

Decision models stop charging for output and undercut LLMs on routine agent calls

Five firms shipped “decision models” in seventeen days: systems that answer a typed question with a choice and a probability, and bill output at zero. On a routing call they are about 280 times cheaper than a frontier model, but only five times cheaper than a small LLM. The shift is real at the expensive end of the agent stack, and the evidence for it is still graded mostly by the vendors.

AUTHOR



Alicante Tech Vibes *Research Labs*

~8 MIN READ
DATED

ALL FIGURES SOURCED &

\$0.042 JEV PER MILLION INPUT TOKENS; OUTPUT FREE · TYPESAFE, 15 SEP 2026	\$0.24 CLEF PRICE PER MILLION TOKENS, WORKERS AI · THE REGISTER, 1 OCT 2026	38.8 ms CLEF-FLASH MEDIAN LATENCY (VENDOR) · CLOUDFLARE, 1 OCT 2026
~13% PAID VERCEL TEAMS TRYING JEV IN 24 HOURS · VERCEL, 18 SEP 2026	\$10 / \$50 FABLE 5.1 PER MILLION INPUT / OUTPUT TOKENS · ANTHROPIC, SEP 2026	66.8% JEV CLASSIFICATION AGREEMENT; FINE-TUNED 74.5% · GLEAN, 29 SEP 2026

I. Five decision models in seventeen days

On 15 September, TypeSafe AI, a start-up led by former OpenAI researcher Diogo Almeida, released [Jev](#), which it calls a “System One” model. Jev takes a state and a typed question and returns a choice, a score or a yes/no answer, each with a probability. It writes no text. TypeSafe priced it at \$0.042 per million input tokens and billed output at zero, describing it as too cheap to meter.

Others moved fast. Fastino Labs shipped the open-weight GLiNER2.5-Decide on 24 September and the hosted GLiDE on 30 September, [according to MarkTechPost](#). Upstage listed [Solar Decide](#) on OpenRouter on 28 September at \$0.10 per million input tokens, currently discounted to \$0.05, with output free. On 1 October, Cloudflare released [Clef](#) and [Clef-flash](#) under Apache 2.0, and AWS released the 1.9-billion-parameter [Strands Decider 2B](#) as open weights.

One closed product became a product category in under three weeks. What the launches share is a billing model: pay for what the model reads, not for what it writes.

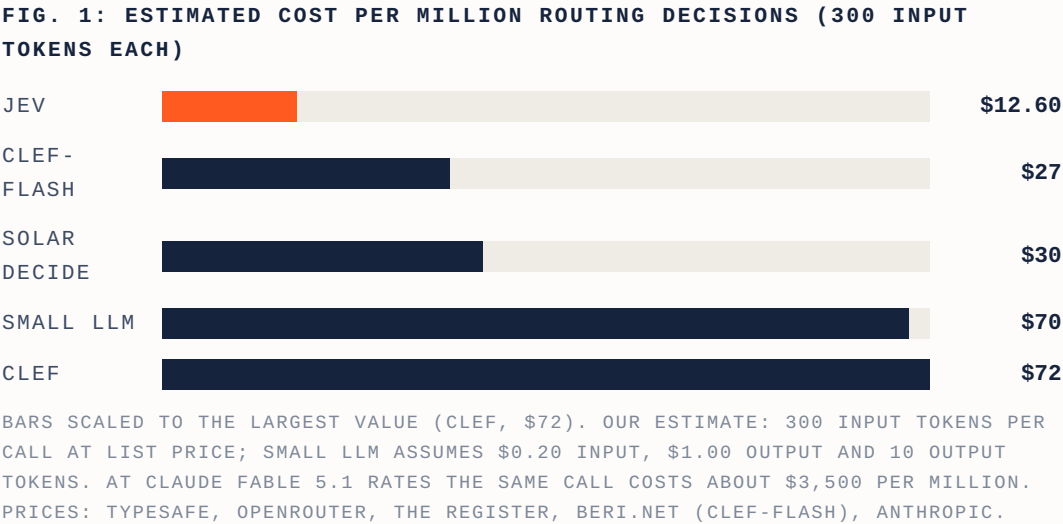
II. The bill moves from what a model writes to what it reads

LLM pricing is built around output. TypeSafe’s [launch post](#) notes that existing models charge between \$0.20 and \$10 per million input tokens, with output typically about five times dearer. Anthropic lists [Claude Fable 5.1](#) at \$10 input and \$50 output. A decision model emits a label and a number, so the output line on the invoice disappears.

