

Everyone talks about Jev. Have you met Kev?

An open model on a consumer graphics card against
TypeSafe's Jev. 542 questions, in Swedish and English.

Swipe 

WHAT IS A DECISION MODEL?

It doesn't write an answer. It chooses.

Which department should get this email?

one choice out of three

Is the customer threatening to cancel?

yes or no, with a probability

How urgent is the ticket?

0, 1 or 2

The answer comes back in a fraction of a second, always in the right format, and you can set thresholds on the probability. Useful for ticket triage, risk flags and voice commands.

THE GAP TO JEV

How many answers it was behind.

23 → 9

Kev-4B out of the box

after two changes

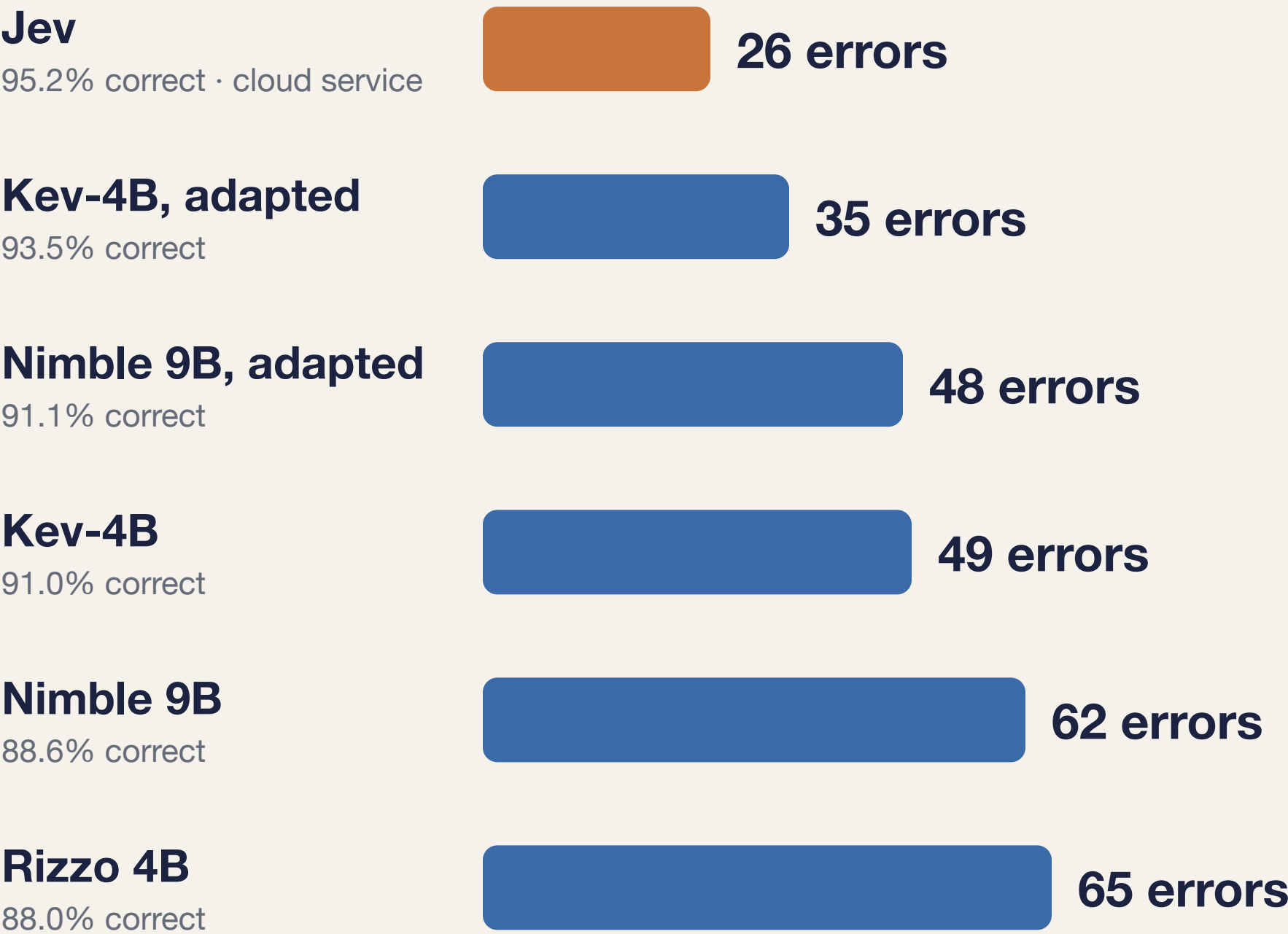
Correct answers behind Jev, out of 542 questions. Jev got 516 right. Kev-4B got 493, then 507.

No training. Same model, same weights.

RESULTS

Errors out of 542 questions

Fewer is better.



The local models ran on our own graphics cards (RTX 3060 and 5060 Ti). Jev is a cloud service. Eight scenarios, Swedish and English.

WHAT WE CHANGED

Two changes. No training.

1

Label the text

Every voice command got the prefix “Swedish voice command (may contain speech recognition errors):”.

2

Move one threshold

The question “is the customer angry?” counted as yes only at probability 0.94, not 0.5.

49



35

errors out of 542 for Kev-4B, 29% fewer
Chosen on half of the texts, checked on the other half.

SIZE

Bigger wasn't better.

Nimble 9B

9 billion parameters

18 GB

two graphics cards

62

errors out of 542

Kev-4B

4 billion parameters

9 GB

one graphics card

49

errors out of 542

Less than half the size. 13 fewer errors.

Kev-4B ran on one 12 GB graphics card. Nimble 9B needed both cards to fit.

SPEED

Under 100 milliseconds.

74 ms

Kev-4B on an RTX 5060 Ti. 87 ms on an RTX 3060. Median time per case in the support scenario.

Jev: 230 ms

That includes the network. Jev is a cloud service, so it isn't pure GPU time and the comparison isn't entirely fair.

CAVEATS

What doesn't hold up yet.

Swedish voice commands

Jev got 32 out of 32. The best open models, with adapted questions, got 29 and 31.

Small test sets

16 to 32 texts per scenario. One case is worth 3 to 6 percentage points, so one or two cases are noise.

Selection and control

The changes were chosen on half of the texts and checked on the other half. The result needs testing on more texts before you build on it.

YOUR TURN

Are you using decision models yet?

Or are you still prompting a chat model with a thousand words to get a yes or no?

Our own measurements, October 2026. RTX 3060 and 5060 Ti. 542 questions, eight scenarios.