A benchmark measures relative performance, a decision-maker must govern its provenance

A performance is not a property of an agent, but of a relation. What deserves to be certified is not a score, but the domain in which that score retains a meaning.

The figure and the question it conjures away

In June 2026, the MIRA model posts 87.8% diagnostic accuracy against 78.1% for physicians on 311 emergency clinical cases evaluated in a comparative protocol published in Nature. The result is remarkable; its scope nonetheless remains conditioned on the data regime under which it was obtained.

That same week, AMIE matches or surpasses clinicians on several indicators of management and adherence to guidelines in simulated scenarios of longitudinal disease management, evaluated against physicians under an experimental protocol published by Google Research and Nature.

These results are real, but the debate they spark often misses its target. The reflex is to ask whether these scores are high enough, robust enough, or close enough to the real world. The question is legitimate, and secondary. The prior question is simpler: what exactly does a performance score measure?

We often speak of a system's performance as a property it would possess, the way an object possesses a mass or a length. The way of speaking seems natural; it is misleading. A performance is not an intrinsic property of an agent.

A performance is a property of the relation between an agent, a data regime and a deployment context. The term relation is used here deliberately. A condition modifies a performance assumed to pre-exist; a relation asserts on the contrary that the performance exists only through the elements that constitute it.

This idea connects in particular with recent debates on the evaluation of AI systems, which show that the meaning of a metric depends on the evaluation framework that produces it (for instance: "Is Your AI Model Accurate Enough? Accuracy Requirements and the EU AI Act", 2026). Accuracy then no longer appears as an intrinsic property of the system but as a property conditioned by the evaluation framework that renders it

observable. The question is no longer only "what is the performance?" but "in what context does this performance retain a meaning?".

The idea is not specific to artificial intelligence. It is present in the sciences of distributed cognition and in ecological rationality, more broadly in any discipline that holds that a capacity cannot be understood independently of the environment in which it is exercised. A heuristic is not performant in itself, it is performant relative to a structure of the world, and a measure of performance is therefore never absolute: it is conditioned by the regime in which it was produced. The same score can then correspond to radically different realities depending on the population observed, the quality of the data, the prevalence of events or the operational conditions of deployment. The benchmark does not lie, it faithfully describes what it measures; the problem is that it describes only that.

Relational performance

Let us set down the word that will keep returning. A regime is the effective distribution of cases on which a system operates: population concerned, quality of data, frequency of events, context of use. The measurement regime is that of the benchmark, the deployment regime that of the field, the validity regime the one the supplier claims to cover. The whole question of governance plays out in the gap between these three regimes.

Robustness then appears in a different light. The dominant discourse treats it as an additional quality that comes on top of accuracy: a system would be accurate, then robust. The representation is misleading.

Robustness is not one more property, it is the measure of what survives when the regime varies.

Put more simply, so that no decision-maker drops off: a robust system is not one that stays performant in absolute terms, it is one whose performance varies little when conditions change. The shift is not verbal, it changes the very meaning of benchmarks. A benchmark ceases to be a general verdict on the quality of a model and becomes a local observation, produced in a given regime; a leaderboard ceases to be a universal ranking and becomes a partial photograph, whose scope depends on the regime in which it was established. The pertinent question is no longer "which is the best model?" but "which model stays performant when the regime changes?".

Provenance as explanatory hypothesis

Once we accept that performance is relational, a question arises at once: why does the same score behave differently when one changes regime? To answer it, I propose the concept of provenance of performance. By provenance, one must not understand a

hidden recipe that could be read off the model, but an explanatory hypothesis bearing on the dominant mechanisms that produce the observed performance.

Three mechanisms weigh particularly, and they must be situated where they act along the path of inference.

Case information acts at the input: the decision depends directly on the characteristics proper to the individual observed.

The memorised structure acts in the model: the decision relies on regularities learned in training, completeness and contamination included.

The population prior acts at inference: when individual information becomes insufficient, the prediction is completed by statistical regularities inherited from the training population.

These mechanisms are neither exclusive nor independent, and every decision mobilises several at once; the question is therefore not to determine which one is present, but which one dominates the observed behaviour. Provenance does not claim to reveal the essence of a performance, it seeks to identify the dominant mechanism that explains its behaviour when conditions change. Its justification is not ontological but pragmatic: different dominant mechanisms produce different failure modes.

Observable signature and validity regime

Provenance is a latent object, and governance therefore cannot bear directly on it: to be governable, it must produce observable consequences. I call signature the set of these consequences, which can be read in the degradation profile under ablation, the out-of-distribution behaviour, the sensitivity to missing data and the stability between laboratory and deployment.

Provenance explains, the signature is measured.

The validity regime forms the third link of the chain: the domain in which the supplier asserts that the observed performance retains its meaning. The complete relation then reads in a single stroke, from the dominant mechanisms to the observable signature, then to the claimed validity regime. Governance acts neither on the latent cause nor on the performance itself, but only on the observable elements of this chain.

The counter-fact that reveals the problem

Recent work from Mass General Brigham illustrates the distinction: twenty-one large models, submitted to twenty-nine clinical vignettes, reach more than 90% correct final diagnosis with complete data, but fail in more than 80% of cases on early differential

reasoning when information is partial (JAMA Network Open, April 2026). The authors interpret this result as a dissociation between knowledge and reasoning.

