So what is Jev?

It's a new AI model.

So what? A lot of models come up every other week.

But Jev is different.

It's not an LLM.

QUESTION

Is this ticket urgent?

LLM

one token at a time

token 10

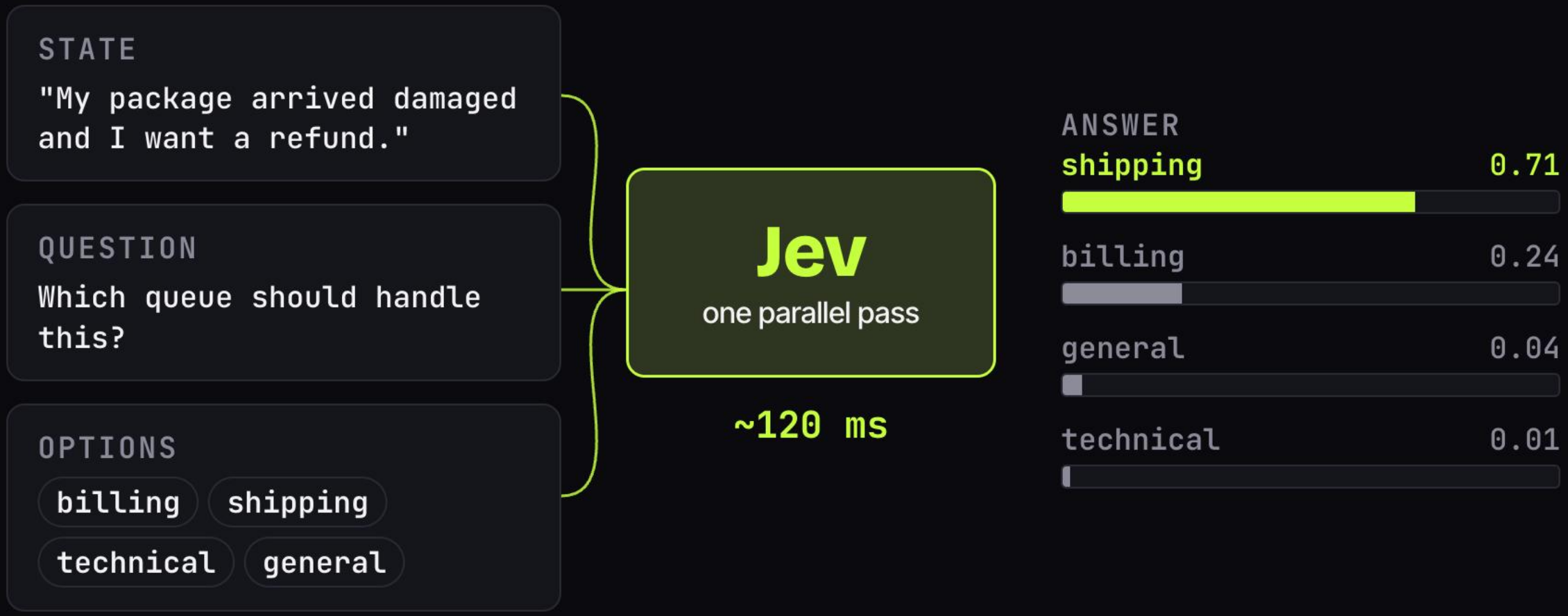
Jev does not generate text.

ANSWER

Yes, this ticket seems urgent because the customer …

Illustrative token stream — not model output.

Jev is an AI model built to make fast, structured decisions that software can use directly.



No text. No parsing. Just a typed answer with a probability.

Illustrative output — not a live model call.

So what? Any LLM can do this with structured output.

True. It can. Two things change.

01 **Speed**

02 **Cost**

Both measured on the same job: triaging a support inbox with Jev and with an LLM.

LIVE

First difference: speed.

