A SAU5 BUYER'S GUIDE · 2026

Sau5

AI testing isn't AI observability isn't *AI guardrails.*

A vendor-neutral map of the four categories of AI quality vendor. What each does, when you need it, and how to pick what you actually need.

Sau5
GLOBAL AI TESTING CONSULTANCY

Edition 01, May 2026
SAU5.AI

Why every AI quality conversation *ends in confusion.*

Here is a conversation that happens in enterprise AI teams every week. An engineering leader says, *"We've decided we need AI testing."* What they describe next is three different products from three different vendor categories. None of them solve the problem they think they're solving.

The market for "AI quality" looks like one thing. It's actually four separate categories. Four different vendor types. Four different commercial models. Four different times in the AI lifecycle when each becomes relevant.

Buyers who don't know the map tend to make one of three mistakes. The first is buying a category they don't need yet, running it for six months, and finding the original problem still sitting there. The second is buying a "do-it-all" platform that turns out to do one category well and the other three at a feature-checklist level. The third is buying nothing at all, shipping the AI feature anyway, and finding out which category they needed when something breaks in production.

This guide is the map. It's short, opinionated, and vendor-neutral. Sau5 sits in one of the four categories, and the guide is honest about which one and what we don't do.

> **HOW TO USE THIS GUIDE**
>
> Pages 3 to 6 cover the four categories. Page 7 has the lifecycle map. Page 8 helps you figure out which category you actually need. Page 9 covers common buying mistakes, and page 10 explains where Sau5 fits.
>
> Total reading time: around 12 minutes.

CATEGORY 01

AI *Testing*

WHAT IT IS

A methodology applied at a defined point in the AI lifecycle. Usually pre-launch, post-incident, vendor evaluation, or periodic re-runs. The deliverable is a runnable harness and a findings report. The buyer gets quantitative scores against defined thresholds: retrieval accuracy, grounding rate, hallucination rate, adversarial robustness.

WHAT IT ISN'T

Not a dashboard. Not always-on. Not a runtime filter. Testing is a project, not a stack.

A NOTE ON SCOPE WITHIN AI TESTING

AI testing in 2026 is not one homogeneous discipline. The mature methodology applies to **RAG and LLM-mediated systems**, where a model retrieves context and produces an answer. **Agentic AI** (multi-step reasoning, tool calls, state across turns, multi-agent coordination) requires different methodology, different tooling, and different skills. Sau5's scope is the first. Agent and workflow testing is covered on page 10.

WHO DOES IT

Specialist consultancies and a small number of in-house teams that build their own methodology. Sau5 is in this category. AI Testing is the smallest of the four categories by vendor count, partly because the methodology requires depth that doesn't transfer directly from other software testing disciplines.

WHEN YOU NEED IT

When you're about to launch an AI feature and need a defensible answer to *"how do we know it works?"*. When something has broken in production. When you're evaluating a vendor's claims before signing a contract. When you've been running for six months and want to know what has drifted.

COMMERCIAL SHAPE

Fixed-scope project (4 to 6 weeks typical) or embedded testers on retainer. Output is methodology plus harness plus handover.

CATEGORY 02

AI *Observability*

WHAT IT IS

Runtime monitoring of production AI traffic. It sits in your stack continuously and shows you what's actually happening: token usage, latency, cost per call, model drift, error rates, alerts when something looks wrong. The dashboard is the product.

WHAT IT ISN'T

Not a methodology. It tells you *what is happening*, not *whether what is happening is correct*. Observability shows you a 30% increase in token usage. It doesn't tell you whether the answers got better or worse.

WHO DOES IT

Arize, WhyLabs, Helicone, Datadog LLM Observability, Langfuse, and a growing roster. Heavy overlap with traditional APM (Application Performance Monitoring) vendors extending into LLM territory.

WHEN YOU NEED IT

Once you're live and need visibility into what's happening. Mature AI teams running large query volumes. Compliance teams that need an audit trail.

COMMERCIAL SHAPE

SaaS subscription, usage-based pricing. Lives in your stack continuously once installed.

CATEGORY 03

AI *Eval Platforms*

WHAT IT IS

A dashboard for managing evaluation runs. Teams write their own test cases, define their own scoring functions, point the platform at their model, and get aggregated results over time. Sometimes called eval orchestration.

WHAT IT ISN'T

Not methodology. The platform doesn't tell you *what* to test. It only helps you organise the testing you've already decided to do. A team using one of these platforms still has to build the golden dataset, write the scoring prompts, and decide the thresholds themselves.

