BUILD IN PUBLIC

REAL CASE

My evaluator got all four right.

What mattered was where it hesitated.

Claude Code

Codex

Jev

Roberto J. Aguirre G. · Head of AI

THE STACK

Three roles. Three tools.

ORCHESTRATOR

Claude Code

Orchestrates and verifies

Builds every brief and checks the evidence before anything is marked done.



BRAIN AND HANDS

Codex

Designs and implements

At high effort it plans the solution. At medium effort it writes the code.



EVALUATOR

Jev

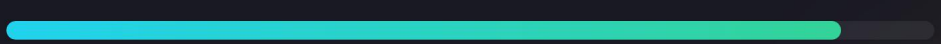
Decides and reviews the diff

Writes no code. Reads the change and the tests, answers closed questions, and says how confident it is.

REAL CASE

One real change. Four questions.

Do the same tests fail as in the previous version?

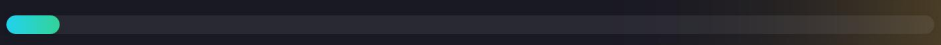


0.90

What caused the failures?

environment **0.90** · already there **0.10** · change **0**

How risky is the change, from 0 to 3?



0.17

Is it safe to merge without further review?



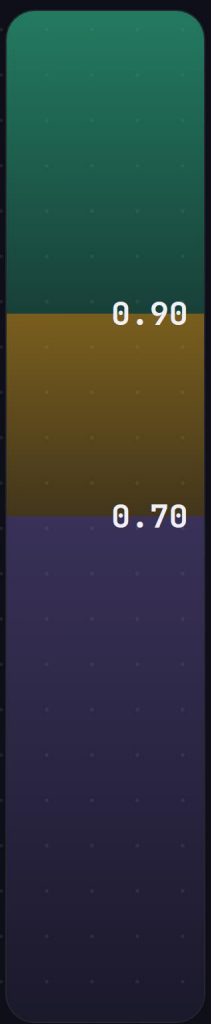
0.71

The only judgment question. Its lowest confidence.

All four, correct. And it hesitated where it should.

THE RULE

Confidence decides what happens next.



`0.90 or higher`

Move on.

No other model gets involved.

`Between 0.70 and 0.90`

The orchestrator decides.

And writes down why.

`0.70 or lower`

Escalate to the brain.

The expensive model only steps in when there is real ambiguity.

MEASURED

That full evaluation cost this much.

961
input tokens

112
output tokens, not billed

919 ms
all four questions in a single request

$0.00004
per evaluation

±0.02 variation across identical runs. No decision changed.

WHAT IT DOES NOT DO

Knowing its limits is part of using it well.

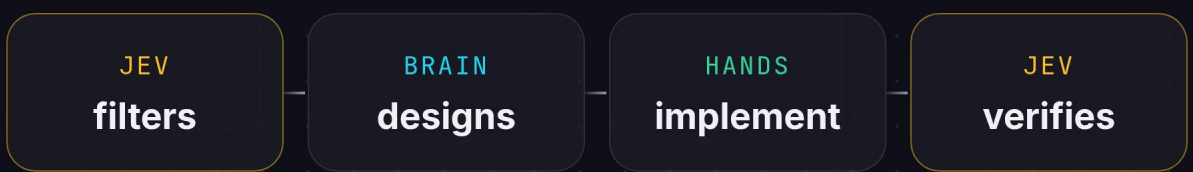
It does not design.

It only accepts `yes or no`, `choice`, or `score` questions. It picks among options and never proposes the missing one.

It does not find the evidence.

It reasons very well over what you already gathered. Finding it is still your job.

THAT IS WHY IT NEVER TAKES THE BRAIN'S SEAT



WHAT I TAKE AWAY

A model that is right is useful.

One that knows when to doubt is trustworthy.

YOUR TURN

Which AI agent decision do you still sign off by hand, and what would it have to prove for you to let it go?