

The Verification Asymmetry: On the Impossibility of Re-Checking a Probabilistic Decision

Causum

Abstract

Every mature engineering discipline rests on an asymmetry that favors the verifier: checking a result is easier than producing it. Probabilistic decision systems invert this asymmetry. A conclusion produced by a large language model cannot be re-checked at all — re-running the process may yield a different answer with nothing changed, and neither run is thereby shown right; asking the process to review itself solicits testimony from the party under review. We examine this inversion from first principles and survey the places an external correctness signal could, in principle, come from: decidable predicates, eventual outcomes, agreement across repetitions, and conformance to expectations stated in advance. Each fails or falls short in instructive ways. We conclude that the most promising and least developed candidate is the fourth: deriving, before a decision runs and from a source independent of the decision process, a constraint on where an acceptable answer must land. We do not know how to derive such constraints reliably, how to state them at the right granularity, or what discipline must govern their use. This paper argues that these unknowns constitute the central open problem of high-stakes AI deployment.

1 The Asymmetry That Built Engineering

Verification being cheaper than production is not a convenience; it is the load-bearing wall of technical civilization. A proof is hard to find and easy to check. A bridge design is costly to produce and tractable to review. A compiler’s output can be tested against the semantics of its source. In every such case, a party other than the producer can gain justified confidence, at reasonable cost, that the work is right — and the entire apparatus of professional accountability (review, audit, certification, liability) is built on that possibility.

Probabilistic decision systems break the wall. When a language model determines that a claim is valid, that a transaction is anomalous, that a treatment is indicated, the determination is a sample from a distribution. What would it even mean to check it?

2 Two Doors That Look Like Exits

2.1 Run it again

The natural first move is repetition. It fails for a reason worth stating carefully: agreement between two runs of the same process over the same inputs demonstrates *stability*, not correctness. A process can be stably wrong — systematically, confidently, and reproducibly — and the failure modes of large models are frequently of exactly this character. Repetition

estimates variance. It cannot touch bias. Voting schemes across samples [3] sharpen the estimate of the distribution’s mode; they do not certify the mode.

2.2 Ask it to check its own work

The second move is self-review: prompt the system to critique, grade, or verify its own conclusion, perhaps in a fresh context. This has become a de facto industry standard [1], and we regard the practice as structurally unsound in high-stakes settings, for a reason older than computing: *the reviewer is the party under review*. The self-critique is generated by the same weights, the same training distribution, the same blind spots. Worse, models’ articulated reasoning is known to diverge from the factors that actually determined their outputs [2]: the review is not even a reliable report of the process it claims to review. Every profession that takes decisions seriously has independently converged on the same rule — the judge may not have an interest in the verdict — and current practice violates it at scale.

We do not claim self-review is useless; it catches some errors, as any additional pass might. We claim it cannot *ground* correctness, because it imports the very reliability question it is asked to settle.

3 Where Could an External Signal Come From?

If correctness cannot be established from inside the process, it must come from outside. Working from first principles, we can enumerate only four candidate sources. We examine each.

3.1 Candidate 1: Decidable predicates

The classical answer: reduce the standard to formal conditions and check them mechanically. Where this applies, it is the gold standard, and a long tradition from program analysis to SMT solving [4] shows how far it reaches. Our own work argues for structuring decision spaces precisely to enlarge the checkable share [5].

But we must be honest about coverage. Almost nothing a real organizational standard says reduces to a decidable predicate over available structures. “Retries must be bounded” — perhaps. “Handle escalations the way our senior adjusters handle them” — no. Forcing the reduction yields a small set of trivially checkable conditions while everything that mattered is silently discarded; the expert-systems generation learned this at the cost of the paradigm [5]. A verification story that covers five percent of the standard is not five percent of a verification story; it is a misleading sense of safety wearing the costume of rigor.

3.2 Candidate 2: Outcomes

Wait, and the world will grade the decision. True — and useless at decision time, which is when governance must act. Outcome data arrives late, entangled with intervening causes, and often never (the denied claim that was wrongly denied generates no counterfactual record). An enterprise that governs by outcomes alone has agreed to learn of its failures from its victims.

