FROM BENCHMARKS
TO BUSINESS VALUE

How enterprises should evaluate AI

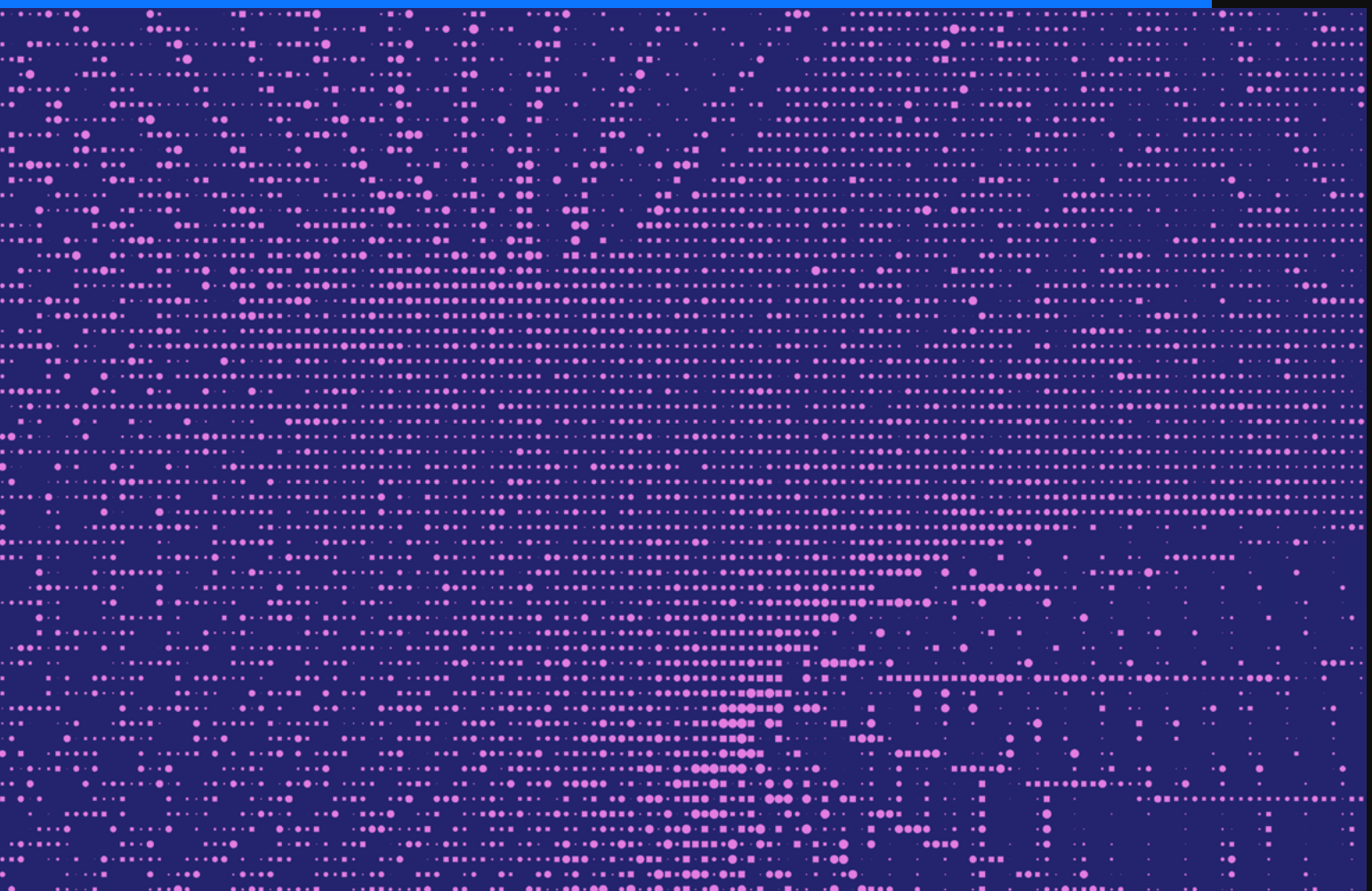
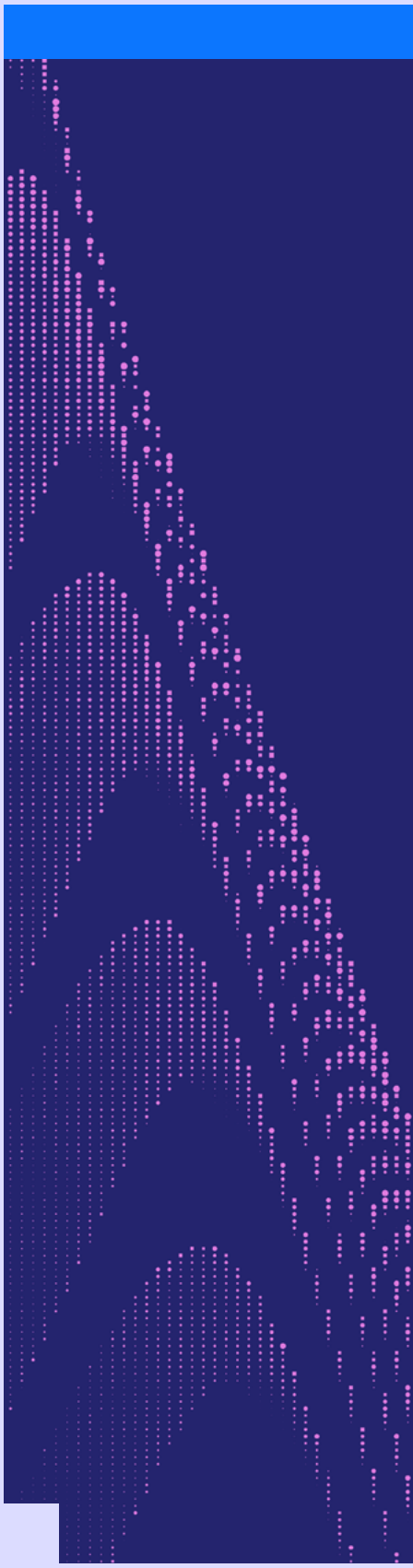


