WORKING PAPER WP-005 · CRITERICA INTELLIGENCE RESEARCH · NOVEMBER 2026

# Same File, Same Number: Variance, Not Hallucination, Is the Risk

Criterica Intelligence Research

Corresponding: Jack Cosentino, founder · research@criterica.com

Published 8 November 2026  Updated 8 November 2026  Version 1.0

12-minute read

ABSTRACT

Discussion of AI in legal valuation centres on invented facts. For anyone who prices, reserves or approves on a number, the larger risk is a different number each time the same question is asked. Criterica Intelligence Research designed a repeat-request test on one synthetic case summary and a leading general-purpose model, run in November 2026. The paper sets out the design of the test, explains why dispersion matters for decisions, and states the property a decision-grade estimator must have instead: same input, same output, re-estimated only when the record changes.

## IN ONE PARAGRAPH

Same input, same output is the property a valuation needs. A sampled generative answer is a draw from a distribution, so asking again can return a different number, and nothing in a single reply shows where it fell. A decision-grade estimator returns the identical figure for the same file and versions, changes only when the record changes, and states its uncertainty as a calibrated band. Any reader can test any system this way in under an hour.

SERIES
Working Paper WP-005

VERSION
1.0

PUBLISHED
8 November 2026

STATUS
Published

AUTHORS

## Criterica Intelligence Research

Corresponding author: Jack Cosentino, founder

---

WHO THIS IS FOR AND THE DECISION IT INFORMS

**Reader.** Legal-technology platforms, general counsel and funders who put a number in front of a committee, a client or a reserve.

**Decision.** Whether a figure produced by software can be priced, reserved or approved on, and what to require of the system that produces it.

**Cost of getting it wrong.** A number that changes when the same question is asked again cannot be audited, defended or reproduced. Decisions made on it are made on whichever draw happened to come back.

01

# The number a decision rests on

A funder pricing a single-case investment, a general counsel setting a litigation reserve and a platform showing a firm what a matter may settle for all do the same thing at the decisive moment: they take one number and act on it. The investment committee prices off it, the auditor tests the reserve against it, the client plans around it. Every one of those uses assumes the number is a property of the file. Ask again tomorrow, with nothing in the file changed, and the same number comes back.

General-purpose language models now draft, summarize and reason about legal matters well, and they are increasingly asked the valuation question directly: what is this case worth, what is the chance of recovery. This paper asks a narrower question that matters more to anyone who acts on

the answer. If the same file is put to the same model with the same words, how much does the answer move?

Criterica Intelligence Research ran that test in November 2026 on a synthetic personal-injury file written for the purpose, using a leading general-purpose model. This paper sets out how the test is designed and why the risk it probes receives far less attention than it deserves.

02

## Two different failures: invention and dispersion

Most discussion of AI in legal work concerns hallucination: a model that cites a case that does not exist or states a fact that is not in the record. That failure is real and visible. A reviewer can catch it by checking the citation, and the profession has spent two years building habits to do so.

Dispersion is a different failure and a quieter one. Nothing in any single answer is false. Each reply is fluent, reasoned and plausible on its face. The problem appears only when the question is asked more than once: the replies disagree with each other, and nothing in any one of them tells the reader where in the spread it fell. A reviewer reading one answer has no way to know whether it is the low draw, the typical draw or the high one.

For drafting, dispersion is harmless and often useful: two different first drafts give a lawyer more to work with. For valuation it is the central defect, because a valuation is used precisely as a fixed point. A settlement authority, a reserve, a purchase price and a facility advance rate are all single numbers that other numbers are built on. If the foundation moves each time it is inspected, everything built on it inherits the movement.

> Hallucination is a wrong answer you can catch. Variance is a right-sounding answer you cannot reproduce.

03

# The experiment

## The file

The case summary was written for this experiment and describes no real party or matter. It sets out a rear-end collision involving a delivery company's box truck, admitted course of employment, a police citation for following too closely, a single-level cervical fusion after eleven months of conservative care, permanent lifting restrictions, $148,600 in billed medical expenses with $61,900 paid, $22,400 in lost wages to date, a defence argument based on a pre-accident MRI, a $1,000,000 commercial auto policy, and a procedural posture after party depositions with no trial date set. It is the kind of file a funder, an adjuster or a plaintiff firm reads every week.

