The Current Landscape for AI Evaluations And Where We Need to Land

By: Mustafa Lonandwala

In 2024, Meta reported that one of its models scored 5% on a set of cyber capability tests. When outside researchers retested the same model with better prompting and scaffolding, the score jumped to 100%. The model hadn't been updated. The difference was entirely in how the test was run.

Meta's safety conclusion was based on that first number. And that number turned out to be an artifact of poor methodology, not a real measurement of what the model could do.

This kind of problem isn't unique to Meta. Every major lab publishes eval reports. Every major lab makes safety claims based on those reports. But when independent analyst Zach Stein-Perlman reviewed the eval reports published by OpenAI, Google DeepMind, and Anthropic in 2025, he found that the reports mostly don't support the safety claims labs make from them. Labs report numbers without defining what thresholds would indicate danger, don't compare results to human baselines, and in some cases use testing methods so poor they systematically underestimate what their models can actually do.

Evaluations, or "evals," are the only systematic tool we have for knowing what we're deploying. If they're not trustworthy, the safety claims built on top of them aren't either.

How AI Evaluations Work

At its most basic, an eval is a test. You give an AI system a task, it produces an output, and some grading logic scores that output. Do this across hundreds or thousands of tasks and you get a picture of what the model can and can't do. That picture is what labs use to make safety claims, set deployment thresholds, and tell the public their models are ready.

There are three broad categories of evals in use today:

- **Capability evals** test what a model can do — coding, reasoning, factual recall. Benchmarks like MMLU, which tests general knowledge across 57 subjects, and HumanEval, which tests coding ability, fall here. These are the most common and most developed.
- **Safety evals** are a subset focused specifically on dangerous capabilities: can the model provide meaningful assistance to someone trying to build a bioweapon, conduct a cyberattack, or replicate itself autonomously? Anthropic's Responsible Scaling Policy uses evals of this kind to determine whether a model crosses a threshold requiring stricter safeguards.
- **Alignment and behavioral evals** test not what a model can do but how it behaves — does it deceive, does it pursue goals it wasn't given, does it try to avoid being shut down. These are the least developed of the three and the least standardized.

Evals are also graded in different ways. Code-based graders are fast and objective but brittle — they can't handle valid answers that don't match the expected pattern exactly. Model-based graders are more flexible but expensive and non-deterministic. Human graders are the most reliable but also the slowest and most expensive to run at scale. In practice most evals combine all three.

Most evals are what researchers call “performance-oriented”. A model gets a fixed set of tasks, produces outputs, and receives a score. But a score on a fixed test and an underlying capability are not the same thing. A model can perform well on a benchmark without actually having the ability the benchmark is supposed to measure. Researcher John Burden at Cambridge puts it plainly: much of AI development is currently building systems that are very good at a game called Chmess — a chess variant with slightly different rules — and then claiming those systems are good at Chess. The benchmarks look similar enough to the real thing that the gap is easy to miss.

What we actually need from evals is prediction. Not how a model scores on a test it was designed around, but how it will behave in situations it hasn't seen before. That gap between what evals currently measure and what we need them to measure is where most of the problems in AI safety evaluation live.

Who’s Running Evals?

There are two groups running AI evaluations right now, and they're doing very different things with them.

Frontier labs use evals as internal release gates. Before a model ships, it gets tested against a set of predefined thresholds. Clear the thresholds, and the model is approved for deployment. Anthropic's Responsible Scaling Policy is the most publicly documented version of this — evals determine whether a model reaches an ASL level that triggers stricter safety requirements. OpenAI and Google DeepMind have equivalent internal frameworks. The key thing to understand about this system is who's running it: the same organizations building the models are also designing the tests, setting the thresholds, and interpreting the results. There's no external check built into the process.

Enterprises and other organizations are in a different position. A hospital, a law firm, a school district deploying AI into real workflows is largely relying on whatever general capability benchmarks already exist. In practice this means benchmarks like MMLU, which tests broad knowledge, or coding benchmarks like HumanEval. These were built to compare models against each other in lab conditions. They were not built to answer the question an organization actually needs answered: will this model behave reliably in our specific context, with our users, under our conditions?

Most enterprises are not running:

- Sector-specific benchmarks tailored to their domain
- Alignment or behavioral evals that test how the model acts, not just what it knows
- Any longitudinal evaluation of how model behavior changes over time in deployment

The result is that deployment decisions across industries are largely being made on the basis of general capability scores. Whether a model is actually safe or reliable for a specific high-stakes use case is a question most organizations don't have the tools to answer.