The second shift is the interface. Cloudflare built Clef to be fully compatible with Jev’s API, so customers can swap one for the other. Upstage’s listing says Solar Decide supports the same /v1/systemone schema. Within a fortnight, a start-up’s request format became the shared plug for a category, which is how commodities form.

Demand signals arrived early. [Vercel reported](#) on 18 September that nearly 13% of paid AI Gateway teams tried Jev within 24 hours, twice as many as any previous model launch on the gateway. Arize’s Laurie Voss [told TechTarget](#) on 22 September that the big labs would follow quickly and that

Jev would find it lacks much of a moat. Cloudflare and Amazon proved the point nine days later.



III. The honest technical read: a better-packaged classifier

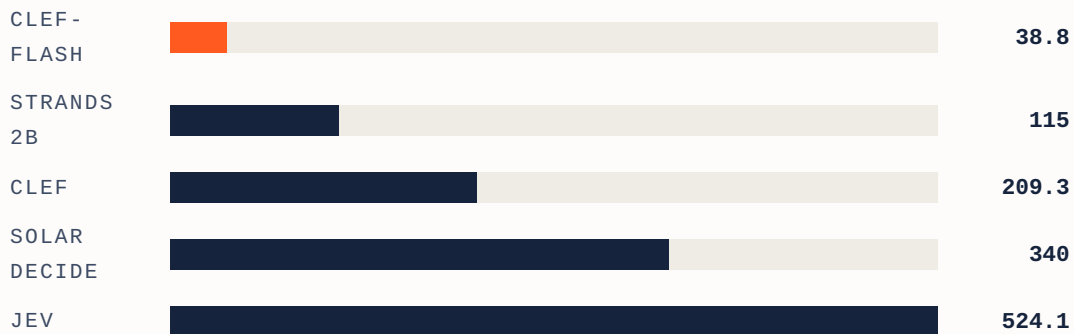
What changed is narrower than the launch language suggests. Decision models cannot write text, code or summaries, and they need a predefined schema. Jev cannot answer outside its schema, but, as one [early review](#) put it, it can still be confidently wrong. The genuine advance is ease of use: zero-shot typed decisions through a hosted API, with calibrated probabilities, and no training run.

That is the skeptic’s strongest point. Glean’s search team [tested Jev on four enterprise tasks](#) and concluded that it does not change the fact that classifiers already existed. On query classification Jev agreed with the production baseline 66.8% of the time, against 74.5% for a fine-tuned model. It underperformed production systems on reranking. It won elsewhere: an 8.1-fold speed-up on model routing, and 2.4% disagreement on citation judging against 17.9% for LLM alternatives.

The headline numbers are vendor-graded. TypeSafe’s claim of results up to 193.6 times faster and 444.6 times cheaper compares Jev with frontier models taking 3 to 329 seconds per task, which includes long chains of reasoning a classifier never runs. TypeSafe concedes it cannot prove its price is not subsidised. [The Register](#) notes Cloudflare’s scores have not been reproduced on the official Decision Index.

The benchmarks also disagree with each other. Cloudflare's tables, summarised by [Beri.net](#), show Clef ahead of Jev on intent detection (94.20 against 79.74 on BANKING77) but behind on When2Call, the test of whether an agent should act at all (72.37 against 80.97). Clef-flash drops to 66.77 on out-of-scope detection where Clef scores 97.43. Routing is solved; judgment is not.

FIG. 2: MEDIAN DECISION LATENCY, MILLISECONDS



BARS SCALED TO THE SLOWEST VALUE (JEV, 524.1 MS). CONDITIONS DIFFER AND ARE NOT COMPARABLE: CLEF, CLEF-FLASH AND JEV MEASURED BY CLOUDFLARE; STRANDS ON AN RTX 3090 (MARKTECHPOST); SOLAR DECIDE IS OPENROUTER P50 INCLUDING NETWORK. TYPESAFE QUOTES 70-500 MS FOR JEV.

Against a frontier model, a decision model is a 280-fold discount. Against a small LLM, it is a rounding error. The disruption lives at the top of the price list, not the bottom.

OUR ESTIMATE FOR A 300-TOKEN ROUTING CALL: \$12.60 PER MILLION DECISIONS ON JEV, ABOUT \$70 ON A SMALL LLM AT \$0.20/\$1.00, ABOUT \$3,500 ON CLAUDE FABLE 5.1 AT LIST PRICE, OCTOBER 2026.

