

Best AI LLM Observability Tools in 2026: 5 Compared

Langfuse, LangSmith, Arize, Braintrust and Helicone compared on pricing, retention, self-hosting and who each one suits best.

Why LLM observability needs its own tooling

Once an LLM feature ships, your logs stop telling you much. A request can return 200 OK and still give a wrong answer, a slow tool call or a prompt that costs four times what it should. LLM observability tools record each step of a model call or agent run as a trace. You can then inspect it, score it with evals and compare runs after you change a prompt or model.

This roundup covers five products that between them handle most of what teams need in 2026. Pricing, retention and licensing details come from each vendor's own pricing pages and repositories as of October 2026. scored.tools has not published ratings for these five yet, so none of them gets a score below and the entries run alphabetically. The comparison table and final picks focus on what you pay, how long your data is kept and whether you can run the tool yourself.

1. Arize (Phoenix and AX)

Score: Not yet scored

Arize ships two products. Phoenix is the open-source one: per its GitHub repo, it is licensed under the Elastic License 2.0, built on OpenTelemetry, and uses the OpenInference project for auto-instrumentation of frameworks and LLM providers. You can start it locally with `pip install arize-phoenix` or deploy it with Docker or a Helm chart. Phoenix covers tracing, evals, versioned datasets, experiments, a prompt playground and prompt management. Arize AX is the hosted platform. Its free tier is the most generous of the hosted options here on users and evals, with both unlimited, but retention is short at 15 days. One caveat on the license: ELv2 is source-available rather than OSI open source. It restricts offering Phoenix as a managed service to third parties, which matters to vendors and not much to anyone else.

Best for: Teams that want OpenTelemetry-native tracing they can self-host, with a hosted upgrade path.

Pricing: Phoenix is free to self-host. AX Free is \$0 for 25k spans, 1 GB ingestion and 15-day retention. AX Pro is \$50/month for 50k spans, 10 GB and 30-day retention. AX Enterprise is custom and adds self-hosting, SSO, audit logs and HIPAA.

2. Braintrust

Score: Not yet scored

Braintrust puts evaluation first, and its pricing shows it. You pay for processed data and scores rather than for traces, so a team that runs heavy eval suites against small payloads gets plenty from the free tier. Teams logging large production prompts may find the per-GB meter adds up faster. Each plan includes model credits (\$10/month on Starter, \$100/month on Pro), which cover LLM-as-judge scoring and playground runs before token rates apply. The jump from free to \$249/month is the steepest entry fee in this group. Braintrust says startups on Pro may qualify for 6 to 12 months free, which takes some of the sting out.

Best for: Product teams whose main loop is changing a prompt, running an eval and comparing scores.

Pricing: Starter is \$0/month with 1 GB of processed data (\$4/GB after), 10K scores (\$2.50 per 1,000 after) and 14-day retention. Pro is \$249/month with 5 GB (\$3/GB after), 50K scores (\$1.50 per 1,000 after) and 30-day retention, extendable to 180 days at \$0.50/GB/month. Enterprise is custom.

3. Helicone

Score: Not yet scored

Helicone began as a proxy-style gateway. You route model calls through it and get request logging, cost tracking and alerts with little code change. The Pro tier adds HQL, Helicone's query language for digging through logged requests. Its plans are built around request volume and rate limits: Hobby allows 10 logs per minute, Pro 1,000, Team 15,000 and Enterprise 30,000, per the pricing page. That makes the free tier fine for a prototype but not for production traffic. The pricing page now carries a "Helicone Joins Mintlify" banner. The page says nothing about what the deal means for the product, so check the roadmap before you commit for the long term.

Best for: Teams that want cost and request logging across providers with minimal instrumentation work.

Pricing: Hobby is free with 10,000 requests/month, 1 seat and 7-day retention. Pro is \$79/month with unlimited seats and 1-month retention. Team is \$799/month with 3-month retention, SOC 2 and HIPAA. Enterprise is custom, with configurable retention, SAML SSO and on-prem deployment. Paid tiers include 10K requests and then bill by usage.

4. Langfuse

Score: Not yet scored

Langfuse is the default answer for teams that want open source and the option to self-host. The vendor's pricing page states plainly that Langfuse "is open source and you can self-host it for free," with Docker Compose for local use and Kubernetes templates for production. The cloud plans are cheap to start. Core at \$29/month includes unlimited users and 90-day data access, which beats the paid entry tiers of most rivals on retention. Pro at \$199/month stretches access to 3 years, which helps if you need long audit trails for regulated work. Overage is simple at \$8 per 100k units, with graduated discounts at higher volumes. Langfuse works with any framework, so you are not tied to one SDK.

Best for: Engineering teams that want self-hosting, or a low-cost cloud plan with long retention.

Pricing: Hobby is free with 50k units/month, 30-day access and 2 users. Core is \$29/month with 100k units and 90-day access. Pro is \$199/month with 100k units and 3-year access, plus an optional \$300/month Teams add-on. Enterprise is \$2,499/month. Extra usage is \$8 per 100k units.

5. LangSmith

Score: Not yet scored

Langsmith comes from LangChain, and if your stack already runs on LangChain or LangGraph, the tracing works with almost no setup. It also accepts traces from apps that don't use LangChain. Pricing is per seat, which suits small teams and gets expensive as more people need access. Retention is the weak point. Base traces are kept for 14 days on every plan, and keeping one for 180 days means upgrading it to an extended trace for an extra fee. The pricing page lists pay-as-you-go overage but no per-trace rate, so model your volume with sales before signing.

Best for: Teams building on LangChain or LangGraph that want tracing and evals inside that stack.

Pricing: Developer is \$0 for 1 seat and 5k base traces/month. Plus is \$39/seat/month with 10k base traces. Enterprise is custom. Base traces are kept 14 days, and extended traces 180 days for an extra fee.

Comparison table

Tool	Score	Free tier	Paid entry	Paid entry retention	Self-host
Arize	Not yet scored	25k spans, 15 days	\$50/month (AX Pro)	30 days	Yes (Phoenix, ELv2)
Braintrust	Not yet scored	1 GB, 10K scores, 14 days	\$249/month	30 days (up to 180 paid)	Enterprise only

Tool	Score	Free tier	Paid entry	Paid entry retention	Self-host
Helicone	Not yet scored	10,000 requests, 7 days	\$79/month	1 month	Enterprise on-prem
Langfuse	Not yet scored	50k units, 30 days	\$29/month	90 days	Yes, free
Langsmith	Not yet scored	5k traces, 1 seat	\$39/seat/month	14 days (180 paid)	Enterprise only

Final picks

Best overall value: Langfuse. It has the cheapest paid tier, the longest retention at that price and free self-hosting, so it is hard to beat on paper.

Best for OpenTelemetry shops: Arize. Phoenix gives you OTel-native tracing you can run yourself, and AX Pro at \$50/month is a cheap hosted step up.

Best for eval-heavy teams: Braintrust. The pricing rewards scoring and experiments, and the startup offer helps with the \$249 entry price.

Best for LangChain users: Langsmith. It needs the least setup on that stack, but budget for per-seat costs and short default retention.

Best for quick cost logging: Helicone. Its gateway setup is the fastest way to see spend per provider, though you should watch how the Mintlify deal plays out.