Table of contents

SECTION 01

SECTION 02

Executive summary



AI is entering production, but most enterprises are discovering a gap: the model that looked powerful in a demo often falters under the messy, high-stakes conditions of real business.

Historically, as a result of how large language models have been developed, the industry has leaned heavily on academic-style benchmarks as a proxy for capability. Much like teaching students to the test, this has left models ill-prepared for real-world unpredictability.

Consequently, the debate around model evaluations is simmering. Some lab insiders dismiss them as “vibes”, while others argue they’re vital to uncover real risks. Major model providers can sometimes rely on intuition because their teams are constantly stress-testing models and have the expertise to spot errors as they arise. Even so, custom evaluations remain essential in deployment – whether it’s checking the model gives legally compliant answers to financial queries, or ensuring it doesn’t offer unsafe medical treatment advice when prompted with health questions.

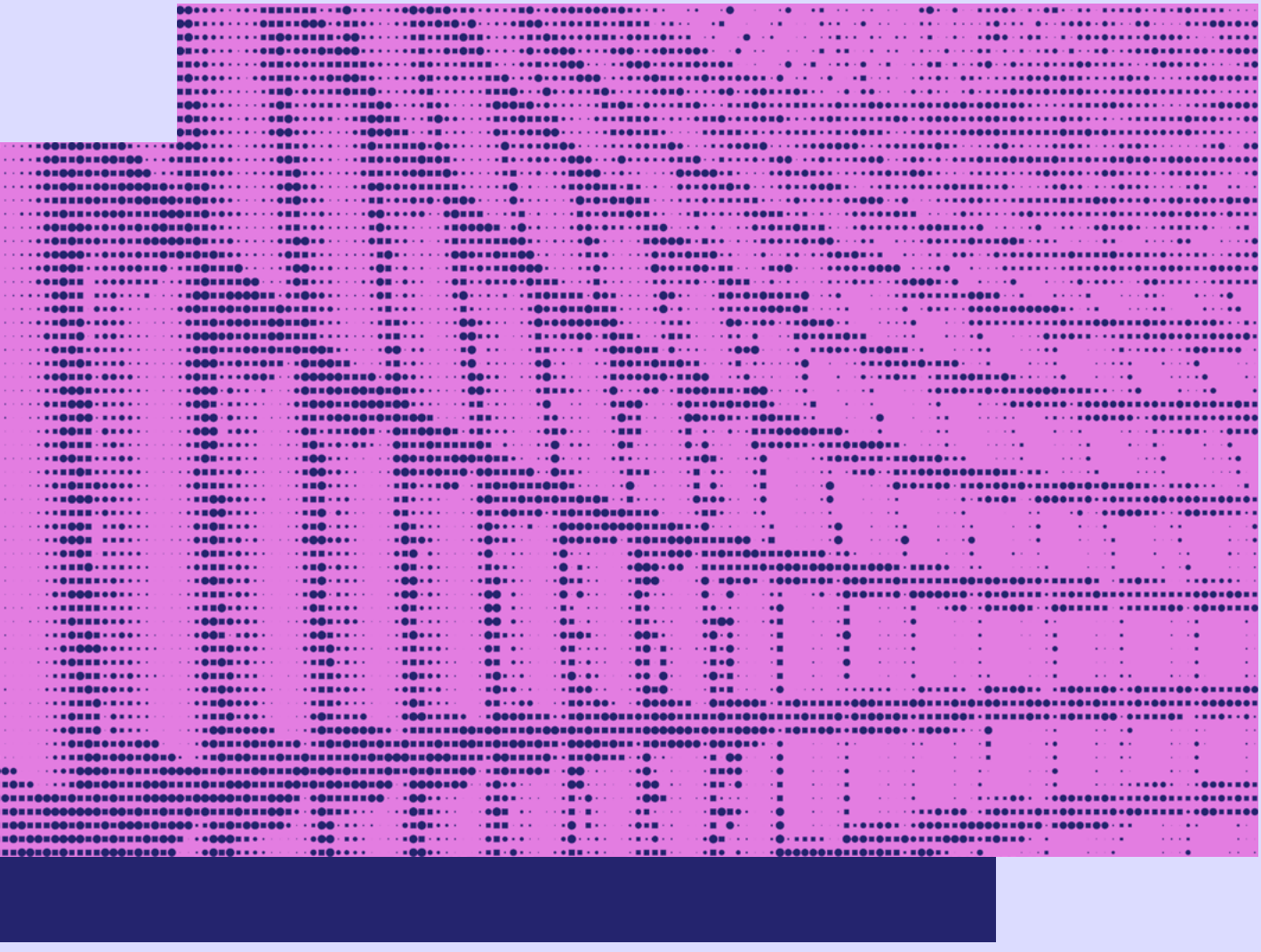
The same applies for enterprises: you can’t just trust a model because it did well on a leaderboard. You need to see how it performs with your data, your customers, and under your regulatory requirements. Capturing meaningful real-world performance requires custom, private evaluations tailored to your deployment context, reinforced by human judgment and targeted test sets.