IV. The economic channels

Prices. Our estimate, using list prices and a 300-token routing call: Jev costs \$12.60 per million decisions, Clef \$72, a small LLM about \$70 and Claude Fable 5.1 about \$3,500. Take an agent platform running one million tasks a month with 20 control calls each. Routing those through a frontier model costs about \$70,000 a month; through Jev, about \$250. The saving is real for teams that default to their best model for everything. For teams already routing to small models, it is closer to \$1,100 a month. Clef, priced at nearly six times Jev, saves almost nothing over a small LLM.

Market structure. Open weights under Apache 2.0 put a ceiling on what any closed decision model can charge, since anyone with the hardware can host Clef or Strands. But hardware is the catch: [The Register](#) puts Clef-flash at 41GB-plus of GPU memory and Clef at 85GB-plus. Strands Decider runs on a consumer RTX 3090, but no hosted provider serves it yet. Hosting, not weights, decides the price.

Productivity. Latency compounds inside agent loops. [Cloudflare says](#) Clef classified websites in its threat-intelligence workflow in 2.2 seconds against 4.7 for the fastest LLM it tested. A loop with ten gating decisions saves seconds per task, which matters more for interactive agents than the cost line.

Capital and labour. The direct labour effect is small; this is infrastructure, not a worker substitute. The capital effect runs against frontier labs: the cheapest, highest-volume slice of their API traffic is exactly the slice decision models target. Engineering effort shifts to writing schemas and setting confidence thresholds, which is where the probability output earns its keep.

v. Second-order: calibration becomes the product

The first consequence is architectural. A decision model that returns a probability lets software escalate the uncertain cases. Fastino already builds this in, with [GLiDE](#) adding extra reasoning on uncertain cases. [Reported results](#) show Strands answers given at 0.9 confidence or higher were right about 95% of the time on unseen short tasks. Expensive LLMs, and humans, move to the tail of hard calls. Ned Bellavance [told TechTarget](#) that human-in-the-loop workflows will still use an LLM.

The second is distribution versus demand. Elena Brooks [argued on 24 September](#) that Vercel's 13% shows reach through an existing gateway, not paid, repeated use. Jev is still in waitlisted early access. Nobody has published retention.

The third is security. Strands Decider's bundled server binds to localhost with no authentication, [MarkTechPost reports](#). A model that gates agent actions is a policy layer. Shipped open, it inherits every deployment mistake its users make.

VI. Precedent: the model cascade, rebranded

Routing cheap traffic to a small model and escalating the rest is an old pattern in machine learning, and intent benchmarks such as BANKING77, which Cloudflare cites, predate the LLM boom. What is new is packaging: a hosted, prompt-only API, a shared schema, and a billing model that zeroes output. Glean's summary fits: Jev makes a good zero-shot classifier much easier to use. Ease of adoption, not a new capability, is what moved 13% of a gateway in a day.

VII. How it plays out: three scenarios

Probabilities are our analytical judgment, not forecasts from a model, and they sum to 100%.

1 · Baseline: a standard layer beneath the LLM ~55%

6-18 MONTHS ·
STEADY

Decision models become a default component in agent gateways for routing, tool selection and guardrails, with prices clustering between \$0.02 and \$0.25 per million input tokens. Frontier labs answer with their own typed-decision endpoints. Watch for a decision endpoint from OpenAI, Anthropic or Google, independent reproduction of vendor benchmarks, and published repeat-usage data from a gateway.

2 · Acceleration: control traffic leaves the LLM ~20%

6-12 MONTHS ·
FAST

Calibrated decision models absorb most agent control calls, hosted prices fall below Jev's, and small-model API revenue at the labs visibly slows. Watch for a hosted Strands or Clef price under \$0.04 per million, cloud marketplaces bundling decision models free with agent tooling, and When2Call-style judgment scores closing the gap to Jev.

3 · Stall: structured output from cheap LLMs is good enough ~25%

3-9 MONTHS ·
ABRUPT

Independent tests confirm Glean's pattern: fine-tuned classifiers and small LLMs with structured outputs match or beat decision models on accuracy, and the price gap stays under an order of magnitude. Jev's price rises after early access. Watch for third-party leaderboards showing no accuracy edge, a Jev price change, and gateway usage falling after the trial window.

