arize

Evals across the lifecycle

What evaluation looks like at every stage of building an AI app

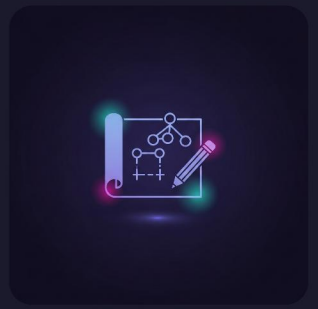
Requirements

THE STAGE

Decide what the app must do and what a good answer actually looks like.

WHERE EVALS COME IN

Turn those requirements into a golden dataset and concrete success criteria.



Design

THE STAGE

Choose the prompts, models, and architecture that will power the app.

WHERE EVALS COME IN

Compare prompts and models head to head on the golden dataset before committing.

Implementation

THE STAGE

Build the application and wire up its tools, retrieval, and integrations.

WHERE EVALS COME IN

Instrument it with tracing so every span, trace, and session is inspectable.

Testing

THE STAGE

Validate behaviour before shipping, the way you would unit-test code.

WHERE EVALS COME IN

Eval-driven development: Arrange, Act, Evaluate, Assert against expected scores.

Deployment

THE STAGE

Ship to real users, where inputs are messier than anything you tested.

WHERE EVALS COME IN

Run online evals on live traffic and alert the moment quality regresses.

BUILD / PRE-RELEASE

PRODUCTION / LIVE