Jérôme Vétillard / Twingital Institute

April 2026

Public Benchmarks Have Lost the Right to Decide Alone

From Score-Based Evaluation to Validation Architecture

Abstract

Public AI benchmarks such as MMLU, HumanEval, or LMSYS Arena have long served as competence signals for comparing foundation models and guiding their pre-selection. This comparative function remains valid. However, their use as a sufficient basis for deployment decisions in regulated contexts has become methodologically fragile, for two distinct orders of reasons.

The first is temporal: the erosion of the effective independence of test sets through contamination, prolonged exposure, or adaptation to evaluation platforms; the declining discriminant power as frontier models approach ceiling performance; and the growing divergence between measured performance and observed behavior in real operational contexts.

The second is structural: dominant public benchmarks are designed to evaluate LLM-type models or close variants. Systems deployed in regulated contexts, which typically rely on heterogeneous architectures combining structured generative models, tabular models, and specialized components, lie outside the evaluated perimeter. What is measured becomes what counts. What is not measured becomes structurally invisible.

The proposed guiding distinction is that between apparent exogeneity and structural exogeneity: public benchmarks often preserve the appearance of an external perspective, even as they are progressively reabsorbed into producers' optimization loops. The response is neither the rejection of benchmarks nor their fetishization, but the restoration of independence through validation architecture. For organizations operating in regulated contexts, the selection and validation of an AI system must be grounded in an internal validation capacity that is documented, reproducible, contextualized to business risk, and structurally separated from the optimization function.

1. Introduction

The question of AI model evaluation is not a peripheral technical subject. It is a governance problem, and, in regulated environments, a problem of proof.

The evaluation paradigm that emerged alongside large language models rested on an implicit assumption: a properly administered public benchmark constitutes a sufficiently robust proxy of general competence to guide model selection. In this configuration, the decision chain was simple. Public benchmarks provided a first-order signal; pre-selection decisions, and sometimes production deployment decisions, aligned with that signal.

This chain is no longer tenable on the same terms. Not because benchmarks have become useless, but because the conditions for their decisional validity have deteriorated over time, through erosion of their independence and discriminant power; and, from the outset, through restriction of their perimeter to a single architectural family. The diagnosis is now clear: evaluation evolves more slowly than the systems it claims to discriminate, and covers only a fraction of those it claims to measure.

The thesis of this note is as follows: for organizations operating in regulated contexts (healthcare, finance, pharma, insurance, compliance, critical legal functions), public benchmarks must be repositioned as secondary signals for monitoring and pre-selection. The deployment decision must rest on an internal, contextualized, documented, and reproducible validation, structured as a function distinct from the model optimization function.

This thesis is part of a broader field developed in earlier work on agentic governance and composite architecture: the reliability of an AI system can never be reduced to the performance of its central component. It plays out in the architecture that surrounds it. What is true for the governance of an agent, the component does not govern the system, applies equally to the evaluation of a model: the score does not govern the decision.

2. Conceptual Clarifications

Five distinctions must be stabilized before proceeding.

- The first opposes contamination and memorization. Contamination denotes a breach of the evaluation protocol: benchmark data, or very close versions of it, re-enter the training or post-training environment. Memorization is a property of the model that may result: the system reproduces learned patterns without the score reflecting genuinely generalizable competence. A benchmark can be contaminated without all performance being regurgitation; conversely, apparently strong performance can mask partial memorization that is difficult to detect. This occurred during the initial training of ToxTwin and led to artificially high AUC scores resulting from training set contamination.
- The second opposes static and dynamic benchmarks. A static benchmark offers the historical advantage of reproducibility. Its defect is that it degrades once it becomes public, commented upon, incorporated into instruction sets, or exploited as an optimization target. Dynamic benchmarks such as LiveBench [8] seek to reduce this problem through periodic item renewal. They do not eliminate all biases. They shift the frontier of validity.
- The third opposes ordinal ranking and absolute validity. A leaderboard may retain comparative sorting value even when its absolute scores cease to be interpretable as guarantees of production robustness. The problem is not that every ranking has become meaningless. The problem is that converting a gap in public scores into a hypothesis of operational reliability is becoming increasingly unjustifiable.
- The fourth opposes temporal erosion and perimeter bias. A benchmark can degrade over time through contamination, saturation, or adaptation. But it can also be structurally inadequate from the outset because its measurement perimeter does not cover the architecture being evaluated. The two problems are distinct. The first is a validity problem that deteriorates. The second is a validity problem that never existed.