WHO DOES IT

Braintrust, Galileo, LangSmith, Vellum, Humanloop, and a fast-growing list. Most have grown out of the LLM application framework space (LangChain, LlamaIndex) and bundle eval as one feature of a larger developer platform.

WHEN YOU NEED IT

When your engineering team wants to run their own evaluations continuously and needs a place to manage them. When you've already figured out what to test and just need infrastructure to do it at scale.

COMMERCIAL SHAPE

SaaS subscription, often a free tier for small volume. Tool, not consultancy.

CATEGORY 04

AI *Guardrails*

WHAT IT IS

Runtime filters that block bad outputs before they reach users. The guardrail sits between the model and the user. Live content filtering, PII redaction, prompt injection rejection, jailbreak blocking, response groundedness enforcement.

WHAT IT ISN'T

Not testing. A guardrail doesn't tell you whether the model is correct on average. It tries to block specific bad outputs at the moment they happen. If the model is unreliable, guardrails are a band-aid rather than a fix.

WHO DOES IT

Lakera, NeMo Guardrails (NVIDIA), Guardrails AI, Patronus, Aporia, and a growing list focused on safety, compliance, or both.

WHEN YOU NEED IT

When you need a runtime safety net, not just pre-launch confidence. Regulated industries where certain output categories must be blocked (legal advice, medical advice, harmful content). Customer-facing AI where a single bad answer carries real cost.

COMMERCIAL SHAPE

SaaS subscription or self-hosted. Usually priced per call or per workspace.

How the four categories *fit together.*

The categories solve different problems at different points in the AI lifecycle. They're complementary, not interchangeable. A mature AI program uses three or four of them. A new AI program usually needs only one or two, and the most common mistake is buying the wrong one first.

PHASE	WHAT YOU'RE DOING	CATEGORY THAT HELPS
Build	Developing the AI system itself	*None, your engineers*
Pre-launch	Verifying the system works against defined criteria	**AI Testing**
Pre-launch	Setting up runtime safety filters	**AI Guardrails**
Pre-launch	Setting up an engineering eval pipeline	**AI Eval Platform**
Post-launch	Monitoring production traffic and quality	**AI Observability**
Post-launch	Continuous regression testing as you iterate	**AI Eval Platform** plus periodic **AI Testing**
Post-incident	Diagnosing what failed and how to prevent recurrence	**AI Testing**
Vendor eval	Independent assessment of a vendor's AI claims	**AI Testing**

THE FOUR QUESTIONS

AI Testing answers *"is this correct?"*. **AI Observability** answers *"what is happening in production?"*. **AI Eval Platforms** answer *"where do we store and compare test results over time?"*. **AI Guardrails** answer *"how do we block bad outputs at runtime?"*

Four different questions. Four different answers. Buyers who confuse the questions buy the wrong answers.

Which one do you actually *need?*

The quickest test, answer the question *"what are you trying to prove?"*. The answer points to one of the four categories. Sometimes two. Almost never all four at once.

If you're saying: *"I'm launching an AI feature in 4 to 6 weeks and need to know whether it actually works."*

You need AI Testing.

If you're saying: *"We launched. Now we want to know what's happening in production."*

You need AI Observability.

If you're saying: *"Engineering wants to run evaluations continuously and needs a place to manage them."*

You need an AI Eval Platform.

If you're saying: *"Something keeps reaching customers that shouldn't."*

You need AI Guardrails.

If you're saying: *"We need to evaluate a vendor before signing."*

You need AI Testing.

If you're saying: *"Something broke in production and we need to know what."*

You need AI Testing, then AI Observability.

If you're saying: *"All of the above and we're not sure where to start."*

Start with AI Testing. The findings tell you which others to layer.

Five common *buying mistakes.*

Mistake 01

Buying an eval platform and calling it testing.

An eval platform is infrastructure for running tests. It doesn't tell you *what* to test. Buying one before you have a methodology gives your team a dashboard with nothing meaningful on it. A platform without a methodology behind it surfaces vanity metrics, not signal about whether the system is actually correct.

Mistake 02

Buying observability before launch.

Observability is post-launch. Before launch you have no production traffic to observe. Teams that buy observability tools pre-launch end up paying for a stack that doesn't fire until months later, and meanwhile they have no answer to the pre-launch *"does this work?"* question.

Mistake 03

Treating guardrails as testing.

Guardrails block bad outputs at runtime. They don't tell you the bad outputs were going to happen, what caused them, or how common they are. A team relying on guardrails alone is operating without instrumentation. They know something was blocked, but not why the model produced it in the first place.

