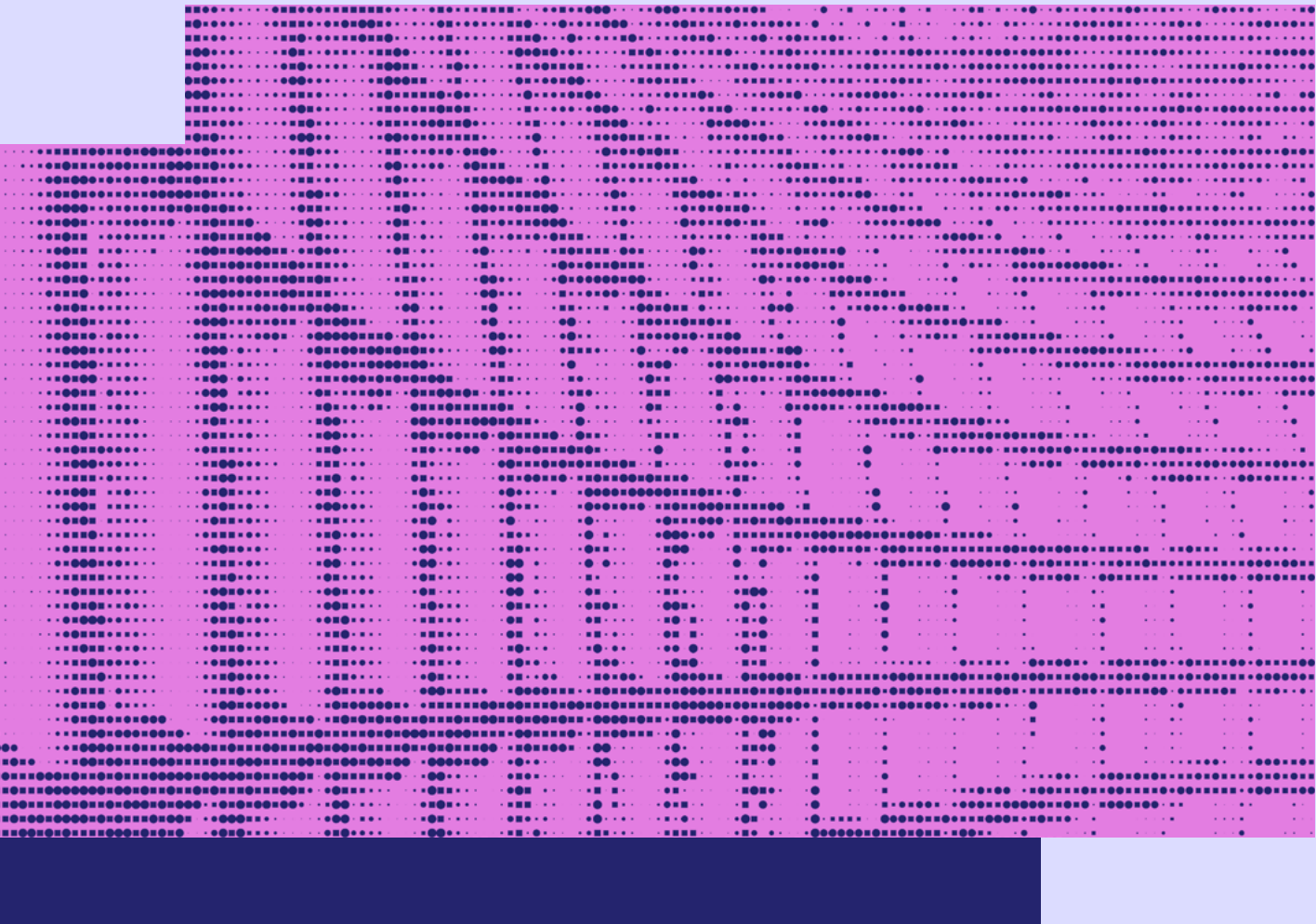


Intro to AI evaluations



AI evaluations explained

Building or buying an AI model is the easy part. The real challenge is making sure it actually works in the day-to-day operations of your business. That's where many enterprise projects falter. Tools that look impressive in a demo often break down when exposed to messy data, complex workflows, or regulatory scrutiny. The result is familiar to many executives: promising pilots that never scale, projects that stall in compliance review, or systems employees simply don't adopt.

Too often, companies judge AI systems against the wrong standards. Vendors showcase performance on public "leaderboards" or highlight benchmark scores that may have little to do with your business. Internally, teams rely on gut feel or limited testing. Neither approach gives leaders confidence that the model will hold up under the pressures of real production use.