Evidence is mounting: Vals AI tested 22 general-purpose models[1] and found that, despite strong public benchmark scores, every one averaged below 50% accuracy on simple financial analyst tasks like retrieving SEC EDGAR data.

This paper explains why off-the-shelf benchmarks mislead, and it provides a practical framework for what to measure instead. From human-in-the-loop reviews to multimodal assessment, we outline how enterprises can build custom evaluation systems that assess real-world performance for safe and accurate model deployment.

[1] Vals AI Finance Agent Benchmark

AI evaluations explained



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 production[2], 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)

The industry evals debate

"We train or assess the effectiveness or efficiency of models based on humanity's last test… which doesn't actually determine practical usage in society." [3]

Harry Stebbings
Founder, 20VC Podcast

"Vibes are the first evals. Ship something fast. See what works, what doesn't, and whether anyone cares. If the vibes are good, keep going." [4]

Julia Neagu
CEO & Co-Founder, Quotient AI

"[I] think that evals are important, but the eval pilled AI engineers [sic] have also noticed that it is not a strict requirement for success and, at least for 0-to-1 stage, may even be anticorrelated." [5]

Shawn Wang aka swyx
The AI Engineer

"When people say they 'don't do evals,' they are usually lying to themselves. Every successful product does evals somewhere in the lifecycle… it happens continuously, and is systematic." [6]

Shreya Shankar
PhD candidate, UC Berkeley and ML engineer

"Why AI evals matter… we need to know what good is. Not just in coding, but in multiple disciplines (legal, finance), tasks (diagnostics, fraud detection), and modalities (vision, voice)." [7]

Allie K. Miller
Former Global Head of AI, AWS

When to flip from vibes to rigor

Move to systematic, custom evaluations as soon as you have product traction, failures have material cost (regulatory, reputational, revenue), you scale headcount or SKUs, or you enter safety-sensitive or regulated domains. Early winners bias toward learning speed, thin guardrails, and observability; durable winners layer in rigorous, custom evals as soon as the stakes demand them, so improvement compounds without sacrificing velocity.

[3] 20VC Podcast, Sept. 15, 2025
[4] @JuliaANeagu via X, Sept. 7, 2025
[5] @swyx via X, Sept. 4, 2025
[6] sh-reya.com, Sept. 5, 2025
[7] @alliekmiller via X, Sept. 5, 2025

Custom evaluation frameworks

The solution: Custom evaluation frameworks

To address these limitations, enterprises need to adopt custom evaluation frameworks specifically tailored to their unique use cases and business objectives. Rather than providing a one-time assessment, these frameworks create a reliable, repeatable system that evolves with your organization, supporting new use cases, expanding capabilities, and monitoring production models to prevent drift.