3.3 Candidate 3: Agreement across independent processes

Stronger than repetition: consult *different* processes — different models, different methods — and treat concurrence as evidence. This is genuinely better; independent errors are less likely to coincide. But independence is the entire question. Models trained on overlapping corpora share failure modes in ways that are difficult to bound, and procuring genuinely independent automated judges is, at present, aspiration rather than architecture. The deeper lesson of this candidate, we think, is directional: *agreement between derivations from different sources carries evidential weight that agreement between repetitions of one source does not*. We will return to it.

3.4 Candidate 4: Conformance to expectations stated in advance

The last candidate is the least developed and, we have come to believe, the most important. Suppose that before a decision runs, one could derive — from the standard in force and the request at hand, and *not* from the decision process itself — a statement of where an acceptable determination must land: which region of the organization’s knowledge it should rest on, which factors it must have consulted, which outcomes are even admissible. Then a completed decision could be checked, mechanically and by an outsider, against an expectation that predates it and owes nothing to it.

This would not be verification of correctness — a wrong standard yields a confident expectation of the wrong thing — but it would be something no current mechanism provides: a per-decision, external, checkable signal, with the attractive property that *when the expectation and the decision agree, two independent derivations have concurred* (the durable insight of Candidate 3, achieved across sources rather than across models).

We state plainly what we do not know:

- **How to derive the expectation.** Deriving “where should this decision land” from a body of encoded knowledge and a natural- language request is itself a hard inference problem. If a model performs the derivation, have we smuggled the untrusted process back in? Under what discipline would that be acceptable?
- **How to state it at the right tightness.** An expectation that admits everything checks nothing; one that admits only the answer is the answer, and deriving it would obviate the decision. The useful object lies between, and we possess no theory of where.
- **What the expectation may do.** May it merely grade the finished decision, or may it steer the decision in flight? Steering seems obviously more useful and just as obviously more dangerous — a wrong expectation that can steer can manufacture its own confirmation. Whether there exists a discipline under which an expectation may influence a decision without contaminating the evidence of agreement is, to us, an open and fascinating question.
- **What happens on disagreement.** A decision that lands outside expectation is a signal — but of what? A bad decision, a bad standard, or a standard that is merely silent there? Distinguishing these seems to require the accountability machinery we have argued elsewhere does not yet exist [7].

4 An Observation About the Shape of Decisions

One further first-principles observation, offered without a full development because we do not have one. Verification of *content* is blocked by the asymmetry; verification of *form* may not be. A determination is typically reached by progressively excluding alternatives — candidates considered and set aside, options narrowed until one remains. The content of each exclusion may be as unverifiable as the conclusion. But the *shape* of the progression — whether alternatives were actually considered, whether exclusions were recorded rather than silent, whether the final answer is among what survived — is the kind of thing that could, in principle, be checked mechanically. Whether checks on form can be made strong enough to carry real assurance, and how a system would record its own progression honestly without doubling its cost, we leave as questions. We note only that in human institutions, process review is precisely what we fall back on when outcome review is impossible — and that this fallback is not nothing: it is most of administrative law.

5 Conclusion

The verification asymmetry is inverted for probabilistic decisions: production is cheap and checking is, at present, impossible. Repetition measures stability, not truth. Self-review asks the interested party. Of the external candidates, predicates cover too little, outcomes arrive too late, inter-process agreement founders on independence — and conformance to independently derived, advance-stated expectation remains undeveloped but, in our assessment, is the only candidate whose obstacles look like research problems rather than dead ends. We have enumerated those problems: derivation, tightness, permissible influence, and the meaning of disagreement. We are pursuing them without certainty of the destination, in the spirit obligatory for first steps. The discipline that solves them will have built something the field currently lacks entirely: a way to hold a single probabilistic decision to account at the moment it is made.

Contact:

Causum

www.causum.com

contact@causum.com

References

- [1] Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*.
- [2] Turpin, M., Michael, J., Perez, E., Bowman, S. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS*.
- [3] Wang, X., et al. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*.

- [4] de Moura, L., Bjørner, N. (2008). Z3: An Efficient SMT Solver. *TACAS*.
- [5] Causum. *Mars[®]: Structured Knowledge for High-Stakes Decision Intelligence*. Causum
- [6] Causum. *Mars[®]: Enforceability as a Property of Data*. Causum
- [7] Causum. *The Accountability Gap: Who Answers for a Single Automated Decision?*
Causum