



PRE-PRODUCTION EVALUATION

Crucible

Crucible is an evaluation harness for agentic systems. Capture a full agent run, score it against a versioned rubric with an automated judge, track quality over time, and export evidence – so you ship what survives, not what you hope works.

WITHOUT IT

An agent's real behaviour is unknown until it happens in production. 'It looked fine in a few manual runs' is the whole quality bar, and a regression only shows up once it's live.

WITH IT

Every agent run is scored against a versioned rubric before it ships. You keep only what clears the bar, catch regressions as scores move, and export the evidence that proves it.

WHAT IT IS

Agentic evaluation harness

SCORES

Full agent trajectories vs a versioned rubric

EXPORTS

Evidence reports for governance

STAGE

Working harness · keyless one-command demo

WHERE IT FITS

Crucible is the gate between a promising agent and a trusted one – the pre-production check that lets you put a governed agent into the stack on evidence, not optimism.