Mistake 04

Buying a "do-it-all" platform.

Several vendors claim to cover multiple categories. Most are strong in one and weak in the others. The practical test, ask which of the four categories they would still sell you if you only had budget for one. That's the category they actually do well. The other three are usually marketing-ware.

Mistake 05

Skipping testing entirely.

The most expensive mistake. Teams that ship without pre-launch testing find out which failure modes they have by reading customer complaints. The cost of fixing a problem in production is somewhere between 10x and 100x the cost of fixing it pre-launch. The cost of *not* fixing it is unbounded.

Where *Sau5* fits in the landscape.

Sau5 is in the first category. AI testing. Methodology, harness, findings, handover. That's what we do, and we don't do the other three. Being clear about that is part of how we position the service.

What Sau5 does

A defined four-week assessment of any production RAG system, run across five domains: retrieval quality, answer grounding, hallucination detection, adversarial robustness, and eval operations. Fixed scope. Fixed fee. Client owns the harness at handover.

We also offer embedded AI testers, qualified people doing the continuous testing work alongside your team. Different engagement shape: ongoing capacity rather than a defined project. Both options are available. The right one depends on your situation.

What Sau5 doesn't do

Two kinds of scope boundary apply. The first is the three adjacent vendor categories from page 5 (observability, eval platforms, guardrails). The second is a scope boundary *within* AI testing: agentic AI workflow testing, which is a different discipline.

- **Agent and workflow testing.** Multi-step reasoning, tool and function call validation, state and memory across turns, multi-agent coordination, trace analysis across LangGraph, CrewAI, AutoGen, OpenAI Assistants API, and similar frameworks. Different methodology, different tooling, different skills. If your RAG layer sits underneath an agent, Sau5 can test that RAG layer; the orchestration on top is out of scope. Specialist agent-testing teams are emerging; eval platforms (LangSmith, Langfuse, Phoenix) have basic agent eval features for engineering teams that want to manage it in-house.
- **Observability.** Arize, WhyLabs, Helicone, Langfuse, or Datadog LLM Observability.
- **Eval platforms.** Braintrust, Galileo, LangSmith, Vellum, or Humanloop. Our harness writes results to any of them. We integrate, we don't replace.
- **Guardrails.** Lakera, NeMo Guardrails, Guardrails AI, Patronus, or Aporia.

How to engage Sau5

1. **Read the methodology.** The full five-domain framework is at sau5.ai/methodology. A free 15-page sample of our Test Cases Manual is linked from the home page.
2. **Join the waitlist.** First engagements are being scoped now. Slots are limited. Form on sau5.ai.
3. **Talk to us.** Not sure whether you need testing or one of the other three categories? Email hello@saucinco.com. We'll tell you honestly, including if you don't need us.

About this *guide.*

About Sau5

Sau5 is a global AI testing consultancy. We assess RAG and LLM-mediated systems across five domains in a fixed four-week engagement, and deliver a client-owned eval harness at handover. We deliver in English and Spanish. Trading entity: Saucinco SAS.

How this guide came together

The four-category framework in this guide describes the AI quality vendor market as it stands in 2026. It isn't a Sau5 invention. The categories already exist. This guide is the first published map laying them out in one place from a vendor-neutral perspective. The vendor lists in each category are illustrative rather than exhaustive. New vendors enter regularly.

Useful external references

For deeper reading on the foundations behind RAG accuracy testing:

- Magesh, V., et al. (2024). *"Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools."* Stanford RegLab / Stanford HAI.
- Es, S., et al. (2023). *"RAGAS: Automated Evaluation of Retrieval Augmented Generation."* arXiv:2309.15217.
- Li, J., et al. (2023). *"HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models."* arXiv:2305.11747.
- Min, S., et al. (2023). *"FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation."* arXiv:2305.14251.
- EU AI Act (Regulation 2024/1689). Articles 13 (transparency) and 14 (human oversight).

Licence and reuse

Sau5

AI testing. *Only AI testing.*

The narrow specialist consultancy for production RAG and LLM systems. Five domains. Four weeks. Fixed fee. Client-owned harness at handover.

If you've read this guide and decided AI testing is what you need, that's what we do. If you've decided you need one of the other three categories, we hope this guide helps you pick the right vendor in that category.

FIND US

sau5.ai · hello@saucinco.com
LinkedIn · X (@Sau5ai) · YouTube (@sau5ai) · GitHub (sau5ai)