HOW TO BUILD AN EVAL

From a golden dataset to evals you can trust

GOLDEN DATASET

WHAT YOU DO

Collect representative real examples your app must handle. Synthesize extra cases when real data is thin.

WHY IT MATTERS

Everything downstream is measured against this set, so it has to mirror real usage.

HUMAN OPEN CODING

WHAT YOU DO

A human reads each trace and writes free-form notes on what went wrong, no fixed categories yet.

WHY IT MATTERS

Surfaces real failure modes from the ground up instead of assuming what to look for.

AXIAL CODING

WHAT YOU DO

Human and LLM together group the open-coding notes into a tight set of recurring failure categories.

WHY IT MATTERS

Turns messy observations into a clear taxonomy – the things your evals will actually measure.

DESIGN EVALS

WHAT YOU DO

Build a targeted eval for each failure category – code checks where you can, an LLM judge where you must.

WHY IT MATTERS

Each category gets the cheapest reliable check, so coverage maps directly to real failures.

TEST vs GOLDEN SET

WHAT YOU DO

Run the new evals against the golden dataset and compare their verdicts to the human labels.

WHY IT MATTERS

Confirms the evals agree with humans before you trust them to grade production traffic.

GATHER / OBSERVE

MEASURE / VALIDATE