- The fifth distinction is central. It opposes apparent exogeneity and structural exogeneity. A public benchmark gives the appearance of an external perspective: it seems to come from a methodological outside independent of the producer. But this externality may be merely apparent once the benchmark is known, integrated, optimized, anticipated, or circumvented by the actors it is meant to measure. Structural exogeneity presupposes effective independence between the function that optimizes a system and the function that evaluates it. This is the core of this note: the problem with public benchmarks is not merely their imperfection; it is the erosion of this effective independence. A benchmark absorbed into the optimization loop no longer measures. It reflects.

3. Diagnosis: The Erosion of Predictive Validity

The counter-thesis is reasonable: despite their documented defects, public benchmarks retain ordinal value for screening, and methodological progress might suffice to preserve part of their utility. That is not the point. The point is that this value is no longer sufficient to authorize deployment.

Three mechanisms converge to explain this erosion.

- The first is contamination and internal defects. The very existence of the NIST AI 800-2 draft [3] is revealing: if benchmarks could still be treated as self-evident instruments, such a level of methodological precision would not be necessary. To this is added the case of SWE-bench Verified, whose validity as a measure of frontier coding capabilities was publicly contested by OpenAI: growing contamination, a significant number of defective tests [10]. When the producer of a frontier model itself judges the benchmark unusable, the signal is difficult to ignore.
- The second is saturation. As the best models approach the ceiling on certain evaluations, a few points of difference become harder to interpret as operationally significant distinctions. The Stanford AI Index 2026 documents this convergence at the top across several major benchmarks [1, 4]. The benchmark continues to rank, but ranks less reliably than before.
- The third is the gap between benchmark and real-world usage. The Ribeiro and Lundberg (2022) study on NLP model failures in operational contexts [12], the Liang et al. analyses on HELM dimension divergences [11], and field feedback on coding agents reveal a recurring pattern: a model can excel on a standardized benchmark and fail unexpectedly on multi-step, interactive, or strongly domain-dependent tasks. The relationship between public score and deployment reliability is too unstable to serve alone as the basis for a high-stakes decision.

The problem is therefore not that benchmarks measure imperfectly. It is that they measure an unstable mixture of genuine competence, test adaptation, contamination, and strategic optimization. The score is a signal. The decision is an architecture.

4. Perimeter Bias

The three preceding mechanisms describe an erosion over time. But there is a prior problem, one that does not derive from degradation: dominant public benchmarks evaluate only a single architectural family.

MMLU, HumanEval, and LMSYS Arena are designed to measure LLM-type models or close variants, specifically textual reasoning, generation, and programming. This perimeter is not a neutral technical choice. It reflects an industrial dynamic: competition among laboratories has structured itself around generalist foundation models, and benchmarks have aligned with that race. What is measured becomes what counts. What is not measured becomes structurally invisible.

Yet the systems deployed in regulated contexts rarely rely on an LLM alone. They mobilize heterogeneous architectures: structured generative models (CT-GAN, TVAE), tabular models (Random Forest, CatBoost, XGBoost), domain-specialized components, whose relevant properties (statistical robustness, stability under extrapolation, coherence of generated distributions, behavior on small-cohort populations) are captured by no public leaderboard.

The case is concrete. In the context of TweenMe, our digital twin generation platform mobilizing more than 25 specialized models, public benchmarks evaluate only a marginal fraction of the system. Validation necessarily relies on internal protocols, aligned with use cases, business data distributions, and operational constraints. These protocols are more demanding on certain dimensions than public benchmarks (robustness, stability, statistical coherence, behavior under extrapolation). They are also less readable, less comparable, and less mobilizable as an external credibility signal.

The tension is structural. Public benchmarks offer a common positioning language at the cost of a reduced evaluated perimeter. Internal benchmarks offer operational relevance at the cost of lesser external recognition. This tension is not resolved by choosing sides. It confirms the necessity of the two-level architecture described below: public benchmarks as common language, internal validation as the basis for decision.

