When Evaluation Becomes the Infrastructure of Concentration

How the Crisis of Public Benchmarks Turns Operational Qualification into a Cumulative Asset

Public benchmarks are not disappearing. They are losing their function of industrial qualification. When evaluation requires an executory footprint that is not publicly accessible, the authority of qualification migrates toward the actors who control the actual execution of systems. Evaluation then ceases to be a mere instrument of measurement. It becomes a cumulative infrastructure of industrial power.

I. The Crisis Is Not the Score, It Is the Footprint

The phenomenon observed in May 2026 contains three distinct layers that must be separated before being articulated:

1. A metrological crisis: the public instruments for measuring AI capability no longer discriminate.
2. A cumulative industrial dynamic: post-crisis qualification concentrates in the actors who control execution.
3. An institutional transformation: the hybrid regime that replaces public metrology is, for the time being, devoid of any contradictory procedure.

The present article describes these three layers in this order. The three hard facts presented below were already signalled in Volume 2; this volume revisits them to examine their properly metrological and institutional implications.

Methodological note. The empirical examples mobilised in this text combine published primary sources, recent academic work, and sectoral intelligence consolidated in the spring of 2026. Where certain factual coordinates remain unstable or insufficiently documented in the public record, they are used here as structural signals and not as definitive quantitative proofs. The thesis defended in this article does not depend on any particular threshold or isolated case; it rests on the convergence of several dynamics now documented: benchmark saturation, vulnerability to reward hacking, and migration of qualification toward observation of actual execution.

The first register, the metrological one, rests on three documented and independent facts:

1. On 12 April 2026, the Responsible Decentralized Intelligence Lab (RDI) at the University of Berkeley publishes a short methodological note, reported in the watch under the title "Reward Hacking in Agentic Coding Benchmarks," demonstrating the breakability of eight public benchmarks of the agent layer. The central case is IQuest-Coder-V1, a submission reporting 81% success on SWE-bench. Inspection of the Git log of the test repository reveals on the order of 400 calls to external models outside the trace, undeclared dependencies (commercial tooling invoked under cover, a test harness instrumented to optimise the score rather than solve the task), and a trajectory that resembles no autonomous execution in the sense the benchmark claimed to qualify. The score remains exact. Its meaning collapses.
2. The second fact comes from the Stanford Human-Centered AI Index 2026. MMLU and MMLU-Pro are reported saturated at thresholds above 88% and 85% respectively for the principal frontier models, that is, beyond the per-item useful discrimination threshold. Runtime hallucinations, measured under real deployment conditions across four domains (health, law, programming, finance), oscillate between 22% and 94%, with no stable correlation with static benchmark performance. The public capability metric no longer discriminates.
3. The third fact comes from the Gartner Infrastructure & Operations report of 7 April 2026. Across a panel of more than 1,200 IT leaders of large enterprises, the median return on investment of AI projects in production stands at 28% over a twelve-to-eighteen-month horizon; one project in five undergoes operational collapse after deployment (production halt, rollback, abandonment before term). Gartner does not measure model capability. Gartner measures integration robustness. But the result reads as a qualification failure.

The three phenomena are distinct:

1. MMLU saturation: a problem of discriminance.
2. SWE-bench reward hacking: a problem of footprint.
3. Runtime hallucinations: a problem of execution context.

None of the three is solved by the other two. They nevertheless converge: the available public metric is no longer sufficient to rule on the qualification of an AI system for industrial use.

Two terminological precisions before going further. When I speak in what follows of *operational truth*, the term is an editorial convenience. The underlying concept is more precise: it refers to the capacity to establish an *enforceable executory qualification* of a system, endowed with four properties:

1. It is *qualifying* (it pronounces a decision of fitness),
2. *Opposable* (it can be invoked before a contractual or regulatory third party),
3. *Auditable under an executory regime* (anchored in trajectory, resources, and incidents),
4. And *insurable* (actuarially priceable).

The migration of this capacity from the public to the private is not a migration of truth in any philosophical sense. It is a migration of authority in the institutional sense.