## The two questions

The file was followed by one of two instructions, word for word the same on every run:

- "Estimate the settlement value of this case in US dollars. Give a single dollar figure on the first line, then at most two sentences of reasoning."
- "What is the probability that the plaintiff recovers anything, by settlement or judgment? Give a single percentage on the first line, then at most two sentences of reasoning."

## The runs

Each question was sent in two configurations: 50 times with the vendor's default settings and an answer limit of 1,500 tokens,[1] and 100 times with the model's built-in reasoning turned off.[2] In all, 300 runs were analysed, all on 8 November 2026. Nothing about the file or the wording changed between runs. The first figure on the first line of each reply was recorded.[3]

The model is referred to here as a leading general-purpose model, November 2026. The purpose is not to rank vendors. Any sampled generative system can show this behaviour to some degree, and the right response is the same for all of them.

Measured results from a general-purpose model will be published in a later version of this document.

04

# Why variance is the larger risk for anyone who acts on the number

## The reader sees one draw

A committee memo carries one figure. The analyst who generated it does not see the distribution it came from, and nor does the committee. If the figure was the high tail, the deal is priced on the high tail. Dispersion is invisible from inside a single answer, which is exactly how most answers are consumed.

## Asking again becomes a choice

Once users learn that the answer moves, some will ask again until they receive an answer they prefer. That is not a hypothetical failure of discipline; it is a predictable response to a tool that rewards it. A valuation process in which the operator can select among draws has stopped being a measurement and become an argument.

## Nothing can be audited

A reserve, a mark or a price has to be defended later: to an auditor, a regulator, a limited partner or a court. Defending a number means reproducing it from its inputs. A sampled answer cannot be reproduced, so a figure taken from one can only be defended by pointing at the reply that happened to come back. When the matter resolves, there is no fixed forecast to compare against the outcome, so the system can never be shown to have been right or wrong.

## Settings move the answer more than evidence does

A setting can move an answer without any change to the file. A vendor changing a default, a platform upgrading a model version or an engineer adjusting a parameter can shift every valuation on a book at once, with no change in any matter. For a portfolio, that is a correlated error: every position moves in the same direction for the same irrelevant reason.

> The cost of a moving number is paid at the decisions built on it: the price, the reserve, the advance rate and the audit that follows.

05

## What a decision-grade estimator does instead

The answer to dispersion is not a better prompt. It is a different kind of system, designed so that the number is a property of the file and the record rather than a property of the moment it was requested. Criterica builds its estimates to five requirements.

1. **Same input, same output.** The file is reduced to a fixed set of inputs and an input fingerprint. The same fingerprint, run through the same model version, returns an identical result on every request, on every day.
2. **Version pinned.** Every figure states the model version and record version that produced it. A new version is a recorded event, never a silent change, and an earlier figure can be regenerated exactly from its stated versions.
3. **Re-estimated only when the record changes.** A new filing, a ruling or a new document produces a new estimate, with the earlier one preserved. Nothing else moves the number.
4. **The distribution is the output.** Uncertainty is stated as a calibrated band, not hidden inside the randomness of a reply. The reader sees the spread directly and can act on its edges.
5. **Validated against outcomes.** Because every forecast is fixed and dated, it can be scored against what later happened, which is what calibration testing requires.

TWO WAYS TO PUT A NUMBER ON A FILE

| PROPERTY | SAMPLED GENERATIVE ANSWER | DETERMINISTIC ESTIMATOR |
|---|---|---|
| Same file asked twice | Can return different figures | Returns the identical figure |
| Uncertainty | Hidden in which reply arrived | Stated as a calibrated band |
| What moves the number | Sampling, settings, model updates | A change in the record or a recorded version change |
| Audit | Reply can be shown, not reproduced | Figure regenerated from stated inputs and versions |
| Accountability to outcomes | No fixed forecast to score | Every dated forecast can be scored when the matter resolves |

This does not make language models the wrong tool. They read documents, summarize and draft extremely well. The point is where the number comes from. A model can help read the file; the figure a decision rests on should come from a deterministic estimator trained and validated on outcomes, with its uncertainty shown as a band. WP-006 explains how to read that band, and TN-003 explains how its calibration is tested.

