arize

Four ways to run evals

From deterministic code checks to fully agentic harnesses

Code

WHAT IT DOES

Deterministic rule-based checks on the output — format, schema, exact match, regex.

ADVANTAGES

Fast, cheap, fully objective and repeatable. Use it whenever you can: a unit test for your model.

LLM as a judge

WHAT IT DOES

An LLM scores subjective, nuanced qualities like tone, helpfulness and groundedness.

ADVANTAGES

Handles judgement code can't, reaching about human-level agreement with raters. Reserve for the fuzzy stuff.

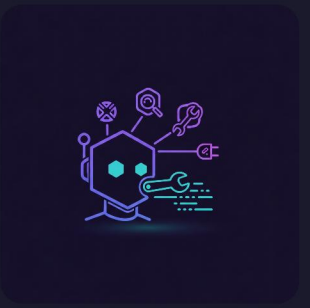
Code + LLM

WHAT IT DOES

Code deterministically extracts the right trace data first, then the LLM judges that clean signal.

ADVANTAGES

Cheaper and far more reliable than pointing a judge straight at raw, messy traces.



Harness as a judge

WHAT IT DOES

An agent like Claude Code pulls traces via API or CLI, inspects spans and uses tools to investigate.

ADVANTAGES

The most flexible — explores dynamically, follows leads, and can even propose new evals itself.

DETERMINISTIC / FAST / CHEAP

AGENTIC / FLEXIBLE / POWERFUL