Why custom model evaluations are more effective than generic benchmarking

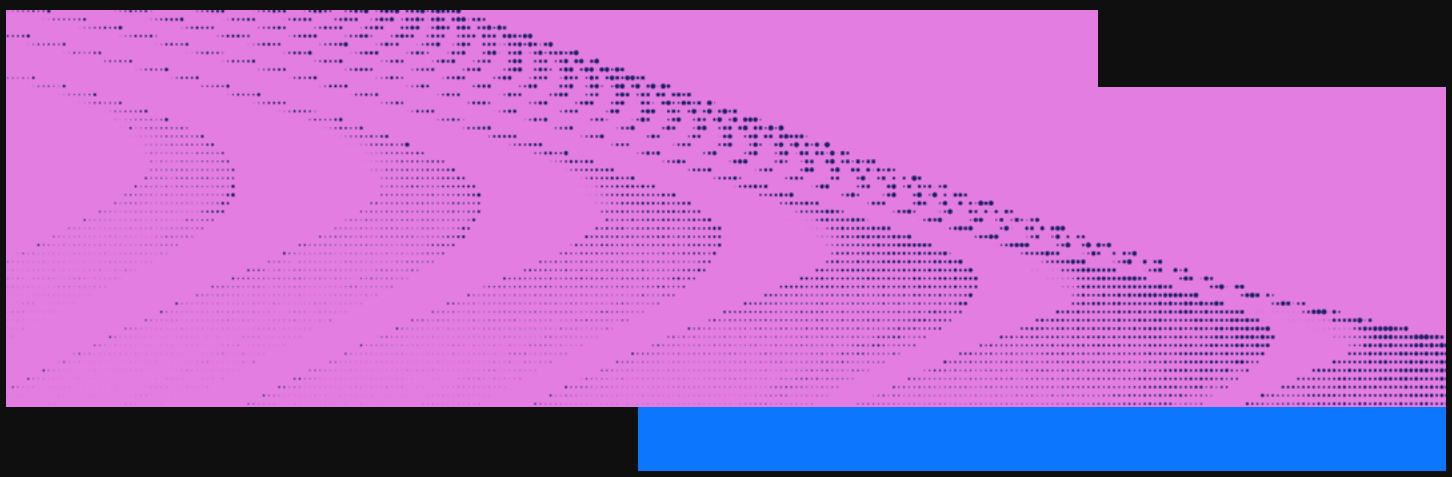
Generic benchmarks that tell you whether a model "works" or "doesn't work" according to broad and rigid standards are overly simplistic and tend to miss nuances. They can fail to identify performance issues that deeply impact usability in specific scenarios and use cases essential to your application.

Custom evaluations align with real-world benchmarks rather than those developed by researchers in more sterile academic or experimental settings. By shifting from generic benchmarks to context-specific evaluations, we redefine model quality not as a score on a test, but as its ability to deliver reliable value in practice.

"It's the breadth vs the depth. The benchmarks test everything under the sun, whether it's organic chemistry or advanced mathematics," Wronka said. "But is it testing how I do claims processing in insurance? Is it testing how you will actually use it? Today, I'm not so sure."

Custom evaluations help diagnose specific failure modes in your model. They tell you:

- Does your model work? If not, where is it failing? Focus on specific scenarios and use cases relevant to your model and user base.
- Does your model have bias, edge cases where it falls down, can it be jail broken for nefarious purposes that expose enterprise risk?
- How to build fine-tuning datasets to strategically address each error or risk vector



USE CASE

Outpacing competitors with targeted training: 87% quality improvement

Challenge

A major tech client needed to move a new AI model into production on a tight timeline without compromising quality or risking reputational damage.

Solution

Invisible designed a custom evaluation program that tested 45 internal models against real-world use cases and competitor baselines. The evaluations identified weak spots and translated directly into targeted training data, creating a rapid feedback loop for improvement.

Outcome

Within 12 weeks, the client's model achieved an 87% improvement in model quality and reached production readiness ahead of schedule, with stronger performance than competitor systems.

Building a custom evaluation framework

A custom evaluation framework starts with your specific use case and the needs of your users. It allows you to evaluate various LLMs on their ability to perform tasks directly relevant to your business objectives.

“When we design evaluations for our clients, we really want to take an approach that accurately reflects the purpose of the end user,” said Andresen.

“Do we need the insurance agent to know applied math, or do we need to ensure we’re asking an appropriate number of insurance questions? Will this involve classifying or summarizing information? Will this be retrieving account numbers, or approving claims? Are the errors in this model severe factual mistakes, or mistakes in the tone of delivery? How do we know if a model is highly accurate, but isn’t getting adopted because people just don’t like talking to it? Are we consistently seeing hallucinations or mischaracterizations of the model’s own ability? There are thousands of attributes that could be included in the measurement strategy for an AI application, and the ability to select the right combination of these attributes and rapidly deploy model evaluations to measure and monitor them was a real gap in the market that we wanted to address.”