That's where evaluations, or "evals", come in. Think of them as quality control for AI. Just as cars are crash-tested and financial systems are audited, AI systems need structured evaluations before they're trusted with high-stakes business processes.

Evaluations give you answers to the questions that matter most at the executive level:

- Can we trust this system with sensitive data and compliance requirements?
- Will it actually save costs, improve accuracy, or increase throughput?
- How will we know it's getting better, not worse, as we use it?
- Will our customers and employees adopt it or avoid it?

“

Think of evaluations as quality control for AI. Just as cars are crash-tested and financial systems are audited, AI systems need structured evaluations before they're trusted with high-stakes business processes.”

An evaluation is a structured way of measuring whether an AI system performs as intended in the context where it will actually be used. It goes beyond academic benchmarks and focuses on the specific tasks, data, and conditions relevant to your business. For example, if an enterprise is deploying AI to process insurance claims, an evaluation would test whether the model can interpret claim forms accurately, apply business rules consistently, and meet compliance requirements.

In practice, evaluations provide decision-makers with evidence that an AI system is reliable, safe, and aligned with organizational goals. They bridge the gap between abstract performance scores and operational confidence in production.

A BRIEF HISTORY OF BENCHMARKS

1990s - 2000s



Hello world

Benchmarks were first developed in academia as a way to measure progress on narrow, well-defined AI tasks. Classic examples include MNIST (handwritten digit recognition, 1998) and ImageNet (image classification, 2009). These datasets became the gold standard because they gave researchers a shared test set and leaderboard for comparing algorithms.

2010s



Rise of leaderboards

With ImageNet, benchmarks became competitive. Annual competitions like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) created a public scoreboard for AI research. Similar benchmarks followed in natural language processing, such as SQuAD (Stanford Question Answering Dataset, 2016) and GLUE (General Language Understanding Evaluation, 2018). Success on these benchmarks often defined the reputation of new models and labs.

LATE 2010s - 2020s



From narrow to broad


As large language models (LLMs) emerged, researchers built benchmarks to test across multiple domains, e.g. SuperGLUE (2019) and MMLU (2021), which measure performance across dozens of subjects from history to math. These became a shorthand for general intelligence.

TODAY



The benchmark crisis

Benchmarks are now showing their limits. Many tasks are effectively “solved” — top models all score above 90%. Datasets have leaked online, contaminating training data. And most importantly: excelling at benchmarks doesn’t predict performance in messy, real-world enterprise contexts. This has led to what some call “benchmaxxing” — chasing higher scores for bragging rights rather than meaningful capability.



Why standard benchmarks and evaluation frameworks miss the mark

New models are released and updated at breakneck speed, dropping seemingly every week. Each release comes with impressive benchmark scores, claiming superiority over its predecessors. On paper, they're getting better each time, but in execution, things are more complicated.

What are these evaluations and benchmarking frameworks testing for? What makes a model or use case “good?” It really depends on what you measure, how rigorously you test it, and when in the model's lifecycle the evaluation happens.

Benchmarks were initially developed to provide a common yardstick for measuring AI capabilities. But as the technology has evolved and business applications have become more sophisticated, these benchmarks have revealed significant limitations, particularly in enterprise contexts. Even where the tests are rigorous, accurate, and comprehensive, they often ask the wrong questions and are insufficiently targeted.

Benchmarks assess how the model performs on standardized tests in research or academic settings. They provide a single score that allows models to be compared against one another on tasks like math, coding, or reading comprehension. They are useful for spotting broad improvements across the industry, but they often measure skills far removed from enterprise use cases. Because benchmarks are public and widely known, many models are trained to “ace the test,” which can inflate scores without reflecting real-world performance. Benchmarks were never designed for business deployment — they were designed to measure research progress in controlled conditions. They are still useful as rough comparisons across models, but they should not be mistaken for indicators of whether a model will succeed in your enterprise environment.

“A lot of these benchmarks are extremely academic. But do they relate to my business? How do we make sure that what we put into production, what we put in front of our users, really delivers the value they expect?” asked Alexius Wronka, CTO of Data and Growth at Invisible Technologies.

“

A lot of these benchmarks are extremely academic. But do they relate to my business? How do we make sure that what we put into production, what we put in front of our users, really delivers the value they expect?”



Alexius Wronka

CTO of Data and Growth
Invisible Technologies

Referencing the MMLU (Massive Multitask Language Understanding), a common industry benchmark, Lydia Andresen, Invisible Technologies, Executive Director of Applied AI, said, “This is very widely used before launching foundation models and is widely respected in the academic community. It covers 57 different subjects across STEM, humanities, and other disciplines. But for our clients to make models real for their users, only a small subset of what’s measured in this benchmark is relevant to their organization. Furthermore, we see a ton of things they need to measure that aren’t in the benchmarks at all, or are underrepresented.”