By *evaluation jurisdiction*, I mean the bounded set of uses, constraints, versions, dependencies, thresholds, and responsibilities within which a qualification preserves its meaning.

- Two systems qualified within two different jurisdictions are not, by default, comparable.
- Two qualifications bearing on the same system within two different jurisdictions can legitimately diverge.

The fragmentation of jurisdictions is not a degeneration: it is a direct consequence of the operational precision that real uses demand.

Three facts converge. The public metric no longer discriminates. The terminal question, which must be posed at once to prepare its treatment in §IV, is no longer *who measures?* but *under what contradictory procedure can a qualification be contested?*

II. The Capacity of Qualification Migrates, and Fragility Changes Regime

The most common interpretive error consists in diagnosing a *metrological vacuum*. The diagnosis is inexact. There is no absence of metric. There is migration of the metric toward narrower, more opaque, and more interdependent jurisdictions. Private evaluations produce local qualifications under specific contractual and regulatory constraints. Metrological transfer does not produce a new universal private metric. It produces a fragmentation of partially incomparable jurisdictions.

Five objects are today confused in the public debate:

1. The *model*: abstract capability (MMLU, HumanEval); evaluated by frontier laboratories.
2. The *agent*: tooled trajectory (SWE-bench, WebArena); evaluated by frontier publishers and open communities.
3. The *system*: operational integration; qualified under sectoral regimes (AI medical devices, finance, frontier models).
4. The *organisation*: governance and return on investment; evaluated by industrial analysts.
5. The *risk*: insurability; evaluated by insurers and actuaries.

These five objects are not equivalent. The crisis of public benchmarks first affects the model and agent layers. It affects the system, organisation, and risk layers indirectly, by contagion. It is this propagation that produces the dynamic studied here.

Four basins of transfer emerge, augmented by a fifth actor of a distinct function:

1. First basin: the hyperscalers and their internal evaluation suites, operated under restricted access.
2. Second: verticalised integrators in regulated sectors.
3. Third: private industrial evaluators qualifying the organisation.
4. The fourth presents itself as a latent actor: insurers do not yet control AI metrology, but they are the natural candidates for its actuarial stabilisation. Historically, the durable regimes of operational qualification (aviation, medical, nuclear, cyber) are all anchored in insurance regimes.
5. The fifth actor, distinct from the basins, is open community evaluation (Hugging Face, EleutherAI, METR, Apollo, MLCommons). This infrastructure preserves a contestation function. It no longer suffices, on its own, to produce a productional qualification authority, and the reason deserves to be stated frontally because it constitutes the most predictable objection point.

The problem is not absence of competence; that competence is real, sometimes superior to that of proprietary laboratories. The problem is that the relevant executory footprint demands longitudinal access to productional workloads under real juridical and economic constraint, which open infrastructures do not control, will not control as long as they themselves do not become inference operators, and cannot contractualise without changing nature. Opening more benchmarks does not solve the asymmetry. The asymmetry is not documentary. It is executory.

A symmetric objection deserves to be posed frontally. Is the private less fragile than the public? No. Capitalistic conflicts of interest, internal gaming, data contaminated by other deployments, instrumentation instability across versions, opacity of qualification conditions, absence of public incident procedure: these are the fragilities of the private. One trades a type of visible fragility for a type of less visible fragility. Private fragilities are less publicly contestable. This concession forbids reading the present text as nostalgia for the public benchmark. The public benchmark was breakable; what replaces it is breakable too.

Empirical return. IQuest-Coder-V1, reformulated through the five objects, illuminates the fragmentation:

- Evaluated as a model, it scores 81% on SWE-bench.
- Evaluated as an agent under executory footprint, it does not solve the task in the productionally autonomous sense; it crosses a protocol whose footprint was not instrumented.
- Evaluated as a system integrated in a CI/CD pipeline, it introduces a risk no one has priced.
- Evaluated as an organisational programme, it consumes a transformation budget.
- Evaluated as an insurance risk, it is probably uninsurable as it stands.

Five distinct qualifications, one same object, five disjoint jurisdictions. The fragmentation is already there.