Above all, this perimeter bias renders obsolete the spontaneous objection that benchmarks will improve. Even a renewed, decontaminated, methodologically impeccable public benchmark remains beside the point for a system that does not primarily rely on an LLM. An improved LLM-centric benchmark does not cease to be LLM-centric. Improvement of the protocol does not correct the inadequacy of the benchmark to evaluate a different deep neural network architecture (non-LLM).

5. From Goodhart to the Governance Problem

Once a public measure acquires strong reputational, commercial, or financial value, it becomes an optimization target. When leaderboards influence valuation, funding, recruitment, or client adoption, it becomes rational for laboratories to optimize not only their models, but also the way their models present themselves in evaluation frameworks.

The intuition is that formulated by Goodhart in 1975 [9], a statistical regularity tends to degrade when subjected to control pressure, then more broadly reformulated by Strathern (1997) [13]

in the context of audit. The measure that succeeds too well ends up measuring its own influence.

This reasoning must not be transformed into indiscriminate accusation. There is no need to assume systematic bad practices. It suffices to recognize that once a public benchmark becomes a signaling asset, there is a structural incentive to optimize for it or adapt to it. This incentive undermines its status as an independent measure for high-stakes decisions.

The decisive point is architectural, not moral. Even assuming good-faith producers, a public, known, and valued evaluation system ends up being absorbed into the optimization loop of the system it evaluates. It then loses its function as a credible external perspective. Exogeneity ceases to be structural. It persists only as appearance.

This mechanism combines with perimeter bias to produce a double blind spot: public benchmarks measure increasingly poorly what they claim to measure (temporal erosion), and do not measure at all what they do not claim to measure (perimeter bias). The conjunction of the two renders their decisional status untenable for heterogeneous systems in regulated contexts.

6. Proposal: A Three-Layer Validation Architecture

The consequence is not the abolition of benchmarks. It is their repositioning within a broader architecture where each layer has a distinct role, a defined scope, and an explicit level of authority.

- First layer: monitoring and pre-selection. Public benchmarks retain genuine utility here. They allow tracking the state of the art, identifying model families, observing progress dynamics, and assembling an initial candidate set. LiveBench [8] shows that it is possible to push the contamination frontier through periodic renewal. HELM [11] shows that a useful evaluation can be multidimensional rather than compressed into a single convenient but misleading score. What matters is not the adoption of a single tool, but the shift toward evaluations that are harder to contaminate and richer analytically. This layer orients. It does not decide.
- Second layer: contextualized internal validation. This is where the decision is made. Evaluation sets must be aligned with the domain, the risk level, actual workflows, the distribution of easy and difficult cases, expected failure modes, and robustness and reproducibility requirements. This layer must document what it measures, what it does not measure, its stability, and its domain of validity.

Concretely, this implies building internal test sets derived from real or realistic business data, including domain-specific adversarial cases, multi-step scenarios reproducing production action chains, and metrics aligned with business risk rather than generic accuracy. Validation protocols must be documented with the same level of rigor as a clinical trial protocol or a regulatory validation dossier: model version, test set version, execution conditions, acceptance thresholds, exclusion criteria.

For multi-architecture systems (those combining LLMs, tabular models, structured generative models, and specialized components) this layer becomes the only place where evaluation is meaningful. Public benchmarks, by construction LLM-centric, cover neither the robustness of

synthetic distributions, nor the stability of tabular models, nor the coherence of orchestration pipelines. Internal validation is then not a complement. It is the primary signal.

The interest of the NIST AI 800-2 draft [3] is not to impose an already-binding standard, but to make the direction visible: evaluation must become methodologically explicit, statistically documented, and interpreted with caution. The movement is underway. It will not be reversible.

- Third layer: organizational separation. The function that tunes, selects, optimizes, or integrates the model must not be the sole judge of the validity that authorizes its use. This separation does not establish perfect institutional exogeneity. It restores an exogeneity by design that is sufficient to reduce role conflicts, document arbitrations, and render deployment authorization contestable.

The logic is that of any serious control system: the one who prescribes does not validate themselves. In finance, this is the segregation of duties principle. In clinical trials, it is the separation between investigator and evaluation committee. In AI model deployment in regulated contexts, the same reasoning applies, and it is not yet integrated by the majority of organizations.