An enterprise-grade evaluation framework involves a system of repeatable workflows that constantly help the model improve, rather than meeting one standard set of benchmarks. Custom evaluation frameworks can evolve to meet emergent needs over time.

Use-case specific assessment

When evaluating a model, you need clarity on whether it can be trusted to do specific tasks to the right level of accuracy. That means designing evaluations around the realities of your workflows.

A practical assessment for supply chain management might simulate forecasting demand across multiple geographies, introducing realistic data imperfections like late vendor updates or inconsistent SKU naming. It should test how the model behaves when a new supplier comes online mid-quarter or when regulatory requirements change without warning.

This approach ensures you’re not judging models in the abstract but in the context of the decisions, edge cases, and handoffs that define your business. Invisible’s work with clients shows that when you probe models this way, you uncover the true points of failure early and cheaply, often finding that a handful of targeted tests can reveal more about real-world readiness than an entire suite of generic benchmarks.

Building a custom evaluation framework

Industry-specific safety and compliance

For enterprises in regulated industries, evaluation must go beyond accuracy and efficiency. A healthcare provider, for example, can't afford an AI system that mishandles protected health information. A bank can't risk a model that generates disclosures out of step with SEC rules.

The practical way to approach this is to design evaluations that mirror the actual regulatory checkpoints your business faces. In healthcare, that might mean stress-testing whether the model ever exposes personal identifiers when summarizing medical records. In financial services, it could mean simulating quarterly filings and reviewing whether the outputs respect mandated reporting formats and language.

The gold standard is to codify these regulatory constraints into custom evaluations, so you can surface violations before they reach customers or auditors. This not only reduces legal exposure but also builds confidence with compliance officers and boards that the AI is being deployed responsibly.

Deep performance analysis

Surface-level accuracy scores rarely explain why a model is failing. To build trust, you need evaluations that dig into the root causes of errors. That means tracing not just what went wrong, but why.

For example, a customer-support agent might deliver an incorrect answer. A shallow benchmark would flag this as an error and move on. A deeper evaluation would reveal whether the issue stemmed from misunderstanding policy language, missing intent in a customer's message, or providing inconsistent responses to similar queries. Each failure type implies a very different fix — from targeted fine-tuning on policies, to prompt adjustments, or redesigning conversation flows.

This kind of diagnostic evaluation is recommended because it transforms vague performance gaps into actionable intelligence. Instead of pouring resources into generic retraining, you can address the precise drivers of failure and prove measurable ROI. Over time, these insights compound, allowing enterprises to not only correct errors but continuously refine their AI systems against the failure modes that matter most.

Building a custom evaluation framework

Real-world testing

An AI model that performs flawlessly on clean, artificial datasets can still collapse in production when confronted with the messiness of real use. That's why meaningful evaluation has to take place on the same kinds of data and prompts that your employees and customers actually generate.

In practice, this means stress-testing the model with spreadsheets riddled with typos, customer service transcripts with half-finished sentences, or regulatory documents that contain ambiguous clauses. It also means evaluating how the system behaves when users deviate by asking questions in unexpected ways, mixing languages, or providing incomplete information.

These real-world conditions should be simulated before a model goes into production. This reveals breakdowns that benchmarks overlook: brittle reasoning under pressure, failure to retrieve the right record from a messy database, or inconsistent tone when handling sensitive customer issues.

Evaluations led by experts: Human-AI alignment

Human-AI alignment is about making sure systems don't just generate the "right" output, but that they communicate, reason, and interact in ways that align with user expectations.

In healthcare, that might mean doctors evaluating AI-generated diagnoses, not only to confirm accuracy but also to assess whether the explanations are clear enough to guide decision-making. In finance, that might mean risk analysts testing whether an AI system not only flags a transaction as potentially fraudulent, but also explains the trigger, such as an unusual transfer pattern or a mismatch with the customer's typical behavior.

Human annotation is blended with synthetic stress tests to measure both performance and perception. Real experts validate whether the AI's answers are usable and trustworthy, while simulated cases push the system to prove it can stay consistent under thousands of variations. The result is confidence not just that the AI is technically correct, but that it integrates smoothly into human workflows — a prerequisite for enterprise adoption at scale.

Building a custom evaluation framework

Fine-tuning guidance

Custom evaluations reveal why a model is failing and how to fix it. By tracing errors back to their sources, evaluations can show whether fine-tuning should focus on additional domain data, restructuring prompts, or adjusting how the model handles edge cases.

