

Train, dev, test

How to split your labelled traces to build an evaluator you can trust



Train

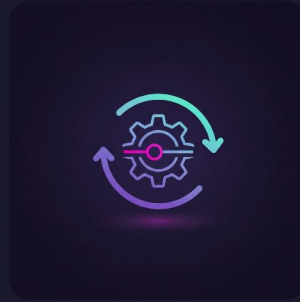
10-20% OF DATA

WHAT IT'S FOR

A handful of hand-labelled traces that teach the judge what good and bad look like.

HOW YOU USE IT

Drop them into the prompt as few-shot examples so the judge mirrors your calls.



Dev

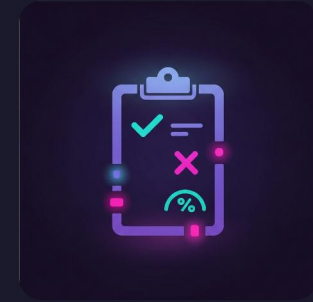
20-30% OF DATA

WHAT IT'S FOR

A held-out set you iterate the prompt against, run after run.

HOW YOU USE IT

Tweak the wording until the judge's scores line up with your own labels.



Test

50-70% OF DATA

WHAT IT'S FOR

Unseen labelled traces the judge has never touched, scored exactly once.

HOW YOU USE IT

Measure true-positive and true-negative rates — your evidence the eval can be trusted.