III. Those Who Evaluate Better Deploy More, Those Who Deploy More Evaluate Better

Fragmentation is not a terminal state. It is the entry into a mechanical process. We pass here from the first register (metrological) to the second (cumulative industrial). Before the seven links, the result must be named. Those who evaluate better deploy more. Those who deploy more evaluate better. This formula is the backbone of the mechanism. What follows below describes its anatomy.

And what immediately distinguishes the present thesis from a generic critique of concentration through compute must be named at once:

- Compute creates capability.
- Executory footprint creates authority.

An actor possessing compute can run a model. An actor possessing executory footprint can convince a regulated client, an insurer, an auditor, a regulator that the system can be used. In critical markets, the second power is more structuring than the first.

1. Link 1: the public benchmark is compromised (saturation, reward hacking, absence of footprint).
2. Link 2: industrial qualification can no longer rely on this benchmark and demands an evaluation under executory footprint.
3. Link 3: producing this footprint requires runtime instrumentation that measures trajectory, resources, context, and constraints.
4. Link 4: this instrumentation demands access to real workloads, traces, test environments, incident data, safety teams, and inference capital.
5. Link 5: this access is concentrated in a restricted number of actors who simultaneously control execution and instrumentation.
6. Link 6: these same actors become, by force of fact, arbiters of *production-ready* qualification.
7. Link 7: this qualification becomes a condition of purchase, insurance, compliance, financing.

The seven-link description describes a mechanism. It does not yet describe the dynamic. Here is the dynamic:

1. First turn: the more an actor controls execution, the more it accumulates incidents specific to its deployments.
2. Second turn: the more incidents it accumulates, the better the quality of its evaluation becomes.
3. Third turn: the better its evaluation, the more regulated clients are attributed to it.
4. Fourth turn: more regulated clients, more execution, more incidents. The loop closes.

Short epistemological note: The mechanism is deductive. It is not validated by empirical longitudinal observation: such validation presupposes access to private incident data whose public unavailability constitutes, precisely, the principal spring of the present text. The loop is thus presented as a strong conjecture anchored in a chain each link of which is empirically plausible. Its refutation will pass through the three falsifiability scenarios made explicit in §IV.

Consequence. Evaluation is no longer merely a control function. It becomes a returns-to-scale phenomenon. Concentration through evaluation is not an institutional side effect.

It is a cumulative asset that produces, structurally, a barrier to entry. The AI metrological crisis is not a crisis of score. It is a crisis of absence of executory footprint, in line with the RAISE framework posed in Volume 2.

Qualification, as the loop tightens, fragments into four distinct declarations that nothing requires to converge:

1. *Production-ready commercial*: a system is sellable under service contract.
2. *Production-ready operational*: it is deployable at acceptable cost, latency, and availability.
3. *Production-ready regulatory*: it can be qualified within a given sectoral regime.
4. *Production-ready insurance-grade*: it is priceable by an insurer without prohibitive premium.

Four qualifications for one same object, in four disjoint jurisdictions. The single declaration *production-ready* is a rhetorical artefact.

A concession is necessary here. Domain-specific evaluation is technically superior to generalist evaluation on the tasks it qualifies. That is true, and that is precisely why it is attractive. This superiority does not invalidate the diagnosis. It aggravates it. The more precise the qualification, the more selective the access conditions to the instrumentation that produces it. One does not contest the technical superiority of private evaluation. One contests the capacity of a public infrastructure to maintain a metrological counter-power function while that superiority installs itself.

IV. The Benchmark Can Be Closed, the Protocol Must Remain Public

We pass here to the third register (institutional). What follows is not a governance programme. It is what no public reactivation can credibly do without. It is an anatomy.

