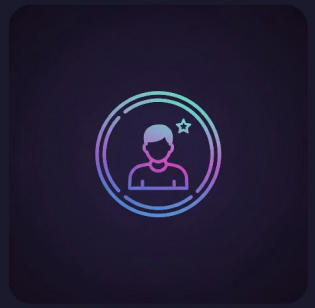
arize

Anatomy of an eval prompt

What you write to tell the judge how to judge



Role

WHAT IT IS

Set the judge's persona and expertise up front, framing it as the kind of reviewer who should assess this output.

WHY IT MATTERS

A clear role steers the model to reason like the right expert instead of an average, generic responder.

Criteria

WHAT IT IS

Spell out exactly what you are judging in plain terms, such as whether a response is factually correct given the retrieved context.

WHY IT MATTERS

Vague prompts give vague scores. Concrete, explicit criteria are what make a judge reliable.

Rubric

WHAT IT IS

Define the scoring scale and what each level means, so the judge knows the difference between a pass, a partial, and a fail.

WHY IT MATTERS

A rubric turns a fuzzy gut-feel judgement into a consistent, repeatable score you can trust across runs.

Examples

WHAT IT IS

Show a handful of labelled good and bad cases drawn from your own data so the judge can pattern-match to them.

WHY IT MATTERS

Few-shot examples anchor the judge to your standards rather than its own priors, sharply lifting agreement.

WHO IS JUDGING

HOW THEY SCORE