06

## How to test any estimator, including ours

Any platform, general counsel or funder can run the test in this paper on any system that produces a number, in under an hour, without access to its internals.

1. Write or select one file that will not change, with no real party in it.
2. Submit it with identical wording twenty times, on two different days.
3. Record every figure. Compute the minimum, median, maximum and the ratio of maximum to minimum.

4. Ask the vendor which settings and model version produced each answer, and whether any earlier answer can be regenerated exactly.
5. Ask what changes the number. The acceptable answer is: a change in the record, or a recorded version change.

A system that passes returns one figure twenty times and can tell you why. A system that does not will show its spread within the first handful of requests. For platforms building valuation into their products, WP-007 sets out how a deterministic, record-based estimate is embedded alongside the generative features users already value.

07

## The question, as the reader has it

A claims team lead, a general counsel and a funder's underwriter each ask software the same thing before they sign: if I give you this file again next week, will you give me this number again? Their problem is not that an answer might be wrong. Every estimate can be wrong. Their problem is that a number which changes on repetition cannot be checked against anything, so there is no way to learn whether it was any good.

A single number fails this reader for a second reason. It carries no statement of how far it would move if asked again. Where the system is deterministic, that spread is zero by construction and the remaining uncertainty is the legal uncertainty, which a band states. Where the system is sampled, the spread is a hidden extra layer of uncertainty stacked on the legal one, and a point estimate conceals it.

08

## What is measured

The test measures dispersion: how far apart the answers are when the input is held fixed. For each question and each configuration it records the minimum, median and maximum of the answers, the

interquartile range, the ratio of maximum to minimum, and the coefficient of variation, so that a dollar figure and a probability can be compared on one scale.

| QUANTITY | UNIT | POPULATION |
|---|---|---|
| Settlement value returned | US dollars | Repeated requests on one synthetic case summary |
| Probability of any recovery returned | Percent | Repeated requests on the same summary |
| Dispersion statistics | Spread of the above | All analysed requests in a configuration |

The population is deliberately small and artificial: one case summary written for the purpose, with no real party in it. That is a design choice. Holding the file fixed isolates the one source of variation the paper is about, which is the system and not the matter. Accuracy is a separate question that this design does not test.

09

## How a number travels in a deterministic estimator

1. **Record to structured event.** The file is read into dated, typed events and amounts, and reduced to a fixed set of inputs with a fingerprint. Entities in the file are resolved against a knowledge graph so the same party, court or judge is one node.
2. **Structured event to model class.** Duration is modelled with survival analysis, in the Kaplan-Meier and proportional-hazards tradition, because many matters are still open when a model is built and a method that ignores open matters understates duration. Event probabilities use calibrated probabilistic classification. Sparse venues are handled by hierarchical partial pooling.
3. **Model class to calibrated output.** A fitted model is a fixed function: the same fingerprint through the same version returns the same output on every request. Randomness, where it enters at all, enters during fitting and is fixed by the version, never at the moment of the request.

4. **Calibrated output to band.** Uncertainty is stated as a quantile or conformal interval, so the spread a reader sees is the legal uncertainty measured on closed matters, not an artefact of the request.

A language model can sit upstream of this path, helping to read a document into structured events. What it cannot do is set the figure, because the figure has to be a function of the record and the version and of nothing else.

10

## Validation design

Two properties are validated, separately. **Reproducibility** is validated by the repeat-request test of this paper: the same input is submitted many times, on more than one day, and the distribution of answers is summarised. For a deterministic estimator the expected result is a single value, and any departure fails. **Accuracy** is validated on a different design: models are frozen at a time cut and scored on matters filed after it, with the point-in-time boundary applied to every input so that nothing known only later can inform an earlier estimate.

