TypeSafe's Jev told me it was 75% confident.

Human annotators agreed 10% of the time.

I audited whether a decision model's confidence scores mean what they claim. 8,000 judgments, measured against human labels.

0.91	0.156	$0.05
AUC ranking	ECE calibration	total cost

Swipe →

WHY IT MATTERS

Jev doesn't write text. It returns a number.

It's a "decision model": you hand it state, it hands back a probability. Which means your code ends up doing this:

- `if p > 0.9:  auto_approve()`
- `elif p < 0.1: auto_reject()`
- `else:  send_to_human()`

> Every threshold you write is a bet that the number is honest. So I checked.

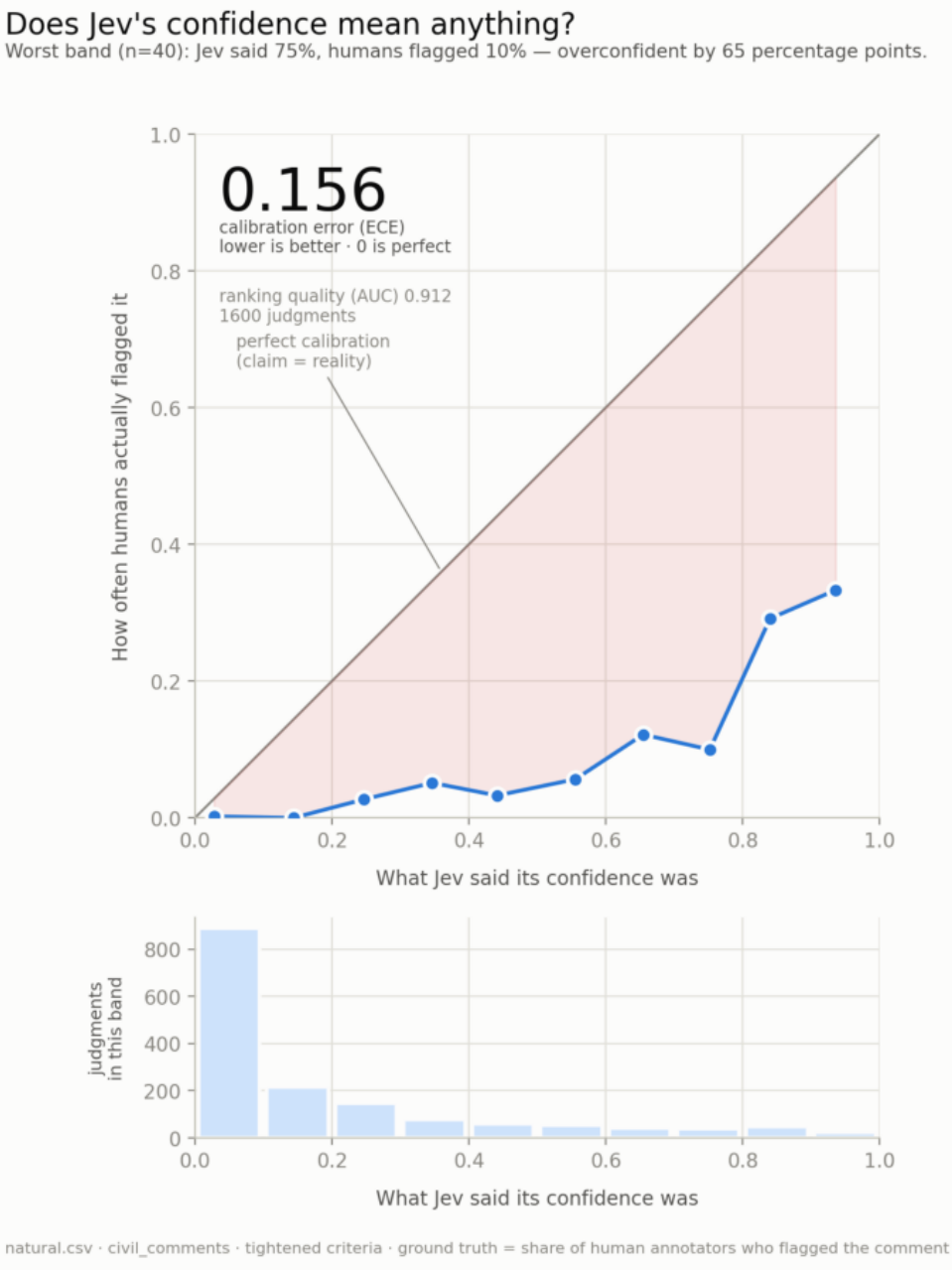
THE METHOD

Ground truth from humans, not models.

civil_comments shows each comment to a panel of annotators and records what fraction flagged it. A toxicity of 0.7 means 7 in 10 real people agreed.

- 8,000 judgments across two base rates
- Four wordings, tested head to head
- One pinned model version (jev-1.13.0)

Total cost of the entire study: five cents.



Every band sits below the diagonal: a systematic lean toward "yes", not an artefact of where I drew the threshold.

THE SPLIT

The ranking is good. The numbers are not.

A model that emits probabilities has two ways to be good, and they fail separately.

- AUC 0.91 — it reliably sorts worse content above better. It understands the task.
- ECE 0.156 — the probabilities attached to that ranking don't mean what they say.

