# Benchmark performance is not deployability: three reliability ports, not three metrics

*Three structural ruptures separate benchmark evaluation from the production regime: split, calibration, applicability domain. Deployability is installed by architecture, not by score.*

## The problem

Every other month, a new model beats state-of-the-art on a public benchmark. The press release circulates. Product teams escalate the news to the executive committee. In half the cases, the model ends up as a frozen proof of concept, or as a silently failing deployment. This is not an execution problem. It is a measurement problem: the industry keeps treating the benchmark score as proof that a model holds up in production, while the two evaluation regimes have almost nothing in common.

The gap is not a recent flaw. It is structural and it is widening. Benchmarks have become more saturated; margins between models are shrinking; leaderboards are mobilized as communication tools all the more aggressively as they discriminate less. Meanwhile, production receives queries the benchmark has never seen, under distributions that the hold-out (the fraction reserved for evaluation) has never simulated.

## Why current solutions fail

The dominant position is simple: evaluate, deploy, monitor. Add observability if needed. This reasoning rests on a rarely explicit assumption: that the benchmark score constitutes a useful measure of operational performance, modulo a degradation factor absorbed by downstream monitoring.

This assumption fails on three ruptures, already named in this week's posts.

1. The first rupture is the split. The benchmark typically uses random partitioning. Production receives its data in temporal order, with population and practice drift. A model that has learned a temporal leak never appears as such on a random split, it appears as excellent. This is precisely the subject of Friday's post: if performance collapses when moving to a temporal split, the model was learning the leak, not the task. A random split in the healthcare industry is a memory test; a temporal split is a generalization test.

2. The second rupture is calibration. The AUC score (area under the ROC curve) measures an order. It can remain very high while the probabilities produced by the model are systematically biased. In clinical use, in pharmacovigilance, in operational triage, we do not exploit an order, we exploit probabilities, whose thresholds trigger costly actions. An AUC of 0.90 without calibration is not a usable model; it is a ranking model.

3. The third rupture is the applicability domain (AD). The hold-out covers the training distribution. Production receives out-of-zone queries (a patient with a rare mutation, a structurally distant

molecule, a comorbidity absent from the training set). Without a signed AD, the model answers. It extrapolates, with no signal that it is extrapolating.

## The alternative model

The alternative is not to abandon benchmarks. It is to stop confusing them with a measure of deployability. Three reliability ports must be installed, and their installation is an architectural decision as much as an evaluative one.

1. First port: disciplined split. Temporal for data that drifts over time (epidemiology, pharmacovigilance, clinical signals). Scaffold-based for molecular data, where structural similarity creates a leak invisible to the random split. The cost of a correct split is an apparent drop in score. That is precisely the cost one seeks to pay upstream, rather than in production.
2. Second port: explicit calibration. Isotonic method or equivalent, on a set independent from the test, with a published reliability diagram. This port transforms an order score into a usable probability.
3. Third port: signed AD, verifiable at inference. An out-of-domain query receives a typed rejection, not a silent prediction. Monday's post stated the formula: *« a score without temporal split, without calibration, without AD, is not a measure of reliability, it is a trinket. »*

On PREDICARE, a pharmacological prediction platform, the three ports are not upstream validation steps, they are the structure of the pipeline. On ToxTwin, the toxicological twin, scaffold split is imposed upstream of any training, and the AD is versioned with the model. These instances do not prove the doctrine is universally good. They show it is implementable, and that its costs are measurable.

## CTO / executive committee implication

The decisional consequence is clear-cut. An AI program whose validation budget is lower than its modeling budget is a program that buys trophies and then finances the production incident. The public benchmark is a necessary condition for deployability; it is never proof of it. Wednesday's post addressed the ML community's objection: if the community is hardening its own protocols (temporal splits, scaffold splits, explicit distribution shifts) it is implicitly recognizing that the raw score did not measure what the market was making it say.

One simple question separates the two regimes at the executive committee level: *« how much does your best model lose, in performance, when you re-evaluate on the latest unseen temporal window, after calibration, restricted to the AD? »* If the number does not exist, the model is not deployable, it is eligible for a test. If the number exists and it is small, you govern your AI. If it is large, you know what needs financing. The rest (press release, leaderboard, slide) is communication, not reliability.

*[Series: Benchmark ⁄ Production / closing article]*