7. Contribution: Apparent Exogeneity, Structural Exogeneity

The distinction between apparent exogeneity and structural exogeneity is proposed here as an analytical framework, not as a formalized regulatory category. Its function is explanatory and architectural.

Apparent exogeneity is that of a public benchmark that appears to come from a methodologically independent outside, but whose effective independence has eroded through exposure, optimization, contamination, or absorption into training practices. Structural exogeneity presupposes effective separation (by design, by organization, by protocol) between the function that optimizes and the function that evaluates. It does not require perfect independence. It requires an independence that is architected, documented, and contestable.

This framework has a reach that extends beyond AI model evaluation. It describes a recurring problem in any high-stakes evaluation system: financial audit absorbed by advisory, credit rating absorbed by structuring, clinical evaluation absorbed by the sponsor. Each time, the mechanism is the same: the measure that acquires signaling value is progressively reabsorbed into the optimization loop of those it is meant to measure. The response, in each case, has been architectural: function separation, rotation, independence by design.

The evaluation of foundation models enters this category. And the response will be the same.

8. Articulation: From Model Governance to System Governance

This analysis extends reasoning developed in two earlier works.

In "Agentic Governance Will Not Come from the Models," the thesis was that the governability of an agentic system does not depend on the quality of the model but on the architecture within which it operates. The action-space / autonomy / reversibility triplet structures the delegation regime, independently of component performance.

The parallel is direct. What is true for the governance of an agent is equally true for the evaluation of a model: reliability is not read in the component's score. It is read in the validation architecture surrounding the deployment decision.

In "Beyond the LLM-Centric Paradigm: Composite Agentic Architecture for Digital Twins in Regulated Environments," the complementary thesis was that the language model is neither abolished nor absolutized. It is repositioned as a component among others in an architecture that surpasses it. The same operation applies to benchmarks: they are neither rejected nor fetishized. They are repositioned within a validation architecture that surpasses them.

The benchmark is neither abolished nor absolutized: it is repositioned. It is the same intellectual movement, applied to evaluation.

The perimeter bias identified in section 4 reinforces this articulation. The composite architecture relies, by design, on models that are not LLMs: tabular models, structured generative models, orchestration components. LLM-centric public benchmarks cannot evaluate these systems, not through imperfection, but by construction. Contextualized internal validation is not a methodological luxury for these architectures. It is the only option that exists.

9. Insufficiency of the Market Paradigm

The solutions currently proposed by the market (guardrails, post-deployment monitoring, RLHF, red teaming) address the problem at the component level. They improve the model or monitor its outputs. They do not build a validation architecture.

A guardrail prevents a model from producing certain outputs. It does not guarantee that the model is suited to the business domain. Monitoring observes performance metrics in production. It does not replace a validation protocol prior to deployment. Red teaming tests the model's robustness against attacks. It does not address the question of the decisional validity of the score that justified the deployment.

These tools are not bad. They operate at the wrong level. Evaluation governance is a system problem. Market solutions are still overwhelmingly component-level solutions. The same level error as for agentic governance, and therefore the same consequence: as long as a system problem is treated as a component problem, it is not solved.

10. Implementation Ground: Validation in PREDICARE and TweenMe

The PREDICARE program (territorial predictive medicine in medically underserved areas) and the TweenMe framework concretely illustrate what a validation architecture separated from the optimization function means. These are not general proofs of the proposed framework. They are grounds where the framework has become operational practice.

In the context of validating a synthetic European oncological cohort (ISPOR 2025 poster), the evaluation question arose in terms that precisely illustrate the distinction between apparent and structural exogeneity. The operational fidelity results obtained (classification metrics and survival tests) have value only because the validation protocol was structurally separated from the generation loop: the synthetic cohort was produced by a pipeline, validation was conducted

on the real cohort, according to a documented protocol, with predefined metrics, explicit interpretation criteria, and a published analysis of limits.

What was validated (the operational fidelity of the synthetic cohort for downstream tasks) was explicitly distinguished from what was not validated: general statistical indistinguishability. The result is not a score on a leaderboard. It is contextualized, reproducible evidence, limited in its domain of validity, and structurally independent of the optimization function.