Why Evals Are Failing

The problems with AI evaluation aren't new. Researchers were raising many of these same concerns in 2016. Nearly a decade later, capability has advanced significantly. Evaluation rigor hasn't kept pace. What follows are four specific failure modes.

Teaching to the Test

When a measure becomes a target, it stops being a good measure. This is Goodhart's Law, and it describes what has happened to AI benchmarks fairly precisely.

Labs train models to perform well on benchmarks. Scores go up. But as MIT Technology Review reported in 2025, improved benchmark scores don't reliably indicate improved underlying ability. The industry has learned to optimize for the test rather than the capability the test is supposed to measure. Data contamination makes this worse — models may have already encountered benchmark questions during training, meaning high scores can partly reflect memorization rather than genuine reasoning.

The result is saturation. SuperGLUE, once a meaningful measure of language understanding, now sits above 90% accuracy across frontier models. The benchmark no longer tells you anything useful. The Brookings Institution flagged the same dynamic in 2025: as AI systems become more capable and general, existing benchmarks become inadequate faster, and new ones struggle to keep pace.

The Elicitation Problem

How you test a model determines what you find.

When Anthropic evaluated Claude Sonnet 3.6 on a subset of RE-Bench, an AI research and development capability benchmark, it scored 0.21. When METR evaluated the exact same model on the same subset, it scored 0.51 — more than double. The improvement Anthropic measured going from Sonnet 3.6 to Sonnet 3.7 was smaller than the gap between those two methodologies applied to the same model. DeepMind ran into a related problem: it gave its model a single attempt on an AI R&D evaluation and reported a score of around 0.15. When it later recalculated allowing multiple attempts — closer to how models are actually used — the score rose to approximately 0.72.

These aren't edge cases. They're documented in the labs' own published materials. And they raise a direct question: if the same model produces scores that vary this much based on how the test is run, how much confidence should a reported eval score actually carry?

Grading Your Own Work

Labs design their own evals, set their own thresholds, run their own tests, and publish their own summaries. There is no external check built into this process.

The 2025 eval reports from the major frontier labs illustrate the problem. On OpenAI's bioweapons evaluation, o3 performed well enough that OpenAI noted the benchmark was approaching saturation — and on remaining tests matched or outperformed the expert human baseline. OpenAI then concluded the model didn't have dangerous bio capabilities. No explanation was offered for how those results supported that conclusion. DeepMind's cyber evaluation reported results without defining what threshold would indicate danger or what evidence would change their assessment. These aren't minor omissions. They're the difference between a safety claim and a documented safety argument.

The structural issue here is straightforward. A lab has financial incentives to ship its model. The same lab is evaluating whether the model is safe to ship. Even with good intentions, that arrangement produces pressure on the conclusions. And when the published reports don't show the reasoning, there's no way for anyone outside to push back.

The Lab-to-Life Gap

Even well-designed evals run under controlled conditions may not predict what a model does once it's deployed. This isn't hypothetical.

Epic, one of the largest healthcare software companies in the United States, built a sepsis prediction model deployed across hospitals nationwide. Internally, it reported an area under the curve of 0.76 to 0.83 — indicating reasonable predictive performance. When researchers at the University of Michigan externally validated the model across 38,455 hospitalizations, the real-world AUC dropped to 0.63. More significantly: despite generating alerts on 18% of all patients, the model missed sepsis in 67% of patients who actually had it. The model was live, in hospitals, influencing clinical decisions, on the basis of internal numbers that didn't hold outside the conditions in which they were produced.

This isn't a generative AI example, but the pattern is the same one the CIRCLE framework identified in 2026 across AI systems more broadly: current evaluation practice treats real-world variability as noise to be eliminated rather than as signal worth measuring. Lab conditions optimize for control. Deployment is where control disappears.

For agentic AI systems the problem compounds further. The Brookings Institution noted in 2025 that agents operating autonomously over long time horizons produce behaviors that no point-in-time benchmark can anticipate. A benchmark is always a snapshot but deployment on the other hand is a process.

Each of these failure modes is a real problem on its own. Together they describe an evaluation system carrying weight it wasn't designed to bear. Benchmarks are being optimized against. Testing methodology produces wildly different results for the same model. Labs publish safety conclusions without showing the reasoning behind them. And the conditions under which models get tested don't reflect how they actually get used. Right now, AI safety governance is built on top of all of this.