This observation does not directly demonstrate that the initial performance rested above all on memorised regularities: the link is inferred. But an assumed inference is not a weakness if it defends itself, and this one defends itself by two properties. It explains the observation simply, accounting in a single movement for both the success on the complete case and the fall on the incomplete case. And it produces testable, therefore refutable, predictions. A theory is not useful because it is certain, it is useful because it explains more with fewer hypotheses and because it can be refuted. One therefore adopts the grid not because it is proven, but because it is, to date, the most economical.

What ablation actually measures

Ablation is the most direct way to observe a signature: progressively remove information, then watch how performance evolves. Substantial gaps between performances measured in the laboratory and performances observed in deployment have been documented in numerous recent works devoted to the reliability of AI systems.

Such variability is precisely the type of signature that the proposed approach seeks to render governable. It is compatible with performances strongly dependent on the measurement conditions that produced them (Narayanan & Kapoor, March 2026).

But one must be precise about what this test actually measures. Ablation does not reveal the origin of a competence, it reveals its dependence: even excellent clinical reasoning can collapse when a critical piece of information disappears, as an expert drops off on a suspicion of pulmonary embolism if one removes the saturation and the D-dimer. It therefore does not directly designate a dominant mechanism, it produces a signature compatible with certain mechanisms rather than others. The nuance is essential: without it, the protocol becomes a claim to read the inside of the system; with it, it becomes a defensible instrument of governance, opposable in review rather than refutable in a single sentence. Eric Topol noted it differently, recalling that these results rest on clean data, text alone, and remain preliminary: the warning is correct, but it lists limitations where a signature, even fallible, gives a hold.

When information becomes scarce

One will object, in the strongest version of the objection, that all this would be settled by widening the measurement distribution: a benchmark broad enough would render provenance superfluous, since the patient benefits from the result, whether reasoned or recited, and one does not require of a human physician that he separate his intuition from his memorisation. The objection wins under one precise condition, that the deployment regime be a simple extension of the measurement regime. But it is not: the rare case is

not a frequent case in greater number, it is structurally absent from any realistic widening of the measurement set, and the average of a broader benchmark masks this extreme behaviour instead of revealing it. This is precisely where different dominant mechanisms produce different failure modes, and where the split ceases to be indifferent.

The relational thesis then produces its sharpest corollary. In data-rich regimes, the three mechanisms tend to converge toward the same decision and knowing which one dominates matters little; in poor regimes, they diverge, and this divergence becomes decisive.

The value of provenance grows as case information decreases.

The proposition is refutable, and that is what makes it something other than an intuition: if, in a poor regime, the three mechanisms still converged toward the same decision, provenance would remain indifferent and the statement would fall. The operational falsifier is therefore the measurement of this divergence, failure modes in support, and not an unobservable endogenous share. It remains to name this regime without jargon: it carries a technical label, high dimension and low sample, but the idea holds in one sentence, there are more potentially relevant variables than cases available to tell them apart. When that is the case, no individual datum suffices to decide on its own, and the population prior takes over for lack of better. A cohort of a few hundred patients carrying a rare mutation, where one measures hundreds of variables, is exactly this situation: a field such as BRAF V600E on 184 patients is an instance of it, and the reader does not even need the field to grasp the mechanism, it suffices to count the variables and the cases. On such a regime, a performance inherited from the completeness of an established dataset does not degrade, it evaporates, because there exists nothing to recall that resembles the present case. Benchmarks are born in rich regimes; the clinical uses that count, rare disease, sub-population, atypical presentation, incomplete record, live in poor regimes, and the decisive shift is therefore not the banal pair benchmark versus field, but the descent along the information curve, there where the decision is most exposed.

A general thesis on predictive systems

Medicine makes the phenomenon visible but does not exhaust it, and it would be lazy to confine the thesis to health for the comfort of the example. The same structure recurs everywhere the information on the case grows scarce: a financial model is judged above all during the crises it has almost never observed, an intelligence system facing unprecedented threats, an industrial forecasting twin in failure regimes rarely encountered, a software copilot as soon as it leaves its training repositories for an atypical codebase. In each case, the same grammar: a performance measured in a given regime, dominant mechanisms that govern its behaviour outside that regime, an observable signature that betrays this dependence. These transpositions are illustrative and not

demonstrated, they show the generality of the structure, not a measurement made in each domain. The clinic is not the object of the thesis, it is its most visible revealer, because the gap between measurement regime and use regime is there maximal, observable, and paid for by a body; elsewhere it is only paid for later.

What must be certified

The consequence is direct: what must be governed is not the observed performance, it is the conditions that render it significant. One cannot require of a supplier that he demonstrate the exact origin of each decision, which would be to demand the unobservable; one can however require that he explicitly declare the validity regime, that he document the out-of-distribution behaviour, and that he characterise the degradation profile under controlled perturbation.

This requirement is coherent with the evolution of the European framework. The EMA Reflection Paper devoted to the use of artificial intelligence in the medicinal product lifecycle already insists on the necessity of documenting the context of use, the model's assumptions, its limits and the conditions under which its performances remain valid. The regulatory consequence remains here conditional: it is not a matter of asserting that the EMA or the AI Act today impose such a reading of performance, but of observing that their documentary requirements converge toward an increasing explicitation of the validity domain of high-risk systems.

These three requirements all bear on observable or declarative elements, never on the latent.

The shift is discreet but profound. A benchmark will always remain what it is, a local measure produced on a given distribution, and the decisive question is not whether a score is high, but in what domain that score retains a meaning. To govern is not to vouch for an observed performance, it is to vouch for the conditions that render it observable, and for the admission, written in black and white, of what happens when one leaves them.

What deserves to be certified is not the performance, it is the domain in which that performance continues to have a meaning.