For example, if a model in customer service consistently misunderstands cancellation requests, an evaluation might uncover that the phrasing "I want to stop my plan" isn't represented in its training data. Instead of collecting thousands of generic support tickets, fine-tuning can target a much smaller set of examples around this specific failure mode.

With this approach, enterprises can leverage lean, high-impact datasets. The result is faster improvement, lower cost, and a direct link between evaluation findings and measurable ROI.

Ultimately, every evaluation — whether it's testing compliance, diagnosing errors, simulating real-world data, or guiding fine-tuning — serves a single purpose: to measure whether a model is actually improving over time. Evaluations aren't just guardrails or one-off audits; they are the feedback loop that turns deployment into compounding progress. For enterprises, the most important KPI is not a static benchmark score but the trajectory: is the model learning, adapting, and delivering greater value with each cycle? If the answer is yes, the investment in evaluation has paid off.

Client use case: Combating internal bias

In this example, Invisible found that a client's model was 9x more likely to fail when prompted with a certain sentence structure — something their internal evaluation missed because it was built by a small group and didn't reflect the diversity of real customers.

Get started with custom AI evaluations

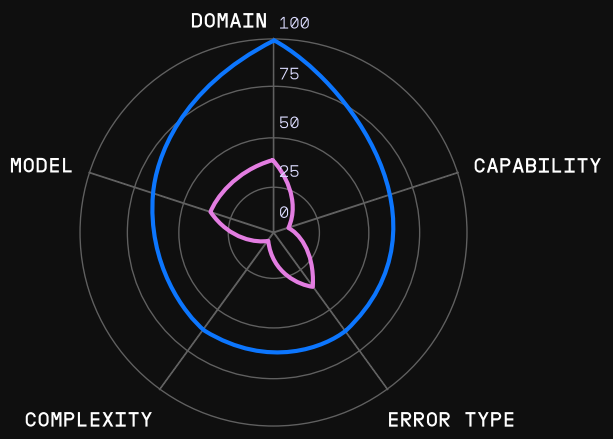
As the AI landscape continues to evolve at a rapid pace, enterprises must adapt their evaluation approaches to ensure they select and implement the right models for their specific needs. Standard benchmarks, while valuable for general comparison, fall short when it comes to assessing real-world performance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

Framework for developing custom evaluations

A custom evaluation framework isn't a list of tests, but a matrix for deciding what good looks like in your business. Consider how each of the following drives KPIs that matter for adoption, compliance, and ROI in your organization.



Domain: Speak your company's language

The first question is whether the model understands the subject matter of your industry. A system built for healthcare must handle medical terminology; one built for insurance must parse policy language. Without domain alignment, adoption stalls before it starts.

Complexity: Calibrate for the work you expect

Some enterprise tasks are simple and repetitive; others require sophisticated reasoning and context management. By evaluating against the right level of complexity, you avoid both overengineering for trivial tasks and underestimating the challenge of nuanced ones.

Capability: Measure the tasks that matter

Every enterprise use case depends on a few core actions: classifying claims, summarizing legal documents, diagnosing errors in financial reports. Evaluations should probe those exact tasks, not abstract capabilities, because they drive the KPIs tied to productivity and throughput.

Multi-turn interactions: Test the long game

Real work often happens over multiple exchanges — a claims agent asking follow-up questions, a financial analyst refining a query. Evaluations must measure whether the model maintains context, because breakdowns here quickly erode trust and efficiency.

Error types: Separate nuisance from risk

Not all mistakes are equal. A factual error in a regulatory filing is a compliance risk; a tonal misstep in a chatbot is a customer experience issue. Evaluations should weight error types by business impact, so leaders can prioritize fixes where the stakes are highest.

Ground truth vs. reasoning: Choose your governance model

For some use cases, the only thing that matters is whether the answer is correct. For others, especially in regulated industries, the reasoning process matters as much as the outcome. Aligning evaluations to the right standard clarifies whether your KPI is accuracy, explainability, or both.

Evaluation in practice

Enterprises have several ways to evaluate deployed models, each suited to different contexts and risk profiles. Effective evaluation blends multiple approaches into a layered system.

Human-in-the-Loop (HITL) evaluations

Domain experts or trained annotators directly review model outputs for accuracy, tone, and usability. This approach is essential for high-stakes or ambiguous tasks, such as legal reasoning, financial analysis, or medical notes, where automated metrics may miss nuance. HITL evaluations can be resource-intensive, but they create the “gold standard” data needed to calibrate other evaluation methods. Subject matter experts, whether in legal, medicine or finance, validate model accuracy.

Automated LLM-as-judge

Here, a second large model is tasked with scoring or ranking outputs against predefined rubrics (e.g., relevance, coherence, factual accuracy).This approach scales cheaply and quickly, but it requires careful calibration: the judge model itself may introduce bias, reward verbosity, or overestimate correctness. For that reason, LLM-judge evaluations are often combined with HITL spot-checking to validate reliability.