Published: LLM 3 s – 329 s · Jev 70 – 500 ms · claim 40–200× faster

INBOUND EMAIL · WHAT IS THE NATURE OF THIS CUSTOMER REQUEST?

"Hi - we were charged twice for our May invoice and the second charge still hasn't been refunded. Can someone look into this today?"

LLM gpt-5-2025-08-07

4.07 s

billing — confidence 0.99 · 137 in / 216 out · last measured

Jev jev-1.13.0

650 ms

billing — confidence 1.00 · 387 in / 45 out · last measured

Jev answered 6.3× faster — same answer.

Figures are TypeSafe's own, from their launch post (Sep 2026). Independent benchmarks are still limited.

LIVE

Second difference: cost.

Per million tokens — in: LLM ₹119.50 · Jev ₹4.02 out: LLM ₹956.00 · Jev free

Same 1,000-token input for both · 10,000 emails a day.

	LLM	Jev
Per email	₹0.326	₹0.004
Per day	₹3,260	₹40
Per year	₹11,89,885	₹14,655

81.2× cheaper — ₹11,75,230 a year saved

Vendor headline 444.6× on their evals; here 81.2×, almost all of it the LLM's reasoning tokens.

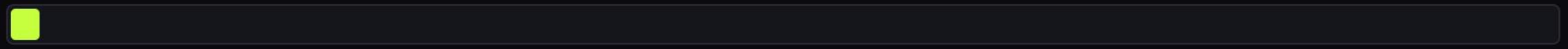
Figures are TypeSafe's own, from their launch post (Sep 2026). Independent benchmarks are still limited. LLM list prices converted at ₹95.6 / USD (23 Sep 2026).

Other benefits

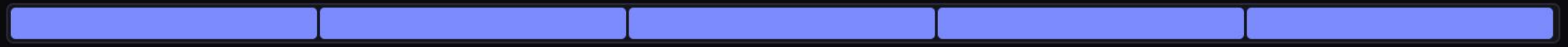
9.1 Many questions, one pass 9.2 Confidence you can act on 9.3 It can't make things up

Shared axis · full width = 7.0 s

Jev — 5 questions ≈ 130 ms



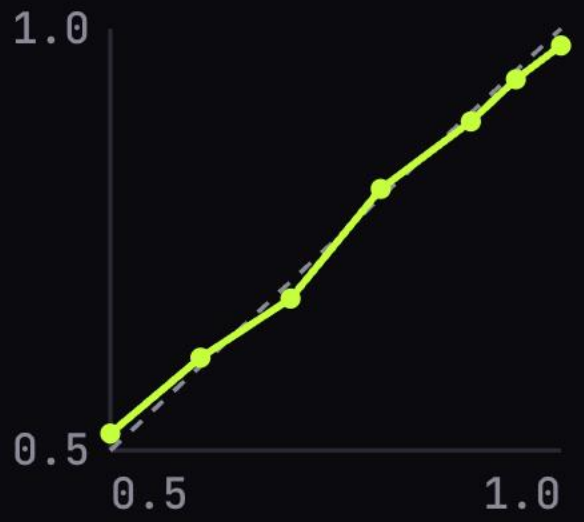
LLM — one call per question ≈ 7.0 s



Adding a question adds almost no time or cost.

Illustrative output — not a live model call.

Other benefits



Trained so that 90% confident means right about 90% of the time.

x: stated confidence · y: actual accuracy
— target against the dashed ideal diagonal

Calibration is Jev's training objective (RLCD). TypeSafe has not yet published a calibration curve.

Illustrative output — not a live model call.

It can't invent an answer. It can still pick the wrong one.

That's why the confidence score matters (9.2).

Illustrative output — not a live model call.

Who built Jev?

DA

Diogo Almeida

ex-OpenAI · InstructGPT · ChatGPT

Now building AI for software, not people.

TypeSafe AI

Software needs System 1 thinking, but we keep building System 2 thinking.