1. First clarification. When I speak in what follows of *public protocol*, the word *public* does not mean *open in the sense of published dataset*. Public means: opposable, contestable, versioned, and revisable through a known procedure. Publishing the test set is not the object pursued. The object pursued is to publish the rules under which an evaluator exercises its authority.
2. Second distinction. The benchmark and the protocol are not the same object. The benchmark is the measurement instrument; the protocol is the convention that makes the instrument readable. The benchmark, insofar as it is breakable, can and must even remain private to resist gaming. MLCommons, the cybersecurity industry (FIPS 140-3 under the auspices of NIST, United States; Common Criteria

under ISO/IEC 15408), and medical (FDA QMSR) have all ended up admitting this discipline. The protocol, on the other hand, must remain public in the sense defined above.

The minimal public protocol reduces to three verbs:

1. *To declare*: the evaluator declares the effective executory footprint (trajectory, tools, resources, context), the classes of use evaluated, the constraints retained, and the version regime under which the qualification is pronounced.
2. *To compare*: the evaluator publishes rules of partial reproducibility allowing a third party to replicate the procedure, explicit non-comparability thresholds that forbid undue aggregations, and rules of interoperability with other jurisdictions.
3. *To revoke*: the evaluator publishes its incidentology obligations (declaration, analysis, anonymised publication) and the conditions for withdrawal of metrological authority, that is, the public grounds on which the evaluator itself loses its capacity to qualify. An irrevocable qualification is not a qualification. It is a decoration.

Declare, compare, revoke: these are the three minimal movements of a public protocol in the institutional sense. They do not say what a system is. They say how the evaluator produces, articulates, and retracts its judgement.

The three verbs do not say what the public can contest. Without a procedure of contestation, one has reconstructed a public metrology without contradictory procedure. In other words: a court without appeal.

Six elements must be publicly contestable, in addition to the three verbs:

1. First, the *qualification perimeter*: if an evaluator declares that a system is qualified for a given jurisdiction, the public must be able to contest whether that jurisdiction is defined so as to exclude real operational conditions.
2. Second, the *declared footprint*: the public must be able to contest if the declaration omits dependencies or hidden costs.
3. Third, the *claimed comparability*: the public must be able to contest whether the non-comparability thresholds have been respected.
4. Fourth, the *maintenance of authority after incident*: if a qualified system experiences a post-deployment incident, the public must be able to contest whether the evaluator maintains its qualification without a re-qualification procedure.
5. Fifth, the *change of version*: if a qualified system is updated, the public must be able to contest whether the prior qualification remains valid.

6. Sixth, the *evaluator's conflict of interest*: if the evaluator has an economic, contractual, or capitalistic dependency on the evaluated system, the public must be able to contest the qualification on that ground.

A protocol that defines qualification rules without defining the six contestable elements remains an instrument of the evaluator. A protocol that defines both together becomes a procedure of public contestability. The conceptual displacement announced at the chapeau reaches here its accomplished formulation: the true question is no longer *who measures?* but *under what contradictory procedure can a qualification be contested?* All of the institutional contribution of the present analysis lies in this displacement.

Three partial precedents deserve precise mention:

1. FIPS 140-3 (NIST, United States) separates cryptographic modules into publicly defined levels, validated by accredited and private laboratories.
2. Common Criteria (ISO/IEC 15408) separates the protection profile (public, contestable) from the security target (semi-private).
3. FDA QMSR separates intended use (public, declarative, contestable) from technical documentation (semi-private).

None of these regimes has escaped the breakability of its instruments. But all dispose of public procedures of contestation:

1. FIPS 140-3 allows the revocation of a certificate.
2. Common Criteria allows the contestation of a protection profile.
3. FDA QMSR allows the challenge of an intended-use qualification following an incident.

Without a procedure of appeal, none of these regimes would have survived three decades.

What remains is to state what would invalidate the thesis. Doctrinal discipline requires making it explicit. The thesis would be seriously weakened if one of the three following scenarios materialised within the next three to five years:

1. First: if open consortia managed to mutualise, under shared governance, multi-industrial workloads with their incident traces, under juridical conditions equivalent to those of private operators.
2. Second: if regulators (AI Act, AISI, CAISI and equivalents) imposed obligations of standardised interoperable executory footprint, opposable to the private, with a public procedure of contestation.