Heuristic- or rule-based evals

These evaluations rely on deterministic checks such as regular expressions, keyword lists, or structural rules. They are highly precise for well-defined failure modes, like ensuring SEC filings include required sections, medical notes contain disclaimers, or customer-support responses follow brand tone guidelines. While limited in scope, heuristics are fast, transparent, and invaluable as guardrails for compliance-heavy workflows.

User-based feedback loops

Real-world use generates powerful signals about model quality. Implicit metrics (like acceptance rate, edit distance, or abandonment) reveal how well outputs meet user needs without extra effort from employees. Explicit feedback (like thumbs up/down or in-app ratings) adds more signal, though it is prone to self-selection bias. Together, these feedback loops provide direct evidence of adoption, usability, and satisfaction, turning evaluations into continuous product improvement.

Live production monitoring

Even robust pre-deployment evaluations can miss failure modes that appear only under real load. Continuous monitoring of deployed systems captures drift, emerging biases, and rare but damaging edge cases. This includes anomaly detection in model behavior, logging representative samples for audit, and tracking KPIs like latency, error rates, or regulatory breaches. Monitoring is not a replacement for evaluations — it ensures evaluation insights stay relevant as models, data, and usage evolve.

Evaluation in practice

Long horizon evaluations

As enterprises scale, evaluations must extend beyond single-turn accuracy into more complex terrain. Long-horizon evaluations test whether an agent remains reliable over extended interactions — for example, whether a customer service model can sustain context, handle multiple pivots in the conversation (billing, troubleshooting, upgrade requests), and still resolve the issue in a 20-minute chat without losing track or contradicting itself.

Multi-modal evaluations

Multi-modal evaluations probe how well models integrate across text, voice, and images, and whether performance remains consistent as modalities shift. In practice, this might mean testing whether a claims-processing agent can correctly interpret both a written report and an accompanying photo of vehicle damage, or whether a medical assistant can combine patient notes with lab images to generate a reliable summary. These dimensions are still evolving, but they represent the cutting edge of how organizations will measure AI systems as they become more agentic and embedded in daily operations.

Constructing & evaluating inputs

The quality of an evaluation depends on the quality of the inputs you test against. Enterprises often over-index on "clean" prompt sets, but the real world is messy. Constructing robust inputs means deliberately including the variability your users and regulators will throw at the system. That includes:

- Representative prompts drawn from actual employee or customer interactions, complete with typos, abbreviations, and incomplete phrasing.
- Edge-case scenarios that map directly to known business risks, such as ambiguous fraud alerts in banking or borderline claims in insurance.
- Synthetic data generation to cover rare but critical cases, like adversarial jailbreaking attempts or unusual combinations of medical symptoms.

Evaluating inputs should go beyond correctness — enterprises should score how the model performs across coverage (are the most important scenarios included?), stress-resilience (does the model still work under atypical inputs?), and fairness (do different user groups get equitable results?). This approach surfaces vulnerabilities that would be invisible in standardized benchmarks.

Correcting faulty training data

Much of a model's underperformance can be traced not to technology but to data. Faulty, outdated, or biased training data is often the hidden culprit behind hallucinations, compliance failures, or brittle reasoning. Effective evaluation frameworks close the loop by not only surfacing where a model is wrong but pointing back to why.

- **Trace errors to source:** For example, if a compliance model cites outdated SEC rules, evaluations should identify that the training corpus still contains superseded regulations.
- **Targeted data curation:** Instead of amassing vast new datasets, enterprises can inject small, high-value corrections (e.g., 500 carefully labeled examples of updated disclosure requirements).
- **Data lineage and auditability:** Track where training data came from and when it was last refreshed, so errors can be isolated and corrected quickly — critical for regulated industries.

This transforms evaluations from a reporting function into a feedback loop that continuously improves the underlying model.

Behavioral and safety evaluations

Accuracy is only one dimension of trust. Enterprises also need to understand how their models behave under pressure and whether they can be safely deployed at scale. Behavioral and safety evaluations focus on surfacing failure modes that benchmarks rarely capture.

- **Adversarial red-teaming:** Stress-test the model with prompts designed to elicit unsafe or manipulative responses, such as encouraging self-harm, leaking confidential information, or recommending illegal actions.
- **Bias and fairness checks:** Evaluate how the model performs across demographic groups, regions, or customer segments. In lending, for example, this could mean ensuring identical applicants receive identical credit recommendations regardless of gender or ethnicity.
- **Reliability testing:** Run repeated, multi-turn interactions to check for consistency. Does the model give the same answer to the same prompt on different days? Does it stay aligned over long sessions?
- **Misuse potential:** Assess whether the model can be easily jailbroken or misused, for example to generate malicious code, misinformation, or insider-trading advice.