On one question, Jev at the default threshold scored 61% where answering "no" every time scores 68% — while ranking at AUC 0.83.

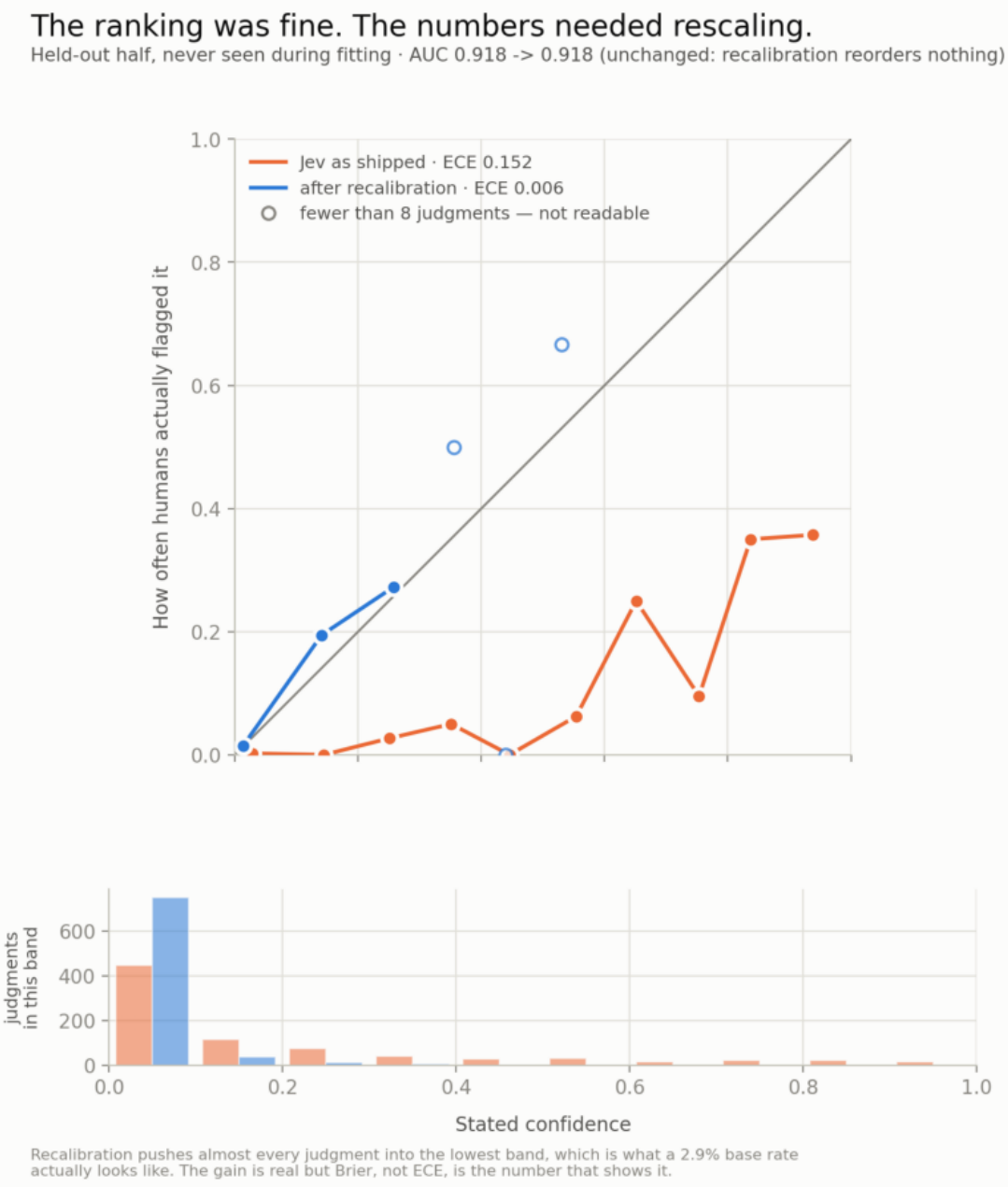
THE OBVIOUS OBJECTION

"Your prompts were just badly written."

Fair. So I tested four, in a single call over identical state, so only the phrasing varied.

- Error ranged 0.21 to 0.52 — and the annotators' own definition scored worst
- AUC barely moved: wording shifts the scale, not the signal
- The strictest still predicted 26% where reality was 5%

Re-ran everything with the best wording. Error fell 0.209 → 0.156. The bias survived.



Two parameters, fitted on one half of the data and scored on the half it never saw.

96%

of the calibration error removed

A logistic fit — about 20 lines — on held-out data. AUC went 0.918 → 0.918, because a monotonic rescale reorders nothing.

The ranking was always there. Only the units were wrong.

HONESTLY

What I'd want asked back at me.

- One dataset, one domain, one model version.
- One question has a single positive in 400. Meaningless, and labelled as such rather than dropped.
- Four wordings is four, not all of them.
- I nearly called 95% accuracy a win — flagging nothing scores 97% here.

This isn't an argument against the model. As a high-recall pre-filter with a calibration layer, it works.

THE TAKEAWAY

Typed output guarantees the interface. Not the truth.

If a model hands you a probability and you write a threshold against it, measure it first. Mine took an evening and cost five cents, and every threshold I'd have written was wrong.

Full code, charts and raw data: github.com/Adilmp/does-jev-confidence-mean-anything