In the following pages, we will discuss the limitations of current benchmark standards and evaluation frameworks and provide insights into more specific benchmarking using our own client use cases.

General limitations of standard benchmarks

- **Irrelevant scenarios:** Unless you need your AI model to play chess or participate in math competitions, benchmarks that measure according to model performance in these areas aren't going to tell you much.
- **Optimized for the test, not real-world use:** Knowing their model must pass muster, lots of developers teach to the test, similar to how students can grind for standardized tests without improving their actual critical thinking skills.
- **Large language models are nearing perfect scores on standard tests:** On popular tests like SuperGLUE, models have already reached or surpassed 90% accuracy, making further gains feel more like statistical noise than meaningful improvement.
- **Data contamination:** Many models may already have seen benchmark questions during training, making results unreliable. Worse, the shelf life of new assessments is short: once released, tests quickly leak online and seep into future training data, turning fresh benchmarks into memorized trivia.
- **Clean vs. real-world data:** Benchmarks typically test models on "clean" lab-grown datasets, which don't reflect the messy reality. In actual deployment, models encounter inputs riddled with human errors, from typos to biases.

Enterprise-specific challenges

- **Organization-specific blind spots:** Off-the-shelf models miss enterprise realities, like a customer service bot that can chat fluently yet fails to follow your company's refund policy.
- **Data security concerns:** Organizations are concerned about training on their proprietary data and putting that data at risk.
- **Rapid obsolescence:** Models are improving so quickly that evaluation frameworks become outdated shortly after they're established.
- **Balance between human alignment and determinism:** Enterprises need evaluation frameworks that are both aligned with human users and deterministic to prevent model drift. However, creating human-aligned datasets at scale is expensive and challenging.
- **Emerging use cases:** Many enterprise AI applications are novel, with no established benchmarks against which to evaluate their performance.
- **No repeatable evaluation framework:** Benchmarks don't provide a sustainable system for assessing models over time or across expanding use cases. Each evaluation requires starting from scratch with new benchmarks, making continuous improvement difficult.
- **Proprietary data challenges:** Enterprises that train on proprietary data require custom benchmarks, creating a resource-intensive process that grows exponentially with each capability being assessed, especially when adapting to new regulations or changing business needs.

This fundamental disconnect between standard benchmarks and enterprise needs has led to significant challenges in AI adoption, with many projects failing to move beyond the proof-of-concept stage.

The cost of doing nothing

For many enterprises, the most common failure mode isn't a scandal or regulatory fine — it's never deploying the model at all. Projects stall in endless pilots, eating capital, staff time, and executive attention with little to show for it. MIT research suggests that as many as 95% of generative AI pilots fail to make it into production², leaving organizations stuck in proof-of-concept limbo. But if models do make it into production without proper evaluations, the risks escalate quickly.



Reputation Errors, hallucinations, or biased outputs can erode customer trust. In a climate where security and accuracy are paramount, a single AI misstep can spark backlash, attrition, and long-term brand damage.



Compliance Regulated industries require AI to meet strict standards. Evaluations ensure models are audit-ready and aligned with laws, protecting against costly penalties.



Fairness Unchecked models can produce discriminatory outcomes in lending, hiring, or healthcare, leading to lawsuits and sanctions. Evaluations surface these risks before deployment.



Reliability Mission-critical applications need AI that performs consistently under real-world conditions, not just in training. Evaluations confirm robustness across accuracy, consistency, and resilience.



Transparency Evaluations also explain why models make certain decisions, building stakeholder trust and meeting growing requirements for explainability.

² MIT Report: 95% of generative AI pilots at companies are failing (Fortune)

Get started with custom AI evaluations

As the AI landscape continues to evolve at a rapid pace, enterprises must adapt their evaluation approaches to ensure they select and implement the right models for their specific needs. Standard benchmarks, while valuable for general comparison, fall short when it comes to assessing real-world performance in enterprise contexts.

Custom evaluation frameworks offer a more reliable and insightful approach, providing organizations with the information they need to make informed decisions about AI deployment and fine-tuning. By focusing on the specific requirements, constraints, and objectives of your use case, you can build evaluation methodologies that truly reflect what matters to your business.

By embracing custom evaluations, enterprises can transcend the limitations of standard benchmarks and develop AI systems that deliver genuine value, align with their unique needs, and uphold the highest standards of safety and compliance.



Take advantage of AI opportunities now

The path to success with AI isn't just building models — it's proving they work where it matters. Custom evaluations are the key to turning pilots into production, and hype into measurable ROI. Don't burn capital on stalled pilots while competitors move ahead with tested, trusted systems.



TALK TO US