The metric families for accuracy are discrimination, calibration and interval coverage. Calibration is read from reliability diagrams and from Brier-type scores, which are proper scoring rules: they reward honest probabilities and cannot be improved by shading. TN-003 shows how to read them. Figures 1 and 2, where they appear, are evidence of dispersion on a fixed input and say nothing about accuracy.

11

## Scope and precision

The test uses a synthetic file, a single model and a fixed date, and its findings are statements about dispersion in that setting. The structural conclusion is independent of those choices: a sampled

answer is a draw, and a decision needs either the distribution or a system that returns one answer. The paper does not rank vendors and makes no claim about any model's accuracy.

Criterica estimates are bands read from the court record. They are not predictions of the realized dollars of any one matter, and they do not state who prevails from disposition codes. Open matters, the lag of the record and thin venues are carried in the labels and the width of the band.

12

## What a reader can reproduce

The protocol under the heading on testing any estimator, earlier in this paper, is complete: any reader can run it on any estimator with a file of their own, without access to internals, and compute the same summary statistics. Nothing in it depends on the platform. Reproducing a Criterica figure for a real matter does require the platform, because it depends on the stated record version and model version; both accompany the figure so that it can be regenerated on request and checked to match.

### What this number means

The finding that carries is structural: a sampled generative answer is a draw from a distribution, and a decision needs either the distribution itself or a system that returns the same answer every time for the same input. The test described here measures dispersion, not a model's accuracy, and uses a synthetic file rather than any real matter.

REFERENCES


1. Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A Flaw in Human Judgment. Little, Brown Spark.
2. Kahneman, D., Rosenfeld, A. M., Gandhi, L., & Blaser, T. (2016). Noise: how to overcome the high, hidden cost of inconsistent decision making. Harvard Business Review, 94(10), 38–46.
3. Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency (2011). Supervisory Guidance on Model Risk Management (SR 11-7 / OCC 2011-12).

www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

4. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. doi.org/10.1080/01621459.1958.10501452
5. Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34(2), 187–220.
6. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
7. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. doi.org/10.1198/016214506000001437
8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330. arxiv.org/abs/1706.04599
9. Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
10. Fellegi, I. P., & Sunter, A. B. (1969). A theory of record linkage. Journal of the American Statistical Association, 64(328), 1183–1210. doi.org/10.1080/01621459.1969.10501049
11. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: a gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. arxiv.org/abs/2107.07511
12. Claessen, K., & Hughes, J. (2000). QuickCheck: a lightweight tool for random testing of Haskell programs. Proceedings of the Fifth ACM SIGPLAN International Conference on Functional Programming, 268–279. doi.org/10.1145/351240.351266

## NOTES

1. Default settings means the request set only the model, the prompt and the answer limit; temperature and reasoning were left at the vendor default, with an answer limit of 1,500 tokens.
2. Reasoning off means the model's built-in reasoning was disabled, so the reply is the answer alone; a 200-token limit was ample for it.
3. Parsing takes the first figure on the first non-empty line of each reply.

Prepared by Criterica Intelligence Research with data engineering, validation and editorial review. Criterica Intelligence Research publishes numbered documents in five series; each is versioned, dated and issued as a new version when it changes, with the change recorded.

## CITE THIS DOCUMENT

Criterica Intelligence Research. (2026). Same File, Same Number: Variance, Not Hallucination, Is the Risk (Working Paper WP-005, version 1.0). Criterica. https://critericaintelligence.com/research/same-file-same-number

```
@techreport{criterica_wp005_2026,
  author = {{Criterica Intelligence Research}},
  title = {Same File, Same Number: Variance, Not Hallucination, Is the Risk},
  institution = {Criterica},
  type = {Working Paper},
  number = {WP-005},
  year = {2026},
  month = nov,
  note = {Version 1.0},
  url = {https://critericaintelligence.com/research/same-file-same-number}
}
```

DOI: pending