

# Jev vs Gemini

TypeSafe's new “System One” model doesn't write text. It picks an answer and tells you how sure it is. We ran it against two Google models on 1,565 business emails.

|                                   |                             |                         |                       |
|-----------------------------------|-----------------------------|-------------------------|-----------------------|
| <b>1,565</b><br>emails classified | <b>3</b><br>models compared | <b>10</b><br>categories | <b>2</b><br>languages |
|-----------------------------------|-----------------------------|-------------------------|-----------------------|

---

## THE TASK

### Sort an email into one of ten buckets

A German B2B supplier gets mail all day: new orders, requests for a quote, invoice disputes, address changes, auto-replies. Something has to decide which team each one goes to.

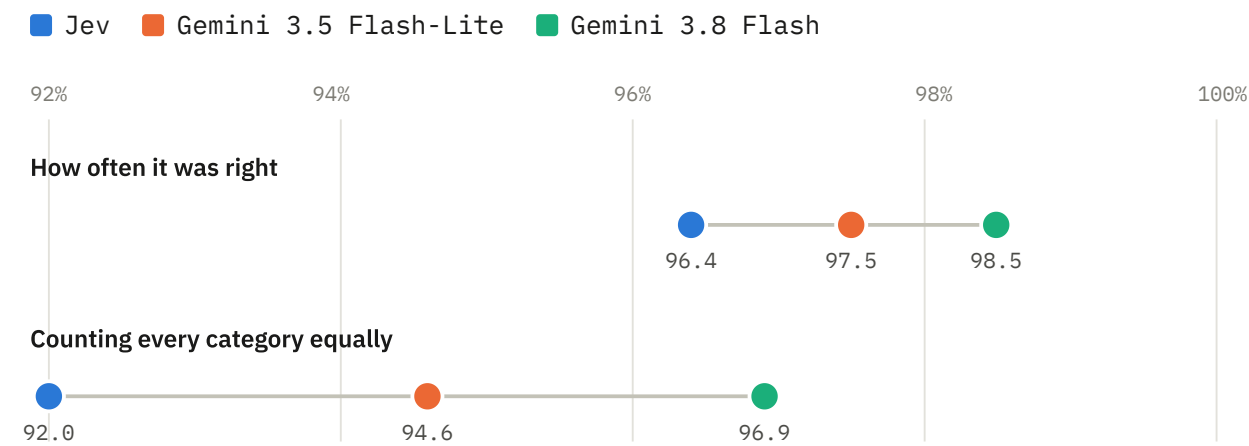
That's the job we gave all three models. Same 1,565 emails, same ten categories, same instructions, one after another. No chat, no summaries. Just pick a bucket.

Most of the emails are real, pulled from four company mailboxes. Real mail is lopsided, though. Nine in ten are quote requests and some categories barely appear, so we added emails written specifically to cover the rare ones. Roughly three quarters are German, the rest English.

---

# Jev is a little less accurate

Not by much.



The second row is the fairer test. Because nine in ten real emails are quote requests, a model that answered “quote request” to everything would look good on the first row and terrible on the second.

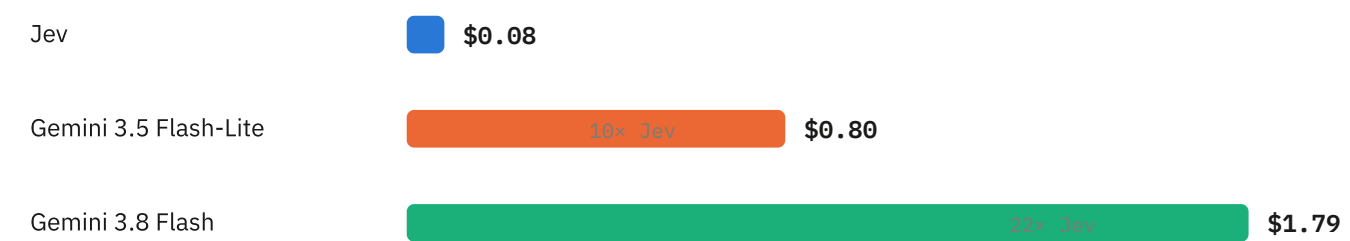
Raw percentages invite over-reading, so we ran the proper test for comparing two models on the same items. It asks whether one wins the disagreements more often than a coin flip explains.