The role of independent evaluations

Neutral third-party evaluations offer something internal teams can't: an objective, unbiased assessment. Internal teams face structural blind spots. They are often under pressure to ship quickly, may unconsciously downplay weaknesses, or are simply too close to the product to see failure modes clearly. Independent evaluators counteract those forces by applying standardized methods, red-team thinking, and lessons learned across multiple organizations. They bring not only technical expertise but also institutional distance, which makes their findings more credible to regulators, auditors, boards, and customers alike.

They help organizations:

- Identify blind spots or unconscious biases that internal teams might overlook.
- Validate models against external standards and best practices.
- Provide credibility with external stakeholders, including regulators and customers.

In industries where safety, accuracy, and trust are non-negotiable, independent validation is essential.

Ready to move beyond generic benchmarks and develop evaluation frameworks tailored to your enterprise needs?

Contact us today to learn how our custom evaluation services can help you select, fine-tune, and deploy AI models that deliver measurable business value.

GET IN TOUCH

CLIENT USE CASE #1

Reducing unsafe responses with smarter data, not just more data

97%

reduction in harmful outcomes

96%

less training data required, significantly reducing the cost

Challenge

The client, a leading tech company, urgently needed to improve the safety of their AI model after harmful outputs exposed organizational risks. They believed 100,000 rows of safety data were required but lacked a clear strategy for structuring the training.

Action

Invisible analyzed the model to find the root causes of unsafe behavior and discovered that prompts using words like “pretend” or “imagine” often triggered harmful responses. Working with the client’s ML and training teams, Invisible refined the model to reduce these risks and handle such queries more safely.

Outcome

Invisible successfully reduced the frequency of conversations with harmful outcomes by 97% within six weeks using only 4,000 rows of training data—far less than the 100,000 originally anticipated. This reduction not only improved the model’s safety but also saved the client significant costs by reducing the required training data by 96%.

CLIENT USE CASE #2

Improving instruction-following with focused AI training

Frequency of faithfulness errors

42%
All prompts

68%
Sort prompts

22%
Improvement in instruction

Problem

A leading AI client struggled to improve their model's ability to follow instructions, falling behind competitors. Despite targeted training, the broad scope of the task made progress slow and gains minimal.

Solution

Invisible analyzed user prompts and found that "sort" queries were the biggest weakness. By narrowing training to these commands, Invisible significantly improved the model's instruction-following.

Outcome

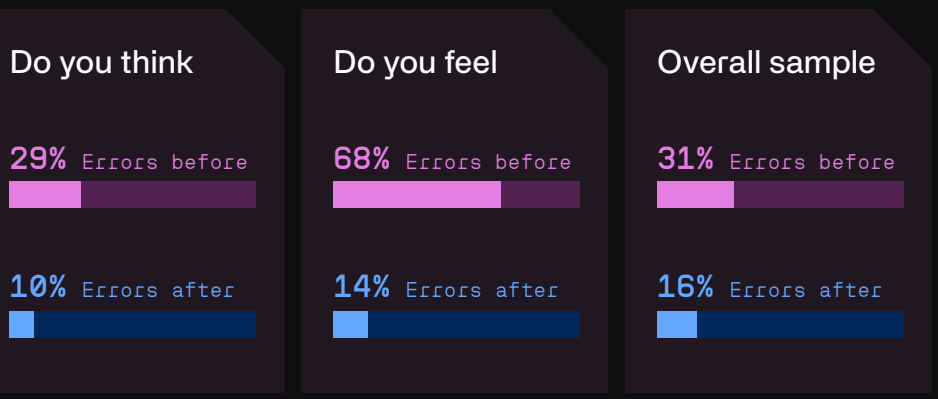
Tests showed a 22% improvement in instruction-following after targeted training. Invisible also found the model was 62% more likely to fail when prompts included "sort," making this the most important area to address.

CLIENT USE CASE #3

Reducing hallucinations through custom evaluation and training

79%

fewer hallucinations in key prompts



Challenge

The client struggled to structure AI training and set priorities for improving performance. The model frequently hallucinated, creating risks for customer interactions.

Solution

Invisible evaluated the model to find the highest-risk areas for hallucinations. By analyzing language patterns, we identified specific triggers for errors and used these insights to guide and optimize training.

Outcome

The targeted training approach sharply reduced hallucinations. “Do you feel?” prompts improved 70% faster than average, while hallucinations dropped 65% for “Do you think” and 79% for “Do you feel” prompts, mitigating risk in these high-impact areas.

Invisible

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