules before training each candidate model. These were not guidelines. They were explicit thresholds a model had to clear to survive.

Then we ran the experiments. The conventional models won. Gradient-boosted trees and a standard time-aware Transformer matched or beat the continuous-time architecture we had expected to add value. The candidate did not approach the survival threshold, so we removed it and published the negative result with the full experimental record.

That could have been the end of the story. It was not, because a failed model leaves an important question: did the idea fail, or did our particular design fail?

We therefore dissected the trained model piece by piece. The clock component mattered, but not because it was acting like a clock. A simpler component given the same case history and

elapsed time performed just as well, and rearranging the order of the time gaps did not change the result. What looked like a clock was really functioning as additional processing capacity.

We retired that design and tested the strongest remaining repair. Instead of giving the model a clock and hoping it learned to use it, we supervised time directly: training the model on the likelihood of each event occurring when it did while also accounting for the silence between events. This is a standard treatment for irregular event streams. We froze new threshold criteria and ran the experiment.

It lost too, by a wide and unambiguous margin against the same conventional baseline.

Two preregistered experiments. Two rejected candidates. The pattern is now difficult to ignore: in the synthetic worlds we built, neither raw elapsed time nor an explicitly supervised timing-surprise signal added reliable predictive value once the procedural events themselves were visible.

The record matters

A model that performs well on data it was trained or tuned against tells you very little. The data may have flattered it, the comparison may have been soft, and the versions that missed rarely make the slide deck. Anyone buying a forecast—or allocating capital based on one—needs to know something harder: how the method behaves when the test is designed to resist flattering it, and what happened to the candidates that did not survive.

The record does real work. It tells us that simply making the model larger is not yet a justified next step. The better question is whether real court dockets carry signal that simpler models cannot already reach, and whether a fair, repeatable test can be built from public records at all.

One caution belongs in plain view. Everything above was measured on synthetic data under controlled conditions; no result yet touches real litigation. The result is also conditional on the synthetic data-generating process. If timing was largely redundant once the event history was known, no architecture could recover much independent timing signal. The real-docket phase will test whether that redundancy holds outside the generator. It is the next phase, and it faces the same rules.

What this means for attaching value to a claim

Even without a winning temporal engine, the framing survives—and it changes how a disciplined valuation process should work.

A claim is a distribution of paths, not a yes-or-no question. Dismissal, settlement, summary judgment, trial, appeal, transfer, and continued survival each carry their own recovery distribution, duration, cost profile, and capital requirement. Valuation starts by mapping those paths, not by picking a winner.

Time and value are coupled. A $10 million recovery at a 60% probability is a $6 million expectation only before you ask when it arrives, what it costs to wait, and what else can happen while you wait. The useful object is a time-conditioned recovery distribution, translated into a risk-adjusted reservation value—not a point estimate of the claim.

Repricing should be event-driven, and silence can become an event. Silence is not automatically information; it becomes information when the observed interval departs from what the case's state would lead us to expect. Eight months is merely a quantity until it is compared with an expected interval of six weeks.

When a material ruling lands, the distribution should move. When an interval becomes unexpectedly long without one, that may justify revisiting the valuation too. In our simulated cases, once the model knew what had already happened procedurally, adding information about how long it had waited for the next event did not make its forecast more accurate. That does not prove delays are irrelevant. It means we could not show that they added information beyond the docket history in these experiments. An unusually long delay may still be a sensible reason for an underwriter to revisit the valuation.

The credible frontier is calibrated ranges. A system that says, "Here is the range, here is what would move it, and here is how often we have been right at each horizon," is worth more than one that produces a confident single figure. Capital discipline—what to fund, what to syndicate, when to exit, and how much apparent diversification disappears when claims share a court or procedural bottleneck—depends on the quality of the ranges, not the precision of the points.

These are not product capabilities, and they are not causal claims about how courts behave. They are the questions a defensible underwriting framework should eventually answer, stated now so that the work can be judged against them.

The bottom line

We began with a thesis about sophisticated models and let the evidence reject it twice, under rules we froze in advance. That is not a detour from building a valuation capability. It is what building one looks like when the standard is methods you can defend, not demos you can sell.

The next phase asks the harder question in the harder arena: whether public docket data can support that standard without confusing what is observable with the legal process itself.

One hypothesis for that phase is that the **informational value of delay may itself be changing**. If AI lowers the cost of generating discovery and procedural work faster than courts expand their capacity to process it, unexpected silence may become more informative—not because calendar time suddenly becomes the case's clock, but because deviations from expected timing may increasingly reveal congestion, strategy, and resource asymmetry. That proposition remains untested.

In litigation finance, the disciplined rejection of a weak signal may be worth as much as the discovery of a strong one.

The full study is open in the liquid-legal research toolkit.

Research, not legal or investment advice. All results to date are based on synthetic data; nothing here values any real claim.