| WHERE THEY DISAGREED     | JEV RIGHT | OTHER RIGHT | VERDICT                  |
|--------------------------|-----------|-------------|--------------------------|
| vs Gemini 3.5 Flash-Lite | 14        | 31          | Flash-Lite better        |
| vs Gemini 3.8 Flash      | 11        | 44          | 3.8 Flash clearly better |

Put in emails rather than percentages, that's Jev getting about one more wrong per hundred than Flash-Lite, and two more per hundred than the bigger model. Consistent enough across 1,565 emails to be a real difference rather than luck, and small enough that nobody would pick a model on it.

FINDING 02

# Jev costs a tenth as much



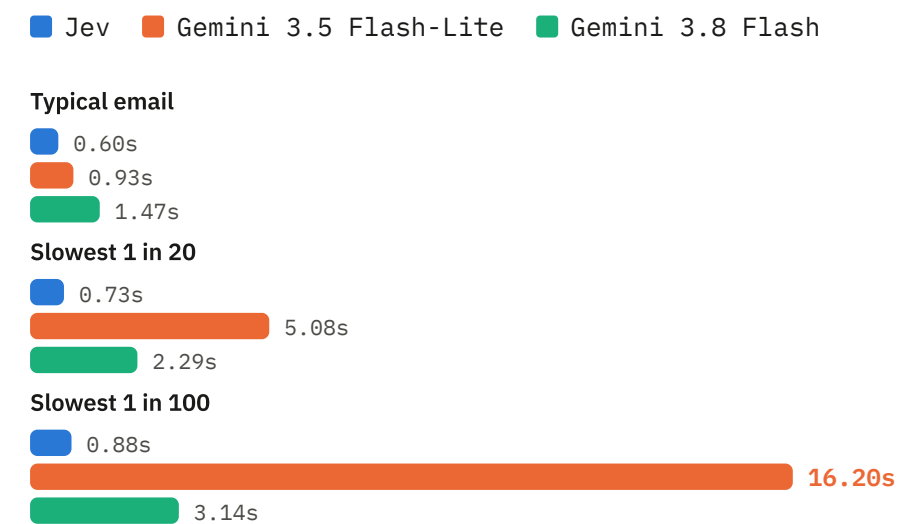
Cost per 1,000 emails, measured from actual token usage rather than list prices.

At those rates, running a thousand emails through Jev costs about **8 cents**. Part of that gap is the input price, which is about seven times lower. The rest is that TypeSafe doesn't charge for output at all, where output is roughly a third of what the Gemini runs cost.

FINDING 03

# The slow answers are where it shows

Typical speed is close. What separates them is the bad days.



Jev's whole range, from typical to worst, spans about a quarter of a second. Flash-Lite's spans sixteen.

Jev's slowest answer out of 1,201 took **1.3 seconds**. Flash-Lite's took **57 seconds**. Flash-Lite was also the only model that hit rate limits, 19 times in total.

If someone is waiting on the screen, the slow tail is what they actually experience.