What Better Looks Like

The problems in AI evaluation are well-documented at this point. What's less discussed is that many of the solutions are also known. They're not simple to implement, but they exist. Here's what serious progress actually looks like.

Independent Evaluation

The most direct response to the self-evaluation problem is moving evaluation outside the labs. METR, an independent nonprofit, already does this — it evaluates frontier models for dangerous capabilities before deployment, using its own methodology rather than the lab's. The UK and US AI Safety Institutes are the policy-level version of the same idea.

But independent evaluation only works if it has real access and real teeth. An external evaluator that receives a model after deployment decisions have already been made, uses whatever methodology the lab prefers, and publishes a summary without showing its reasoning isn't solving the problem. What's actually needed is pre-deployment access, standardized testing methodology, and published reasoning — not just published scores. The Brookings Institution made a further point worth noting: grant funding alone won't sustain a healthy ecosystem of independent evaluators. This needs stable, long-term funding infrastructure.

Capability-Oriented Evaluation

Most current evals measure performance on fixed tasks. What we actually need to measure is something different — the underlying properties that predict how a model will behave in situations it hasn't seen before. A model that scores well on a benchmark it was optimized for isn't necessarily capable in any meaningful sense. It might just be good at that specific test.

The tools for building better evals already exist. Psychometrics and cognitive psychology have spent decades developing methods for measuring latent properties in intelligent systems. AI evaluation doesn't need to start from scratch. It needs to borrow more deliberately from those fields and build tests that actually predict behavior rather than just rank models against each other.

Real-World and Longitudinal Evaluation

Lab evals are snapshots but deployment is ongoing. Those are two different things and they need two different kinds of evaluation infrastructure.

The CIRCLE framework makes the case for treating real-world context as the primary signal of how a system actually behaves. That means field testing with real users, behavioral monitoring post-deployment, and longitudinal tracking of how model behavior shifts over time as conditions change.

The Epic sepsis model is a useful reference point here. External validation before wide hospital deployment would have caught the gap between internal and real-world performance before it became a

clinical problem. The same principle applies to generative AI. Evaluation that stops at the lab door is only telling part of the story.

The Policy Dimension

These three directions are technically achievable. The harder question is whether there's enough institutional pressure to actually require them. Voluntary frameworks depend on the goodwill of the organizations they're asking to comply. That's a fragile foundation for something as consequential as frontier AI safety.

The EU AI Act's conformity assessment requirements are a step toward making some of this mandatory. Burden's argument goes further: developers above a certain capability threshold should be required to make binding commitments to abide by the results of independent evaluations. Without that kind of structural requirement, the incentive to treat evals as a checkbox rather than a safety gate remains.

The Stakes

Evals are the only systematic tool we have for knowing what we're deploying. Every safety claim a lab makes, every deployment decision an organization takes, every regulatory framework being built around frontier AI — all of it rests on the assumption that evaluations are telling us something real about these systems. The evidence in this piece suggests that assumption deserves more scrutiny than it's currently getting.

Independent evaluation with genuine pre-deployment access. Elicitation standards that don't let methodology determine the conclusion. Longitudinal monitoring that follows models into deployment rather than stopping at the lab door. Transparency about what thresholds mean and what evidence would change a safety determination. These are tractable problems.

What's harder is the institutional question. Right now, the organizations with the most incentive to ship models quickly are also the ones deciding whether those models are safe to ship. Voluntary commitments exist, but voluntary commitments bend under competitive pressure. The gap between what current evals actually measure and what we need them to measure isn't a technical accident. It's partly a product of a system where the costs of inadequate evaluation are diffuse and the benefits of moving fast are immediate.

Sources

[1] How can we best evaluate agentic AI? | Brookings
[2] Evaluation of General-Purpose Artificial Intelligence: Why, What & How
[3] Evaluating AI Evaluation: Perils and Prospects
[4] AI companies' eval reports mostly don't support their claims — LessWrong
[5] CIRCLE: A Framework for Evaluating AI from a Real-World Lens
[6] The Epic Sepsis Model Falls Short—The Importance of External Validation | Critical Care Medicine
[7] Anthropic's Responsible Scaling Policy

[8] EU AI Act
[9] Guidelines for capability elicitation - METR