THE 3-JUDGE PRINCIPLE

Don’t trust a single judge

How AI quality assurance borrowed the courtroom jury and why the winning panel is exactly three.

THE LINEAGE

PRE-2023

Human juries

The wisdom of crowds, long before AI: use an odd number of raters and take the majority verdict — never trust a single annotator.

Inter-rater reliability — Cohen’s κ, Krippendorff’s α

JUN 2023

LLM-as-a-Judge

A strong model can grade another: GPT-4 matches human preference >80% of the time — but favours its own text, longer answers, and whatever it reads first.

Zheng et al., “Judging LLM-as-a-Judge”, NeurIPS 2023

APR 2024

Judges → juries

Swap the lone judge for a panel of different model families. It beats a single large judge, shows less intra-model bias — and costs over 7× less.

Verga et al., “Replacing Judges with Juries” (PoLL)

2025–26 · TODAY

Gated in the pipeline

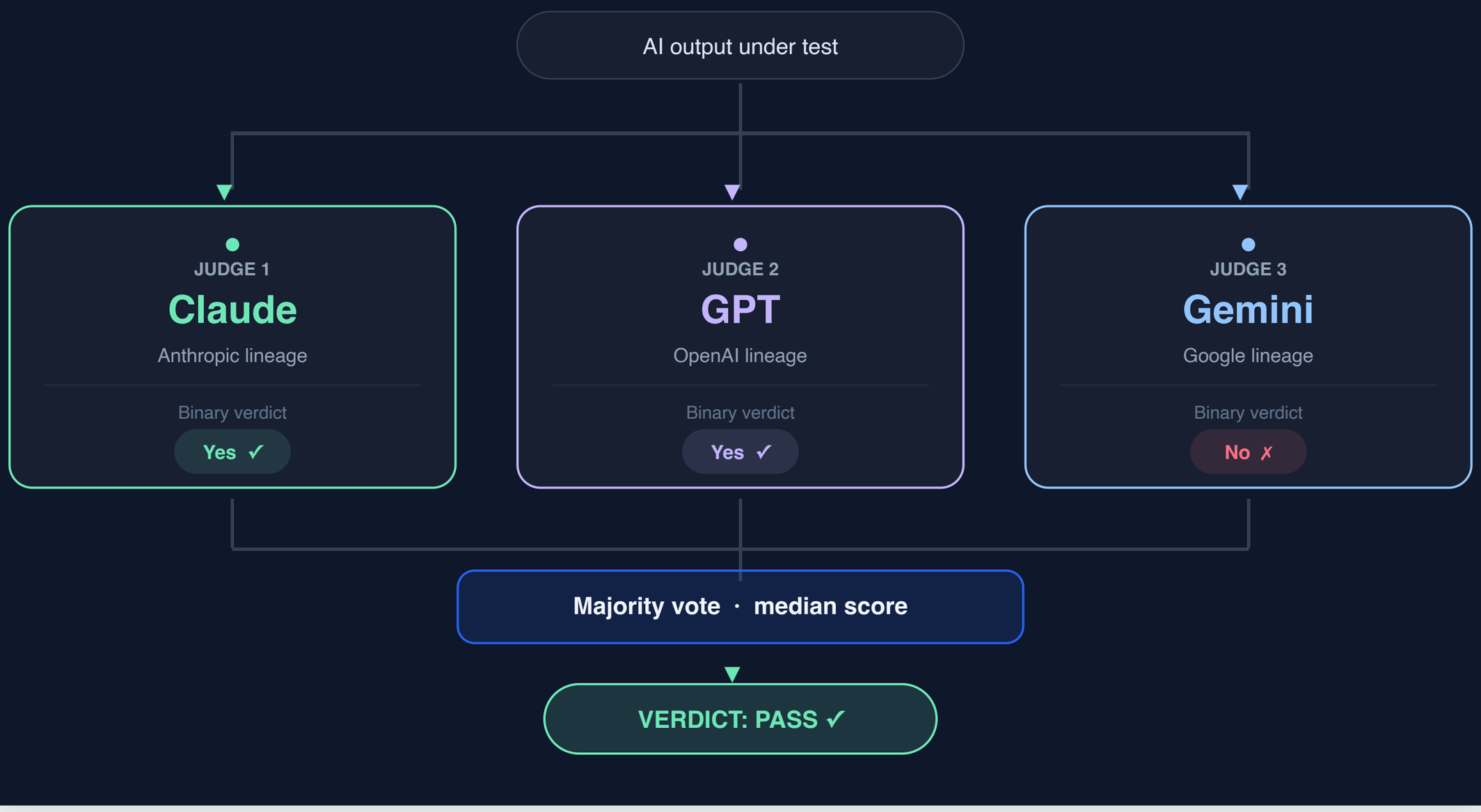
Calibrate each judge against a golden set of known-good and known-bad cases. If the panel drifts after a model update, the release gates — no spend on unreliable runs.

Continuous-evaluation practice — LangSmith, Arize, Braintrust

So how does that panel actually work?

THE PANEL AT WORK

One output. Three lineages. Majority rules.



Three lineages, a binary rubric, majority rules — each model’s blind spots cancel out.

Sources — Zheng et al., “Judging LLM-as-a-Judge”, NeurIPS 2023 · Verga et al., “Replacing Judges with Juries” (PoLL), 2024