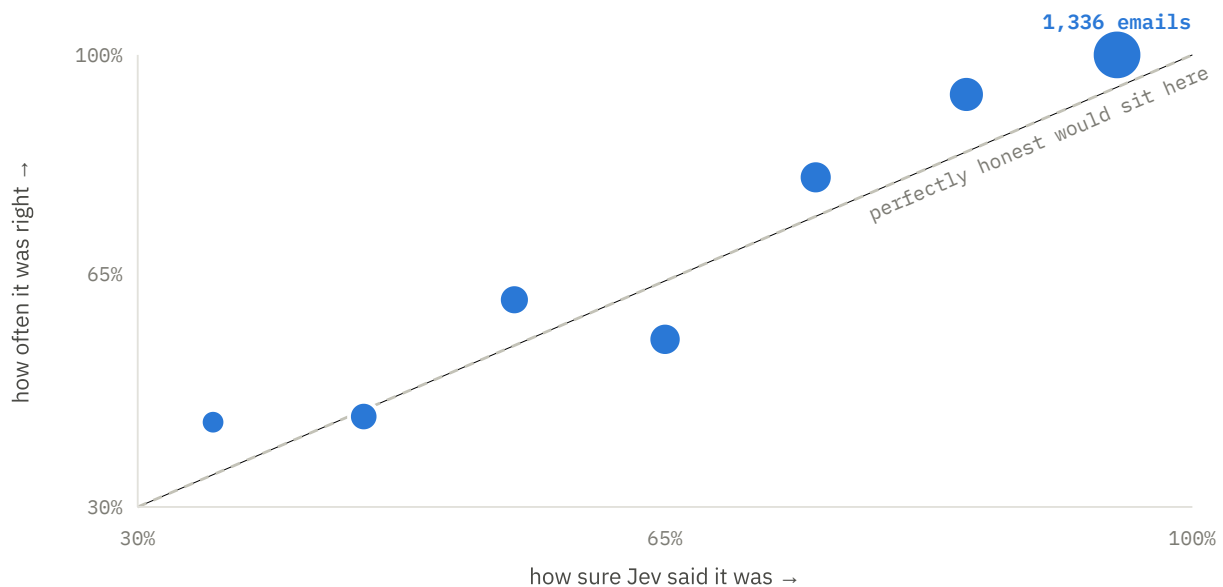
#### FINDING 04 • THE MAIN ONE

## It tells you which answers to trust

Jev doesn't pick one answer. It returns a probability for every one of the ten options, because that's how the model works, not because you asked it to.

A normal language model can't do this. You can ask one to write a confidence number next to its answer, but that number is just more generated text. It's writing what a confident answer looks like, not measuring anything. In our runs both Gemini models wrote **1.0** for the overwhelming majority of emails and never once went below 0.6.

So the question isn't which model has better confidence scores. It's whether Jev's are worth anything at all. They are.



Each dot is a band of answers; bigger dots hold more emails. When Jev says it's 85% sure it's right 94% of the time, slightly pessimistic, which is the safe direction to be wrong in.

**737 of the 1,565 answers came back at 99% confidence or higher. Every single one was correct.** Meanwhile, answers below 70% confidence were wrong nearly half the time. Errors are **13× more likely** down there.

That's the whole point of the thing. The mistakes aren't scattered randomly through the pile; they sit in a corner the model points at.

## So how much can you actually automate?

Sort every answer by confidence, take the most confident ones first, and see how far down you can go before the mistakes start.



Jev holds a perfect record through the most confident 85.5% of the mail. Push to 95% and it is still 98.9% correct. Include everything and it lands at its overall 96.4%.

On this benchmark that means **85% of the mailbox** came back with nothing wrong, and the remaining 15% is where the hard ones sat. That is a retrospective result on our data, not a guaranteed automation rate.

That isn't a claim about Jev being smarter. It's a different kind of output: a measured probability instead of a sentence that sounds sure.

---

HONESTLY

## What this doesn't prove

- **No human checked most of the answers.** For the real emails, a third model, Claude Opus 5, which never saw any result, decided what “correct” meant. Only the written emails have answers we know for certain, and on those Jev scored 96.2%.
- **Attachments were never tested.** 84% of the real emails had files attached. No model was shown any of them. That's the biggest gap between this test and real life.
- **One task.** Sorting email into buckets is exactly what Jev is built for. It can't write, can't do arithmetic, can't compare dates. None of which we tested, because none of it applies here.
- **The written emails are easier than real ones.** One author wrote all 364 of them, so they're cleaner and more consistent than anything a real inbox produces.

- **Real mail is lopsided.** Nine in ten are quote requests, and six categories had fewer than 15 real examples, which is why we wrote more.
- 

#### IN ONE SENTENCE

Jev is a shade less accurate than the cheap Gemini model, costs a tenth as much, never took longer than 1.3 seconds, and it is the only one that hands back a real probability instead of a sentence that sounds sure.

**Method.** 1,565 emails: 1,201 real ones from four company mailboxes, plus 364 written to cover categories real mail rarely produces. All three models got identical text and identical category definitions, taken straight from the production prompt. Jev pinned to `jev-1.13.0` ; Gemini called directly on Vertex AI. Attachments excluded for all models. Reference answers for the real emails from Claude Opus 5, which never saw any model's output. Significance by McNemar's exact test on paired predictions across all 1,565. Overall accuracy is reported for completeness but the category-weighted figure is the fair comparison, since the mix of categories in real mail is so uneven. Confidence measured as top-choice probability for every model.