Arize Calhacks Hacker Starter Pack

Arize is an AI engineering platform focused on evaluation and observability. It helps engineers develop, evaluate, and observe AI applications and agents.
Arize has both Enterprise and OSS products to support this goal:

- Arize AX — an enterprise AI engineering platform from development to production, with an embedded Alyx
- Phoenix — a lightweight, open-source project for tracing, prompt engineering, and evaluation

Why Observability Matters

Observability matters because it gives you insight into what's really happening under the hood with your AI applications/agents. It's how you spot when your agent goes off track, starts hallucinating, or gives a confusing response. For example, if users keep getting frustrated when talking to a chatbot, observability can show you whether the issue came from missing context, bad retrieval, or a broken tool call.

Arize Phoenix

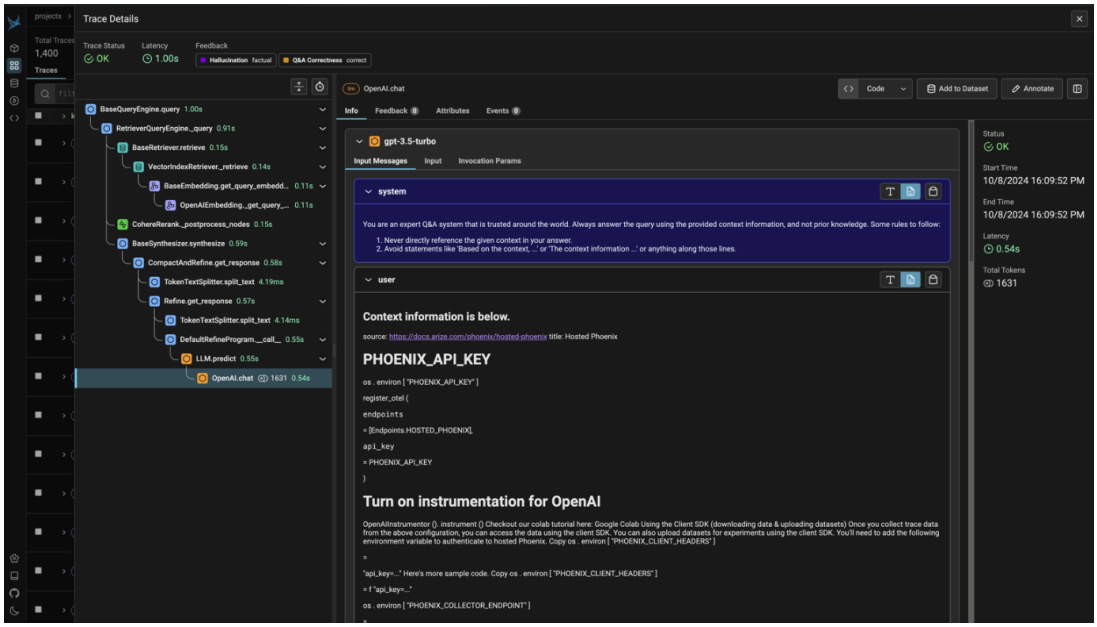
At Calhacks, we'll be looking at Phoenix, our open source observability product! Phoenix Cloud offers free, ready-to-use Phoenix instances preconfigured with 10 GiB of storage, more than enough for any project pre-deployment.

Observability unlocks tracing and evaluation pipelines.

What is Tracing?

[Tracing Quickstart]

LLM tracing records the paths taken by agents or LLM apps as they propagate through multiple steps of a request. For example, when a user interacts with an LLM application, tracing can capture the sequence of operations, such as tools calls, LLM response generation, and document retrieval to provide a detailed timeline of the request's execution.

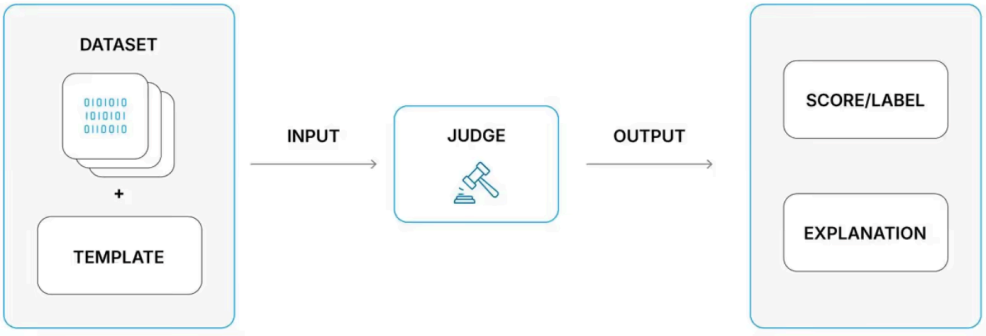


What are Evals?

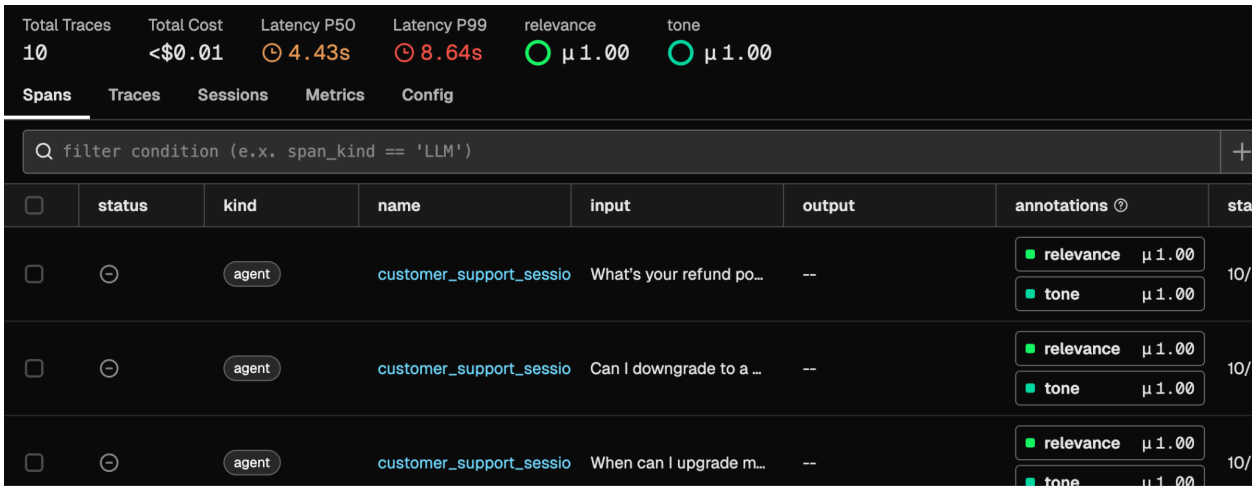
[Evals Quickstart]

LLMs are non-deterministic, and their execution and outputs can vary even with the same input. To ensure application performance, you need a way to judge the quality of your LLM outputs.

There are many ways to go about this, but the most common is using LLM-as-a-Judge. This is an evaluation method that uses an LLM to judge the output of another LLM.



This approach works great for evaluations because you can define any custom criteria and describe it in plain English. From there, you can feed in existing data, run it through the Judge, and review aggregate metrics to understand overall performance.



Cookbooks and Tutorials

These notebooks/tutorials are to help you getting started with implementing Tracing/Evals and building more robust Agents and LLM Applications.

- End-to-End Tutorial
- Creating a custom LLM Judge evaluator

Videos

- Tracing and Evaluating OpenAI Agents
- Trace Level Evals
- Session Level Evals

Career Opportunities

- Open Roles