This case also illustrates perimeter bias. The TweenMe pipeline mobilizes CT-GANs, Fine & Gray models, SurvTRACE, random forests. None of these components is evaluable by an LLM-centric benchmark. Internal validation is not a luxury complement here. It is the only validation that exists for this class of systems.

Similarly, in PREDICARE, the selection of models embedded in the patient digital twin was not grounded in a public benchmark. It was grounded in contextualized internal validation: test sets derived from clinical protocol data, metrics aligned with clinical risk (sensitivity on alerts, specificity on false positives generating alert fatigue), acceptance thresholds defined by the medical committee and not by the development team. Function separation is not an abstract principle. It is an operational architecture.

11. What This Changes for a CTO and for an Executive Committee

For a CTO, the consequence is an inversion of the decision hierarchy. The public benchmark ceases to be the primary support for decision. It becomes an orientation instrument for first-pass filtering. The decision migrates toward a governed internal evaluation function. This implies building an evaluation capacity (not necessarily a laboratory), but a function capable of designing contextualized test sets, documenting protocols, analyzing failures, and maintaining a validation reference frame distinct from supplier marketing materials.

For organizations operating multi-architecture systems, this capacity is not optional. It is the very condition of existence for any relevant evaluation, since public benchmarks do not cover the perimeter of the deployed system.

For an executive committee, the subject is not technical in the narrow sense. It falls under internal control and the quality of evidence mobilized to authorize a system to enter a business process. The question is no longer: which model has the best public score? It becomes: on what methodologically contestable basis have we estimated that this model can be deployed in this precise context?

Four concrete implications follow.

1. The organization must formalize a proof policy: what level of demonstration is required according to use criticality? What combination of public benchmarks, internal evaluations, adversarial tests, and human supervision conditions deployment authorization?
2. It must separate roles: the team pushing model integration does not alone control the validity conclusion.
3. It must document arbitrations in a contestable manner.
4. It must accept that evaluation governance becomes a question of decision architecture, not procurement or benchmark shopping.

The interest of this approach is that it aligns with the direction of institutional developments. NIST is pushing toward documented evaluation practices [3, 6]. The European AI Office is developing tools and methodologies to evaluate GPAI models [7]. These frameworks remain in the process of stabilization. But the direction is clear: requirements of proof, method, and contestability will harden. Organizations that wait for a finalized standard before acting will find themselves behind a movement already underway.

12. Discussion and Limits

This framework has limits that must be made explicit.

5. First limit: ordinal ranking retains value. It would be excessive to claim that a public leaderboard no longer says anything. It continues to provide useful signals for pre-selection and monitoring. The thesis is not that benchmarks are dead. It is that they can no longer decide alone.
6. Second limit: not all degradations are uniform. Not all benchmarks degrade at the same rate, not all domains are exposed in the same way, and not all uses require the same level of proof. The proposed framework targets high-stakes decisions in regulated contexts. It does not claim that screening a general-public conversational tool requires the same rigor.
7. Third limit: the notion of structural exogeneity is proposed as an analytical framework. It is not a formalized regulatory category as such. Its fertility is explanatory and architectural, as it allows naming a problem that everyone recognizes without formalizing it.
8. Fourth limit: the cost of internal validation is not addressed. Building a contextualized internal evaluation capacity has a cost (in skills to develop, in data, in time, in infrastructure). This cost is not negligible, and it can constitute a barrier for mid-sized organizations. The framework identifies what must be done. It does not claim it is free.
9. Fifth limit: the perimeter bias described in section 4, while structurally incontestable, does not affect all organizations identically. Those deploying exclusively generalist LLMs remain within the perimeter of public benchmarks, even if the problems of temporal erosion persist. The perimeter bias argument takes full force for multi-architecture systems in regulated contexts. It must not be over-generalized.
10. Sixth limit: the 2026 institutional frameworks are in the process of stabilization. NIST AI 800-2 is a draft of voluntary best practices. The European AI Office is scaling up. It is appropriate to speak of strong orientation rather than finished standard.

13. Conclusion

Public benchmarks have not ceased to be useful. They have ceased to be sufficient.

Their weakness stems from two orders of reasons. The first is temporal: contamination, saturation, and test adaptation progressively erode their predictive validity. The second is structural: designed for a single architectural family, they do not cover the heterogeneous

systems that constitute the reality of deployments in regulated contexts. An improved LLM-centric benchmark does not cease to be LLM-centric.