WHAT WOULD FALSIFY THIS THESIS

The claim is that decision models create a cheap layer that pulls high-volume routine agent calls away from per-token LLM pricing. It fails if independent evaluations show accuracy losses large enough that escalations erase the savings, as Glean's query-classification result hints. It also fails if TypeSafe ends free output or raises Jev's price after early access, confirming the subsidy it could not rule out. And it fails if frontier labs cut small-model prices to within about twice a decision model's cost, which would make a separate layer not worth the integration.

WHAT TO WATCH, IN ORDER

1. **By 31 October 2026** — independent reproduction of Cloudflare’s Clef scores on the official Decision Index. The cleanest test of whether open weights match Jev.
2. **By 15 November 2026** — whether Jev leaves waitlisted early access at \$0.042 with output still free, which separates a launch price from a market price.
3. **By 30 November 2026** — any gateway or vendor disclosure of repeat paid usage, which tests Brooks’s distribution-versus-demand critique.
4. **By 31 December 2026** — whether OpenAI, Anthropic or Google ships a typed-decision endpoint, as Voss predicted, which would signal defence of the cheap-call tail.
5. **By 31 December 2026** — whether any hosted provider, including Amazon Bedrock, serves Strands Decider, and at what price, which sets the floor for the category.

SOURCES

1. TypeSafe AI, “Introducing System One Models & Jev,” 15 September 2026 · typesafe.ai
2. Flowtivity, “Jev by TypeSafe AI: Is the 200x Faster Decision Model Too Good to Be True?,” 16 September 2026 · flowtivity.ai
3. Vercel, “Jev is the fastest-adopted model in AI Gateway history,” 18 September 2026 · vercel.com
4. TechTarget, “Jev decision model touted as quicker, cheaper LLM alternative,” 22 September 2026 · techtarget.com
5. Elena Brooks, “Jev Adoption: What Fast Distribution Does and Does Not Prove,” Medium, 24 September 2026 · medium.com
6. Glean, “Jev and the return of the zero-shot classifier,” 29 September 2026 · glean.com
7. OpenRouter, “Solar Decide – API Pricing & Providers,” listed 28 September 2026, accessed 4 October 2026 · openrouter.ai
8. Cloudflare, “Clef: open-weight decision models,” 1 October 2026 · blog.cloudflare.com
9. Cloudflare Developers, “Clef models on Workers AI,” changelog, 1 October 2026 · developers.cloudflare.com
10. The Register, “Cloudflare tries to outplay Jev with open-weight Clef models,” 1 October 2026 · theregister.com

11. SiliconANGLE, "AWS debuts Strands Decider 2B, a first lightweight decision model," 1 October 2026 · siliconangle.com
12. MarkTechPost, "AWS Strands Labs Releases Strands Decider 2B," 1 October 2026 · marktechpost.com
13. MarkTechPost, "Decision AI Models Explained: TypeSafe Jev vs Fastino GLiDE, GLiNER2.5-Decide and Open-Source Competitors," 2 October 2026 · marktechpost.com
14. Dealroom, "Cloudflare launches Clef as TypeSafe's 'decision model' idea spreads across big tech," 2 October 2026 · dealroom.co
15. Beri.net, "Cloudflare's Clef Beats Jev at Routing and Loses at Judgment," 3 October 2026 · beri.net
16. Traictory, "Cloudflare ships Clef, an open-weight answer to Jev, with a price critics noticed," 3 October 2026 · traictory.com
17. Anthropic, "Claude Fable," product page (Fable 5.1 announced 1 September 2026), accessed 4 October 2026 · anthropic.com

METHOD NOTE: FIGURES ARE ATTRIBUTED AND DATED INLINE; WHERE SOURCES DISAGREE, THE DISAGREEMENT IS REPORTED RATHER THAN AVERAGED. BENCHMARK AND LATENCY FIGURES ARE VENDOR MEASUREMENTS UNLESS STATED, AND WERE TAKEN UNDER DIFFERENT CONDITIONS. COST-PER-DECISION FIGURES ARE OUR ESTIMATES FROM LIST PRICES WITH STATED ASSUMPTIONS (300 INPUT TOKENS PER CALL; 10 OUTPUT TOKENS FOR LLMS). THE CLEF-FLASH PRICE COMES FROM SECONDARY REPORTING ONLY. SCENARIO PROBABILITIES ARE ANALYTICAL JUDGMENT.

Built by the [Alicante Tech Vibes](#) community, a tope.

2026-10-04