AI moves from a feature to a primitive.

```
if jev("Is this customer angry?") > 0.9:
```

Illustrative

AI stops being the product.
It becomes plumbing.

01

AI agents

Which tool? · Which model? · Is this step safe?

Every small decision, fast and cheap enough to check.

02

Business operations

Support triage · Claims · Invoices · Lead scoring

Automate the clear cases. Send the unclear ones to a person.

03

Trust & safety

Moderation · Fraud · Spam · Jailbreak detection

Score, threshold, escalate, on every single request.

04

Unstructured → structured data

Logs · Catalogs · Tickets · Call transcripts

Stop sampling. Label everything.

05

Real-time systems

Game characters · Live UI · Control loops

Not just cheaper. Newly possible.

So, what are people building using Jev

Demo

Architecture

What's actually under the hood?

Some of this TypeSafe has told us. Some we're inferring.

CONFIRMED UNCONFIRMED

Eight pieces. Three are still open questions.

UNCONFIRMED

01

Transformer-based — but which kind?

Encoders read all at once (BERT).
Decoders write word by word (GPT).

Jev acts encoder-like. Could be a decoder with a scoring layer. Open question.

UNCONFIRMED

02

Broad world knowledge (maybe)

Does it know general facts, like a chat model does?

Its Wikipedia-navigation demo suggests some. How much is unknown.

CONFIRMED

03

Built for System 1 tasks

Quick gut judgments: spam? urgent? which team?

Fast, frequent decisions inside software — not reasoning or writing.

UNCONFIRMED

04

Trained on synthetic data

Training data generated by computers, not collected from people.

TypeSafe hasn't explained how its data is made.

CONFIRMED

05

Non-autoregressive

Chat models write one word at a time, each waiting on the last.

Jev produces the whole answer in one go. Much faster.

CONFIRMED

06

Schema-constrained output

You give it a fixed list of allowed answers up front.

It can only pick from that list. It can't make one up.

CONFIRMED

07

Parallel sampler

Ask five questions, it answers all five at once.

Adding questions barely slows it down.

CONFIRMED

08

RLCD — calibrated confidence

Every answer comes with a confidence number, trained to be honest.

Say '80% sure' → right about 80% of the time. Fixes overconfidence.

01

No independent benchmarks

Every speed and cost number comes from TypeSafe.

They skipped public leaderboards. Independent audits are only starting.

02

Text only

It can't see images or hear audio.

Every real-time demo feeds it a text description of the world.

03

No web search, no outside knowledge

It only reasons over the state you hand it.

By design — it's a decision layer, not a research agent.

04

It explains nothing

Just an answer and a probability.
No reasoning to inspect.

And 'can't hallucinate' isn't 'can't be wrong' — it picks confidently.

05

Known weak spots

Reads literally. Weak at math, counting and dates.

Accuracy drops when you feed it irrelevant state.

06

Closed and early

No open weights. One vendor. Waitlisted early access.

Pricing may be subsidised, and the model can change under you.

SAFE BET

01

Everyone copies it

Qwen clones appeared in days. JSON mode all over again.

Jev keeps the data and calibration. Not the idea.

SAFE BET

02

It vanishes into the stack

Already inside Vercel, LiteLLM, and a Postgres jev() function.

You'll use it without ever calling it.

LIKELY

03

One brain, many reflexes

A big model plans. Cheap decisions run every step between.

The real skill: splitting a goal into small questions.

LIKELY

04

Dashboards for doubt

Wrong answers are silent. Someone has to watch the numbers.

Calibration graphs and drift alerts become standard kit.

LIKELY

05

A new job title

Writing the options and thresholds is the actual work now.

Prompt engineering, round two, backlash included.

WILD CARD

06

Give it eyes

Right now it reads a text description of Doom.

Let it read the screen, and robotics opens up.

WILD CARD

07

Too cheap to not use

A decision costs $0.00004. So you check everything, always.

Per keystroke. Per scroll. Per log line.