In both cases, the deep mechanism is the same: exogeneity becomes apparent. Benchmarks lose the methodological distance that grounded their decisional value, either because they are absorbed into the optimization loop, or because they do not measure what is actually deployed.

The problem is of the same nature as that of agentic governance: as long as reliability is sought in the component, it is not found there, because it plays out in the architecture. An excellent model evaluated by a benchmark absorbed into its own optimization loop is not a validated model. It is a well-ranked model, which is not the same thing.

For a CTO, this imposes building an internal validation function. For an executive committee, this imposes treating evaluation as a matter of decision control, not technology procurement. For the organization, this imposes a validation architecture grounded in three principles: public benchmarks as secondary monitoring signals, contextualized internal evaluation as the primary signal, and structural separation between optimization and deployment authorization.

Public benchmarks have lost the right to decide alone. This is not a critique. It is a diagnosis and a clear architectural indication.

Notes

[1] Stanford Human-Centered Artificial Intelligence. AI Index Report 2026. Stanford University, 2026.

[2] Kontorovich, Aryeh, et al. "Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation." arXiv, 2025.

[3] National Institute of Standards and Technology. AI 800-2: Automated Evaluation of AI Systems: Practices and Considerations. Initial Public Draft, January 2026.

[4] Stanford Human-Centered Artificial Intelligence. AI Index Report 2026, section "Technical Performance".

[5] The distinction between apparent exogeneity and structural exogeneity is proposed here as an analytical framework to describe the loss of effective independence of public benchmarks and the necessity of validation separated by design.

[6] National Institute of Standards and Technology. AI 800-2, sections on documentation, benchmark design, statistical analysis, and interpretation of results.

[7] European Commission, AI Office. Institutional documentation relating to GPAI models, evaluation methods, and the scaling up of AI Act application, 2025-2026.

[8] White, L., Vinitsky, E., and Sridhar, N. "LiveBench: A Contamination-Limited Benchmark for Large Language Models." arXiv preprint, 2024.

[9] Goodhart, Charles A. E. "Problems of Monetary Management: The U.K. Experience." Reserve Bank of Australia, 1975.

[10] OpenAI. "Why SWE-bench Verified No Longer Measures Frontier Coding Capabilities." 2026.

[11] Liang, Percy P., Bommasani, Rishi, Lee, Tony, et al. "Holistic Evaluation of Language Models." arXiv preprint, 2023.

[12] Ribeiro, Marco Tulio, and Lundberg, Scott. "Adaptive Testing and Debugging of NLP Models." Proceedings of the 60th Annual Meeting of the ACL, 2022.

[13] Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." European Review, vol. 5, no. 3, 1997, pp. 305-321.

Bibliography

European Commission. Regulation (EU) 2024/1689 on artificial intelligence (AI Act). Official Journal of the European Union, 2024.

Goodhart, Charles A. E. "Problems of Monetary Management: The U.K. Experience." Reserve Bank of Australia, 1975.

Hendrycks, Dan, Burns, Collin, Basart, Steven, et al. "Measuring Massive Multitask Language Understanding." arXiv preprint, 2020.

Kontorovich, Aryeh, et al. "Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation." arXiv, 2025.

Liang, Percy P., Bommasani, Rishi, Lee, Tony, et al. "Holistic Evaluation of Language Models." arXiv preprint, 2023.

National Institute of Standards and Technology. AI 800-2: Automated Evaluation of AI Systems: Practices and Considerations. Initial Public Draft, 2026.

OpenAI. "Why SWE-bench Verified No Longer Measures Frontier Coding Capabilities." 2026.

Ribeiro, Marco Tulio, and Lundberg, Scott. "Adaptive Testing and Debugging of NLP Models." Proceedings of the 60th Annual Meeting of the ACL, 2022.

Stanford Human-Centered Artificial Intelligence. AI Index Report 2026. Stanford University, 2026.

Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." European Review, vol. 5, no. 3, 1997, pp. 305-321.

White, L., Vinitsky, E., and Sridhar, N. "LiveBench: A Contamination-Limited Benchmark for Large Language Models." arXiv preprint, 2024.