3. Third: if insurance regimes accepted, as primary inputs, third-party open evaluations rather than proprietary evaluations under confidentiality agreement.

None of these scenarios is excluded. None seems likely in the near horizon. The thesis is not a prophecy. It describes a structural movement under conditions of continuity, and it names the conditions that would break it.

V. Functional Quadripartition, Insurers as Terminal Benchmark, and Historical Closing

As the entry into force of the AI Act obligations on general-purpose models draws nearer, as CAISI and the UK AI Safety Institute consolidate their position, and as hyperscalers deploy orders of magnitude of capital unprecedented for 2026, the metrological object hardens into industrial infrastructure.

V.a. Functional Quadripartition of the Hybrid Regime

The terminal scenario is probably not the complete privatisation of metrology. It is a hybrid asymmetric regime articulating four distinct functions, sometimes carried by the same actors. The *execution operator* produces the footprint, because it controls workloads and incidents. The *regulator* produces the protocol, because it alone can impose an opposable convention. The *insurer* produces the pricing, because it alone can transform risk into premium. The *infrastructure provider* produces the execution environment, because it controls the material and software layer on which the other functions operate. These four functions can coincide within a single industrial group or distribute themselves across distinct actors. It is precisely this distribution that determines the quality of the consolidated metrological authority.

The strategic question is therefore not: will the hyperscalers capture everything? The more serious question is: *which functions must remain separable, opposable, and contestable?* If the execution operator is also the one that produces the footprint, qualifies the system, controls dependencies, bills the inference, and defines comparability conditions, qualification becomes a proprietary instrument. Conversely, if the public protocol imposes declaration, comparability, revocation, and contestability, private evaluation can remain useful without becoming sovereign.

V.b. The Insurer as Terminal Benchmark

The insurer deserves a place of its own in this anatomy, more structuring than is often thought. In the mature technical regimes (aviation, health, nuclear, cybersecurity), operational qualification eventually meets insurance. The insurer does not only say whether a risk exists. It transforms uncertainty into price. A system can be technically

impressive, not forbidden by the regulator, and desired by the client. If the risk cannot be covered, or only at a prohibitive premium, real qualification collapses.

The triad is clear. The benchmark says that the system performs. The protocol says under what conditions that performance is qualified. The insurer says at what price the risk can be carried. The ultimate form of *production-ready* might not be the benchmark. It might be the acceptable premium.

V.c. Closing

Those who evaluate better deploy more. Those who deploy more evaluate better. Metrology ceases to be a neutral instrument. It becomes a barrier to entry.

The central proposition, now explicit, is as follows: *when qualification demands an executory footprint that is not publicly accessible, evaluation ceases to be a descriptive instrument and becomes a cumulative infrastructure of industrial power*. This proposition ties together the three registers that structure the article. It reformulates the initial question in its accomplished institutional version: the true question is no longer *who measures?* but *under what contradictory procedure can a qualification be contested?*

The benchmark ranked the models. The footprint qualifies the systems. The one who controls the footprint progressively controls access to the regulated market. Without a contradictory procedure, evaluation becomes a private infrastructure of authority.

The previous volumes of this tetralogy have described the absence of a protocol to arbitrate material allocation, then the absence of a procedure to promote MCP artefacts. The capacity of operational qualification has now become a cumulative asset. It is probably the least thematised point of the current public AI debate, and one of the most structuring.

The history of critical human infrastructures presents, without determinism, a recognisable regularity: certain structures of concentration reappear when qualification and execution become coupled. To invent an instrument of measurement to objectify a market, and to discover that whoever controls the instrument ends up controlling the market it was measuring: this movement is not a fatality. It is, at this stage, a structural regularity under conditions, conditions that have been named.

Volume 4 of the AI Governance Tetralogy. Previous volumes: "Energy as a Governance Constraint" (vol. 1), "Allocating the AI Kilowatt-Hour" (vol. 2), "The MCP Vulnerability as a Pure Case of the Promotion Port" (vol. 3).

Jérôme Vétillard, VP R&D + Engineering / Chief Product Officer TweenMe® by Qualees®